QLSD: Quantised Langevin Stochastic Dynamics for Bayesian Federated Learning

Maxime Vono, Vincent Plassier, Alain Durmus, Aymeric Dieuleveut, Eric Moulines

QLSD: Quantised Langevin Stochastic Dynamics for Bayesian Federated Learning

Maxime Vono
Criteo AI Lab
Paris, France

Vincent Plassier
CMAP, École Polytechnique
Lagrange Mathematics and Computing Research Center

Alain Durmus
Université Paris-Saclay
ENS Paris-Saclay

Aymeric Dieuleveut
CMAP, École Polytechnique

Eric Moulines
CMAP, École Polytechnique

Abstract

The objective of Federated Learning (FL) is to perform statistical inference for data which are decentralised and stored locally on networked clients. FL raises many constraints which include privacy and data ownership, communication overhead, statistical heterogeneity, and partial client participation. In this paper, we address these problems in the framework of the Bayesian paradigm. To this end, we propose a novel federated Markov Chain Monte Carlo algorithm, referred to as Quantised Langevin Stochastic Dynamics which may be seen as an extension to the FL setting of Stochastic Gradient Langevin Dynamics, which handles the communication bottleneck using gradient compression. To improve performance, we then introduce variance reduction techniques, which lead to two improved versions coined `QLSD`* and `QLSD`$^{++}$. We give both non-asymptotic and asymptotic convergence guarantees for the proposed algorithms. We illustrate their performances using various Bayesian Federated Learning benchmarks.

1 INTRODUCTION

A paradigm shift has occurred with *Federated Learning* (FL) (McMahan et al., 2017; Kairouz et al., 2021). In FL, multiple entities (called clients) which own locally stored data collaborate in learning a "global" model which can then be "adapted" to each client. In the canonical FL, this task is coordinated by a central server. The initial focus of FL was on mobile and edge device applications, but recently there has been a surge of interest in applying the FL framework to other scenarios; in particular, those involving a small number of trusted clients (*e.g.* multiple organisations, enterprises, or other stakeholders).

FL has become one of the most active areas of artificial intelligence research over the past 5 years. FL differs significantly from the classical (distributed) ML setup (McMahan et al., 2017): the storage, computational, and communication capacities of each client vary amongst each other. This poses considerable challenges to successfully deal with many constraints raised by (i) partial client participation (*e.g.* in mobile applications, a client is not always active); (ii) communication bottleneck (clients are communication-constrained with limited bandwidth usage); (iii) model update synchronisation and merging.

Many methods derived from stochastic gradient descent techniques have been proposed in the literature to meet the specific FL constraints (McMahan et al., 2017; Alistarh et al., 2017; Horváth et al., 2019; Karimireddy et al., 2020; Li et al., 2020; Philippenko and Dieuleveut, 2020), see Wang et al. (2021) for a recent comprehensive overview. Whilst these approaches have successfully solved important issues associated to FL, they are unfortunately unable to capture and quantify epistemic predictive uncertainty which is essential in many applications such as autonomous driving or precision medicine (Hunter, 2016; Franchi et al., 2020). Indeed, these methods only provide a point estimate being a minimiser of a target empirical risk

Proceedings of the 25$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

Table 1: Overview of the main existing distributed/federated approximate Bayesian approaches. Column *Comm. overhead* gives the scheme employed to address the communication bottleneck. Column *Heterogeneity* means that the proposed approach tackles the impact of data heterogeneity on convergence while column *Bounds* highlights available non-asymptotic convergence guarantees.

Method	Comm. overhead	Heterogeneity	Partial participation	Bounds
Hasenclever et al. (2017)	local steps	✗	✗	✗
Nemeth and Sherlock (2018)	one-shot	✗	✗	✗
Bui et al. (2018)	local steps	✗	✓	✗
Jordan et al. (2019)	one-shot	✗	✗	✓
Corinzia et al. (2019)	local steps	✗	✓	✗
Kassab and Simeone (2020)	local steps	✗	✓	✗
El Mekkaoui et al. (2021)	local steps	✗	✗	✓
Plassier et al. (2021)	local steps	✗	✗	✓
Chen and Chao (2021)	local steps	✓	✓	✗
Liu and Simeone (2021a)	one-shot	✗	✗	✗
This work	compression	✓	✓	✓

function. In contrast, the Bayesian paradigm (Robert, 2001) stands for a natural candidate to quantify uncertainty by providing a full description of the posterior distribution of the parameter of interest, and as such has become ubiquitous in the machine learning community (Andrieu et al., 2003; Hoffman et al., 2013; Izmailov et al., 2020, 2021).

In the last decade, many research efforts have been made to adapt serial workhorses of Bayesian computational methods such as variational inference, expectation-propagation, and Markov chain Monte Carlo (MCMC) algorithms to massively distributed architectures (Wang and Dunson, 2013; Ahn et al., 2014; Wang et al., 2015; Hasenclever et al., 2017; Bui et al., 2018; Jordan et al., 2019; Rendell et al., 2021; Vono et al., 2022). Since the main bottleneck in distributed computing is the communication overhead, these approaches mainly focus on deriving efficient algorithms specifically designed to meet such a constraint, requiring only periodic or few rounds of communication between a central server and clients; see Plassier et al. (2021, Section 4) for a recent overview. As highlighted in Table 1, most current Bayesian FL methods adapt these approaches and focus almost exclusively on Federated Averaging type updates (McMahan et al., 2017), performing multiple local steps on each client. This is in contrast with predictive FL algorithms (which are **not** estimating predictive uncertainty), for which a variety of schemes have been explored, *e.g.* via gradient compression or client subsampling (Wang et al., 2021, Section 3.1.2). Moreover, very few Bayesian FL works have attempted to address the challenges raised by partial device participation or the impact of statistical heterogeneity; see Liu and Simeone (2021b); Chen and Chao (2021). Convergence results in Bayesian FL lag far behind "canonical" FL.

In this paper, we attempt to fill this gap, by proposing novel MCMC methods that extend Stochastic Langevin Dynamics to the FL context. It is assumed that the clients' data are independent and that the global posterior density is therefore the product of the *non-identical* local posterior densities of each client. To meet the specificity of Bayesian FL, each iteration of the proposed approaches only requires that a subset of active clients compute a stochastic gradient oracle for their associated negative log posterior density and send a lossy compression of these stochastic gradient oracles to the central server. The first scheme we derive, referred to as *Quantised Langevin Stochastic Dynamics* (`QLSD`), can interestingly be seen as the MCMC counterpart of the `QSGD` approach in FL (Alistarh et al., 2017), just as the Stochastic Gradient Langevin Dynamics (`SGLD`) (Welling and Teh, 2011) extends the Stochastic Gradient Descent (`SGD`). However, `QLSD` has the same drawbacks as `SGLD`: in particular, the invariant distribution of `QLSD` may deviate from the target distribution and become similar to the invariant measure of `SGD` when the number of observations is large (Brosse et al., 2018). We overcome this problem by deriving two variance-reduced versions $\texttt{QLSD}^\star$ and $\texttt{QLSD}^{++}$ that both include control variates.

Contributions (1) We propose a general MCMC algorithm called `QLSD` specifically designed for Bayesian inference under the FL paradigm and two variance-reduced alternatives, especially tackling *heterogeneity*, *communication overhead* and *partial participation*. (2) We provide a non-asymptotic convergence analysis of the proposed algorithms. The theoretical analysis highlights the impact of statistical heterogeneity measured by the discrepancy between local posterior distributions. (3) We propose efficient mechanisms to mitigate the impact of statistical heterogeneity on

convergence, either by using biased stochastic gradients or by introducing a *memory* mechanism that extends Horváth et al. (2019) to the Bayesian setting. In particular, we find that variance reduction indeed allows the proposed MCMC algorithm to converge towards the desired target posterior distribution when the number of observations becomes large. (4) We illustrate the advantages of the proposed methods using several FL benchmarks. We show that the proposed methodology performs well compared to state-of-the-art Bayesian FL methods.

Notations and Conventions The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\cdot\|$ and we set $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. For $n \in \mathbb{N}^*$, we refer to $\{1, \ldots, n\}$ with the notation $[n]$. For $N \in \mathbb{N}^*$, we use $\wp_N$ to denote the power set of $[N]$ and define $\wp_{N,n} = \{x \in \wp_N : \mathrm{card}(x) = n\}$ for any $n \in [N]$. We denote by $\mathrm{N}(m, \Sigma)$ the Gaussian distribution with mean vector m and covariance matrix Σ. We define the sign function, for any $x \in \mathbb{R}$, as $\mathrm{sign}(x) = \mathbf{1}\{x \geq 0\} - \mathbf{1}\{x < 0\}$. We define the Wasserstein distance of order 2 for any probability measures μ, ν on $\mathbb{R}^d$ with finite 2-moment by $W_2(\mu, \nu) = (\inf_{\zeta \in \mathcal{T}(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\theta - \theta'\|^2 \mathrm{d}\zeta(\theta, \theta'))^{1/2}$, where $\mathcal{T}(\mu, \nu)$ is the set of transference plans of μ and ν.

2 QUANTISED LANGEVIN STOCHASTIC DYNAMICS

In this section, we present the Bayesian FL framework and introduce the proposed methodology called `QLSD` along with two variance-reduced instances.

Problem Statement We are interested in performing Bayesian inference on a parameter $\theta \in \mathbb{R}^d$ based on a training dataset D. We assume that the posterior distribution admits a product-form density with respect to the d-dimensional Lebesgue measure, *i.e.*

$$\pi(\theta \mid \mathrm{D}) = \mathrm{Z}_\pi^{-1} \prod_{i=1}^b \mathrm{e}^{-U_i(\theta)}, \tag{1}$$

where $b \in \mathbb{N}^*$ and $\mathrm{Z}_\pi = \int_{\mathbb{R}^d} \prod_{i=1}^b \mathrm{e}^{-U_i(\theta)} \mathrm{d}\theta$ is a normalisation constant. This framework naturally encompasses the considered Bayesian FL problem. In this context, $\{\mathrm{e}^{-U_i}\}_{i \in [b]}$ stand for the unnormalised local posterior density functions associated to b clients, where each client $i \in [b]$ is assumed to own a local dataset D_i such that $\mathrm{D} = \sqcup_{i=1}^b \mathrm{D}_i$. The dependency of U_i on the local dataset D_i is omitted for brevity. A real-world illustration of the considered Bayesian problem is "multi-site fMRI classification" where each site (or client) owns a dataset coming from a local distribution because the methods of data generation and collection differ between sites. This results in different local likelihood functions, which combined with a local prior distribution, lead to heterogeneous local posteriors.

As in embarrassingly parallel MCMC approaches (Neiswanger et al., 2014), (1) implicitly assumes that the prior can be factorized across clients, which can always be done although the choice of this factorization is an open question. This product-form formulation can be alleviated by considering a global prior on θ and only calculating its gradient contribution on the central server during computations, see Algorithm 1.

A popular approach to sample from a target distribution with density π defined in (1) is based on Langevin dynamics with stochastic gradient which, starting from an initial point θ_0, defines a Markov chain $(\theta_k)_{k \in \mathbb{N}}$ by recursion:

$$\theta_{k+1} = \theta_k - \gamma H_{k+1}(\theta_k) + \sqrt{2\gamma} Z_{k+1}, \quad k \in \mathbb{N}, \tag{2}$$

where $\gamma \in (0, \bar{\gamma}]$, for some $\bar{\gamma} > 0$, is a discretisation time step, $(Z_k)_{k \in \mathbb{N}^*}$ is a sequence of i.i.d. standard Gaussian random variables and $(H_k)_{k \in \mathbb{N}^*}$ stand for unbiased estimators of ∇U with $U = \sum_{i=1}^b U_i$ (Parisi, 1981; Grenander and Miller, 1994; Roberts and Tweedie, 1996). In a serial setting involving a single client which owns a dataset of size $N \in \mathbb{N}^*$, the potential U writes $U = U_1 = \sum_{j=1}^N U_{1,j}$ for some functions $U_{1,j} : \mathbb{R}^d \to \mathbb{R}$, and a popular instance of this framework is `SGLD` (Welling and Teh, 2011). This algorithm consists in the recursion (2) with the specific choice $H_{k+1}(\theta) = (N/n) \sum_{j \in \mathcal{S}_{k+1}} \nabla U_{1,j}(\theta)$, where $(\mathcal{S}_k)_{k \in \mathbb{N}^*}$ is a sequence of i.i.d. uniform random subsets of $[N]$ of cardinal n.

In the FL framework, we assume that at each iteration k, the i-th client has access to an oracle $H_{k+1}^{(i)}$ based on its local negative log posterior density U_i, depending only on D_i, so that $H_{k+1} = \sum_{i=1}^b H_{k+1}^{(i)}$ is a stochastic gradient oracle of U. Note that we do not assume that $H_{k+1}^{(i)}$ is an unbiased estimator of ∇U_i, but only assume that H_{k+1} is unbiased. This allows us to consider biased local stochastic gradient oracles with better convergence guarantees, see Section 3 for more details. A simple adaptation of `SGLD` to the FL framework under consideration is given by recursion:

$$\theta_{k+1} = \theta_k - \gamma \sum_{i=1}^b H_{k+1}^{(i)}(\theta_k) + \sqrt{2\gamma} Z_{k+1}, k \in \mathbb{N}. \tag{3}$$

If for any $i \in [b]$, every potential function U_i also admits a finite-sum expression *i.e.* $U_i = \sum_{j=1}^{N_i} U_{i,j}$, similar to `SGLD`, we can for example use the local stochastic gradient oracles $H_{k+1}^{(i)}(\theta) = (N_i/n_i) \sum_{j \in \mathcal{S}_{k+1}^{(i)}} \nabla U_{i,j}(\theta)$, where $(\mathcal{S}_{k+1}^{(i)})_{k \in \mathbb{N}^*, i \in [b]}$ stand for i.i.d. uniform random subsets of $[N_i]$ of cardinal n_i. However, considering the MCMC algorithm associated with the recursion (3) is not adapted to the FL context. Indeed, this algorithm

would assume that each client is reliable and suffers from the same issues as SGD in a risk-based minimisation context, especially a prohibitive communication overhead (Girgis et al., 2020).

Proposed Methodology To address this problem, we propose to both account for the *partial participation of clients* and *reduce the number of bits transmitted* during the upload period by performing a lossy compression of a subset of $\{H_{k+1}^{(i)}\}_{i\in[b],k\in\mathbb{N}^*}$. This method has been used extensively in the "canonical" FL literature (Alistarh et al., 2017; Lin et al., 2018; Haddadpour et al., 2020; Sattler et al., 2020), but interestingly has never been considered in Bayesian FL; see Table 1.

To this end, we introduce a compression operator $\mathscr{C} : \mathbb{R}^d \to \mathbb{R}^d$ that is unbiased, *i.e.* for any $v \in \mathbb{R}^d$, $\mathbb{E}[\mathscr{C}(v)] = v$. In recent years, numerous compression operators have been proposed (Seide et al., 2014; Aji and Heafield, 2017; Stich et al., 2018). For example, the QSGD approach proposed in Alistarh et al. (2017) is based on stochastic quantisation.

QSGD considers for $\mathscr{C}$ a component-wise quantisation operator parameterised by a number of quantisation levels $s \geq 1$, which for each $j \in [d]$ and $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$ are given by

$$\mathscr{C}^{(s,j)}(v) = \frac{\|v\| \operatorname{sign}(v_j)}{s} \left(l_j + \mathbf{1}\left\{ \xi_j \leq \frac{s|v_j|}{\|v\|} - l_j \right\} \right), \quad (4)$$

where $l_j = \lfloor s|v_j|/\|v\| \rfloor$ and $\{\xi_j\}_{j\in[d]}$ is a sequence of i.i.d. uniform random variables on $[0, 1]$. In this particular case, we will denote the quantisation of v via (4) by $\mathscr{C}^{(s)}(v) = \{\mathscr{C}^{(s,j)}(v)\}_{j\in[d]}$.

The proposed general methodology, called *Quantised Langevin Stochastic Dynamics* (QLSD) stands for a compressed and FL version of the specific instance of SGLD defined in (3). More precisely, QLSD is an MCMC algorithm associated with the Markov chain $(\theta_k)_{k\in\mathbb{N}}$ starting from θ_0 and defined for $k \in \mathbb{N}$ as

$$\theta_{k+1} = \theta_k - \gamma \frac{b}{|\mathcal{A}_{k+1}|} \sum_{i\in\mathcal{A}_{k+1}} \mathscr{C}_{k+1}\left[H_{k+1}^{(i)}(\theta_k)\right] + \sqrt{2\gamma} Z_{k+1}\,,$$

where $(\mathcal{A}_k)_{k\in\mathbb{N}^*}$ denotes the subset of active (*i.e.* available) clients at iteration k, possibly random. Note that we indexed $\mathscr{C}$ by $k+1$ to emphasize that this compression operator is a stochastic operator and hence varies across iterations, see *e.g.* (4). The derivation of QLSD in the considered Bayesian FL context is described in details in Algorithm 1. A generalisation of QLSD taking into account *heterogeneous communication constraints* between clients by considering different compression operators $\{\mathscr{C}^{(i)}\}_{i\in[b]}$ is available in the Supplementary Material, see *e.g.* Section S1. In the particular case of the finite-sum setting where each client owns a dataset of size N_i, *i.e.* for the choice $H_{k+1}^{(i)}(\theta) = (N_i/n_i)\sum_{j\in\mathcal{S}_{k+1}^{(i)}} \nabla U_{i,j}(\theta)$ for $\theta \in \mathbb{R}^d$, $\mathcal{S}_{k+1}^{(i)} \in \wp_{N_i,n_i}$, we denote the corresponding instance of QLSD as QLSD$^\#$.

In this paper, we have decided to focus only on a non-adjusted sampling algorithm (QLSD) since the derivations of non-asymptotic results are already consequent, see the Supplementary Material. In addition, up to authors' knowledge, a general consensus on the choice between Metropolis-adjusted algorithms and their unadjusted counterparts has not been achieved yet.

Algorithm 1 Quantised Langevin Stochastic Dynamics (QLSD)

Input: nb. iterations K, compression operators $\{\mathscr{C}_{k+1}\}_{k\in\mathbb{N}}$, stochastic gradients $\{H_{k+1}^{(i)}\}_{i\in[b],k\in\mathbb{N}}$, step-size $\gamma \in (0, \bar{\gamma}]$ and initial point θ_0.
for $k = 0$ **to** $K-1$ **do**
  for $i \in \mathcal{A}_{k+1}$ // On active clients $\mathcal{A}_{k+1}$ **do**
    Compute $g_{i,k+1} = \mathscr{C}_{k+1}\left[H_{k+1}^{(i)}(\theta_k)\right]$.
    Send $g_{i,k+1}$ to the central server.
  end for
  // On the central server
  Compute $g_{k+1} = \frac{b}{|\mathcal{A}_{k+1}|}\sum_{i\in\mathcal{A}_{k+1}} g_{i,k+1}$.
  Draw $Z_{k+1} \sim \mathrm{N}(0_d, \mathrm{I}_d)$
  Compute $\theta_{k+1} = \theta_k - \gamma g_{k+1} + \sqrt{2\gamma} Z_{k+1}$.
  Send θ_{k+1} to the b clients.
end for
Output: samples $\{\theta_k\}_{k=0}^K$.

Variance-Reduced Alternatives Consider the finite-sum setting *i.e.* for any $i \in [b]$, $U_i = \sum_{j=1}^{N_i} U_{i,j}$ where N_i is the size of the local dataset D_i. As highlighted in Section 1, SGLD-based approaches, including Algorithm 1, involve an invariant distribution that may deviate from the target posterior distribution when $\min_{i\in b} N_i$ goes to infinity, as stochastic gradients with large variance are used (Brosse et al., 2018; Baker et al., 2019). We deal with this problem by proposing two variance-reduced alternatives of QLSD$^\#$ that use control variates. The simplest variance-reduced approach, referred to as QLSD* (see Algorithm S1) and discussed in more details in the Supplementary Material (see Section S2), considers a fixed-point approach that uses a minimiser $\theta^\star$ of the potential U (Brosse et al., 2018; Baker et al., 2019) defined as

$$\theta^\star \in \arg\min_{\theta\in\mathbb{R}^d} \sum_{i=1}^{b} U_i(\theta)\,. \quad (5)$$

In this scenario, the stochastic gradient oracles write for each $i \in [b]$, $k \in \mathbb{N}^*$, $\theta \in \mathbb{R}^d$ and $\mathcal{S}_{k+1}^{(i)} \in \wp_{N_i,n_i}$, $H_{k+1}^{(i)}(\theta) = (N_i/n_i) \sum_{j \in \mathcal{S}_{k+1}^{(i)}} [\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)]$. Although $\mathbb{E}[H_{k+1}] = \nabla U$, note that for each $i \in [b]$, $\mathbb{E}[H_{k+1}^{(i)}] \neq \nabla U_i$ so $H_{k+1}^{(i)}$ is not an unbiased estimate of U_i. We show in Section 3 that introducing this bias improves the convergence properties of QLSD$^\#$ with respect to the discrepancy between local posterior distributions. Since estimating $\theta^\star$ in a FL context might impose an additional computational burden on the sampling procedure, we propose another variance-reduced alternative referred to as QLSD^{++} (see Algorithm 2). This method builds on the Stochastic Variance Reduced Gradient (SVRG): it uses control variates $(\zeta_k)_{k\in\mathbb{N}}$ that are updated every $l \in \mathbb{N}^*$ iterations (Johnson and Zhang, 2013) and at each iteration $k \in \mathbb{N}$ and for any client $i \in [b]$, the stochastic gradient oracle $H_{k+1}^{(i)}$ defined by $H_{k+1}^{(i)}(\theta) = (N_i/n_i) \sum_{j \in \mathcal{S}_{k+1}^{(i)}} [\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\zeta_k)] + \nabla U_i(\zeta_k)$. To reduce the impact of local posterior discrepancy on convergence, we take inspiration from the "canonical" FL literature and consider a *memory term* $(\eta_k^{(i)})_{k\in\mathbb{N}}$ on each client $i \in [b]$ (Horváth et al., 2019; Dieuleveut et al., 2020). At each iteration k, instead of directly compressing $H_{k+1}^{(i)}$, we compress the difference $H_{k+1}^{(i)} - \eta_k^{(i)}$, store it in $g_{i,k+1}$, and then compute the global stochastic gradient $g_{k+1} = \frac{b}{|\mathcal{A}_{k+1}|} \sum_{i \in \mathcal{A}_{k+1}} g_{i,k+1} + \sum_{i=1}^b \eta_k^{(i)}$. The memory term $(\eta_k^{(i)})_{k\in\mathbb{N}}$ is then updated on each client $i \in [b]$, by the recursion $\eta_{k+1}^{(i)} = \eta_k^{(i)} + \alpha \mathbf{1}_{\mathcal{A}_{k+1}}(i) g_{i,k+1}$. The benefits of using this memory mechanism will be assessed theoretically in Section 3 and illustrated numerically in Section S5.2 in the Supplementary Material.

Algorithm 2 Variance-reduced Quantised Langevin Stochastic Dynamics (QLSD^{++})

Input: minibatch sizes $\{n_i\}_{i\in[b]}$, number of iterations K, compression operators $\{\mathscr{C}_{k+1}\}_{k\in\mathbb{N}^*}$, stepsize $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$, initial point θ_0 and $\alpha \in (0, \bar{\alpha}]$ with $\bar{\alpha} > 0$.
// Memory mechanism initialisation
Initialise $\{\eta_0^{(1)}, \ldots, \eta_0^{(b)}\}$ and $\eta_0 = \sum_{i=1}^b \eta_0^{(i)}$.
for $k = 0$ **to** $K - 1$ **do**
  // Update of the control variates
  if $k \equiv 0 \pmod{l}$ **then**
    Set $\zeta_k = \theta_k$.
  else
    Set $\zeta_k = \zeta_{k-1}$
  end if
  for $i \in \mathcal{A}_{k+1}$ // On active clients **do**
    Draw $\mathcal{S}_{k+1}^{(i)} \sim$ Uniform $(\wp_{N_i,n_i})$.
    Set $H_{k+1}^{(i)}(\theta_k) = (N_i/n_i) \sum_{j \in \mathcal{S}_{k+1}^{(i)}} [\nabla U_{i,j}(\theta_k) - \nabla U_{i,j}(\zeta_k)] + \nabla U_i(\zeta_k)$.
    Compute $g_{i,k+1} = \mathscr{C}_{k+1}\left(H_{k+1}^{(i)}(\theta_k) - \eta_k^{(i)}\right)$.
    Send $g_{i,k+1}$ to the central server.
    Set $\eta_{k+1}^{(i)} = \eta_k^{(i)} + \alpha g_{i,k+1}$.
  end for
  // On the central server
  Compute $g_{k+1} = \eta_k + \frac{b}{|\mathcal{A}_{k+1}|} \sum_{i \in \mathcal{A}_{k+1}} g_{i,k+1}$.
  Set $\eta_{k+1} = \eta_k + \alpha \sum_{i \in \mathcal{A}_{k+1}}^b g_{i,k+1}$.
  Draw $Z_{k+1} \sim \mathrm{N}(0_d, \mathrm{I}_d)$.
  Compute $\theta_{k+1} = \theta_k - \gamma g_{k+1} + \sqrt{2\gamma} Z_{k+1}$.
  Send θ_{k+1} to the b clients.
end for
Output: samples $\{\theta_k\}_{k=0}^K$.

3 THEORETICAL ANALYSIS

This section provides a detailed theoretical analysis of the proposed methodology. In particular, we will show the *impact of using stochastic gradients*, *partial participation* and *compression* by deriving quantitative convergence bounds for QLSD, which is detailed in Algorithm 1. We then derive non-asymptotic convergence bounds for QLSD* and QLSD^{++}, and explicitly show that these variance-reduced algorithms indeed succeed in reducing both the variance caused by stochastic gradients and the effects of *local posterior discrepancy* in the bounds we obtain for QLSD$^\#$. We consider the following assumptions on the potential U.

H1. *For any $i \in [b]$, U_i is continuously differentiable. In addition, suppose that the following hold.*

(i) U is $\mathtt{m}$-strongly convex, i.e. *for any $\theta_1, \theta_2 \in \mathbb{R}^d$, $\langle \nabla U(\theta_1) - \nabla U(\theta_2), \theta_1 - \theta_2 \rangle \geq \mathtt{m} \|\theta_1 - \theta_2\|^2$.*
(ii) U is $\mathtt{L}$-Lipschitz, i.e. *for any $\theta_1, \theta_2 \in \mathbb{R}^d$, $\|\nabla U(\theta_1) - \nabla U(\theta_2)\| \leq \mathtt{L} \|\theta_1 - \theta_2\|$.*

Note that **H**1-(i) implies that U admits a unique minimiser denoted by $\theta^\star \in \mathbb{R}^d$.

The compression operators $\{\mathscr{C}_{k+1}\}_{k\in\mathbb{N}}$ are assumed to satisfy the following assumption.

H2. *The compression operators $\{\mathscr{C}_{k+1}\}_{k\in\mathbb{N}}$ 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 $k \in \mathbb{N}^*$, $v \in \mathbb{R}^d$, $\mathbb{E}[\|\mathscr{C}_k(v) - v\|^2] \leq \omega \|v\|^2$.

As an example, the assumption on the variance of the compression operator detailed in **H**2-(ii) is verified for the quantisation operator $\mathscr{C}^{(s)}$ defined in (4) with $\omega = \min(d/s^2, \sqrt{d}/s)$ (Alistarh et al., 2017, Lemma 3.1).

Non-Asymptotic Analysis for Algorithm 1 We consider the following assumptions on the stochastic gradient oracles used in QLSD.

H3. *The random fields* $\{H_{k+1}^{(i)} : \mathbb{R}^d \to \mathbb{R}^d\}_{i\in[b],k\in\mathbb{N}}$ *are independent and satisfy the following conditions.*

(i) For any $\theta \in \mathbb{R}^d$ *and* $k \in \mathbb{N}$, $\sum_{i=1}^b \mathbb{E}[H_{k+1}^{(i)}(\theta)] = \nabla U(\theta)$.
(ii) There exist $\{\mathtt{M}_i > 0\}_{i\in[b]}$, *such that for any* $i \in [b]$, $k \in \mathbb{N}$, $\theta_1, \theta_2 \in \mathbb{R}^d$, $\mathbb{E}\left[\left\|H_{k+1}^{(i)}(\theta_1) - H_{k+1}^{(i)}(\theta_2)\right\|^2\right] \leq \mathtt{M}_i \langle \theta_1 - \theta_2, \nabla U_i(\theta_1) - \nabla U_i(\theta_2)\rangle$.
(iii) There exist $\sigma_\star, \mathtt{B}^\star \in \mathbb{R}_+$ *such that for any* $\theta \in \mathbb{R}^d$, $k \in \mathbb{N}$, *we have* $\mathbb{E}\left[\left\|H_{k+1}^{(i)}(\theta^\star)\right\|\right]^2 \leq \mathtt{B}^\star/b$, *and* $\mathbb{E}\left[\left\|\sum_{i=1}^b H_{k+1}^{(i)}(\theta^\star)\right\|^2\right] \leq \sigma_\star^2$, *where* $\theta^\star$ *is defined in* (5).

We can notice that **H**3-(ii) implies that ∇U_i is $\mathtt{M}_i$-Lipschitz continuous since by the Cauchy-Schwarz inequality, for any $i \in [b]$ and any $\theta_1, \theta_2 \in \mathbb{R}^d$, $\|\nabla U_i(\theta_1) - \nabla U_i(\theta_2)\|^2 \leq \mathtt{M}_i \langle \theta_1 - \theta_2, \nabla U_i(\theta_1) - \nabla U_i(\theta_2)\rangle$. Conversely, in the finite-sum setting, **H**3-(ii) is satisfied by QLSD$^\#$ with $\mathtt{M}_i = N_i\bar{\mathtt{M}}$ if for any $i \in [b]$ and $j \in [N_i]$, $U_{i,j}$ is convex and $\nabla U_{i,j}$ is $\bar{\mathtt{M}}$-Lipschitz continuous, for $\bar{\mathtt{M}} \geq 0$ by Nesterov (2003, Theorem 2.1.5).

In addition, it is worth mentioning that the first inequality in **H**3-(iii) is also required for our derivation in the deterministic case where $H_{k+1}^{(i)} = \nabla U_i$ due to the compression operator. In this particular case, $\mathtt{B}^\star$ stands for an upper-bound on $\sum_{i=1}^b \|\nabla U_i(\theta^\star)\|^2$ and corresponds to some discrepancy between local posterior density functions meaning that $\nabla U_i \neq \nabla U$ for $i \in [b]$. This phenomenon, referred to as *data heterogeneity* in the risk-based literature (Horváth et al., 2019; Karimireddy et al., 2020), is ubiquitous in the FL context.

Finally, we assume for simplicity that *clients' partial participation* is realised by each client having probability $p \in (0, 1]$ of being active in each communication round.

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

A generalisation of this scheme considering different probabilities p_i per client can be found in the Supplementary Material, see *e.g.* Section S1.1. Under the above assumptions and by denoting Q_γ the Markov kernel associated to Algorithm 1, the following convergence result holds.

Theorem 1. *Assume* ***H1****,* ***H2****,* ***H3*** *and* ***H4****. Then, there exists* $\bar{\gamma}_\infty$ *such that for* $\bar{\gamma} < \bar{\gamma}_\infty$, *there exist* $A_{\bar{\gamma}}, B_{\bar{\gamma}} > 0$ *(explicitly given in Section S1 in the Supplementary Material) satisfying for any probability measure* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, *any step size* $\gamma \in (0, \bar{\gamma}]$ *and* $k \in \mathbb{N}$,

$$W_2^2\left(\mu Q_\gamma^k, \pi\right) \leq (1 - \gamma\mathtt{m}/2)^k \cdot W_2^2(\mu, \pi) + \gamma B_{\bar{\gamma}} + \gamma^2 A_{\bar{\gamma}} (1 - \mathtt{m}\gamma/2)^{k-1} k \cdot \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^2 \mu(\mathrm{d}\theta)\,,$$

where $\theta^\star$ *is defined in* (5).

Similar to ULA (Dalalyan, 2017; Durmus and Moulines, 2019) and SGLD (Dalalyan and Karagulyan, 2019; Durmus et al., 2019), the upper bound given in Theorem 1 includes a contracting term that depends on the initialisation and a bias term $\gamma B_{\bar{\gamma}}$ that does not vanish with $k \to \infty$ due to the use of a fixed step size γ. In the asymptotic scenario, *i.e.* $\bar{\gamma} \downarrow 0$, Table 1 gives the dependencies of $B_{\bar{\gamma}}$ for QLSD and its particular instance QLSD$^\#$, in terms of key quantities associated with the setting we consider. Similar to SGLD, we can observe that the use of stochastic gradients entails a bias term of order $\sigma_\star^2\,\mathrm{O}(\gamma)$. On the other hand, the use of partial participation and compression compared to SGLD introduces an *additional bias* of order $(\omega/p)(\mathtt{m}\mathtt{B}^\star + \mathtt{L}\mathtt{M}d)\,\mathrm{O}(\gamma)$, which grows with in particular $\mathtt{B}^\star$, corresponding to the impact of the local posterior discrepancy on convergence.

Non-Asymptotic Analysis for Variance-Reduced Alternatives We assume in the sequel that the potential functions $\{U_i\}_{i\in[b]}$ admit the finite-sum decomposition $U_i = \sum_{j=1}^{N_i} U_{i,j}$ for each $i \in [b]$ and consider the following assumptions.

H 5. *For any* $i \in [b], j \in [N_i]$, $U_{i,j}$ *is continuously differentiable and the following holds.*

(i) There exists $\mathtt{M}_i > 0$ *such that, for any* $\theta_1, \theta_2 \in \mathbb{R}^d$, $\|\nabla U_i(\theta_2) - \nabla U_i(\theta_1)\|^2 \leq \mathtt{M}\langle\theta_2 - \theta_1, \nabla U_i(\theta_2) - \nabla U_i(\theta_1)\rangle$.
(ii) There exists $\bar{\mathtt{M}} \geq 0$ *such that, for any* $\theta_1, \theta_2 \in \mathbb{R}^d$, $\|\nabla U_{i,j}(\theta_2) - \nabla U_{i,j}(\theta_1)\|^2 \leq \bar{\mathtt{M}}\langle\nabla U_{i,j}(\theta_2) - \nabla U_{i,j}(\theta_1), \theta_2 - \theta_1\rangle$.

As mentioned earlier, **H**5 is satisfied if for every $i \in [b]$ and $j \in [N_i]$, $U_{i,j}$ is convex and $\nabla U_{i,j}$ is $\bar{\mathtt{M}}$-Lipschitz continuous. Under these additional conditions, the following non-asymptotic convergence results hold for the two reduced-variance MCMC algorithms described in Section 2. Denote by $Q_{\circledast,\gamma}$ the Markov kernel associated to QLSD* with a step size $\gamma \in (0, \bar{\gamma}]$.

Theorem 2. *Assume* ***H1****,* ***H2****,* ***H4*** *and* ***H5****. Then, there exists* $\bar{\gamma}_{\circledast,\infty}$ *such that for* $\bar{\gamma} < \bar{\gamma}_{\circledast,\infty}$, *there exist* $A_{\circledast,\bar{\gamma}}, B_{\circledast,\bar{\gamma}} > 0$ *(explicitly given in Section S2 in the Supplementary Material) satisfying for any probability measure* $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, *any step size* $\gamma \in (0, \bar{\gamma}]$ *and* $k \in \mathbb{N}$,

Table 2: Order of the asymptotic biases $\{B_{\bar\gamma}, B_{\circledast,\bar\gamma}, B_{\oplus,\bar\gamma}\}$, associated to the three proposed MCMC algorithms, in squared 2-Wasserstein distance for two types of asymptotic. Red dependencies prevent from (quick) convergence while green dependencies ensure convergence of associated MCMC algorithms. $\theta^\star$ is defined in (5).

Algo.	Bias	Dependencies of the asymptotic bias when $\bar\gamma \downarrow 0$					Dependencies of the asymptotic bias as $N_i \to \infty$
		d	$H_{k+1}^{(i)}$	$\mathtt{B}^\star$	partial particip.	ω	
QLSD	$B_{\bar\gamma}$	d	$\sigma_\star^2$	$\mathtt{B}^\star$	$(1-p)/p$	ω	$\mathrm{O}(N_i)$
QLSD$^\#$	$B_{\bar\gamma}$	d	N_i^2	$\sum_{i=1}^b \|\nabla U_i(\theta^\star)\|^2$	$(1-p)/p$	ω	$\mathrm{O}(N_i)$
QLSD*	$B_{\circledast,\bar\gamma}$	d	N_i	-	$(1-p)/p$	ω	$d\,\mathrm{O}(1)$
QLSD^{++}	$B_{\oplus,\bar\gamma}$	d	N_i	-	$(1-p)/p$	ω	$d\,\mathrm{O}(1)$

$$W_2^2\left(\mu Q_{\circledast,\gamma}^k, \pi\right) \le (1-\gamma\mathtt{m}/2)^k \cdot W_2^2(\mu,\pi) + \gamma B_{\circledast,\bar\gamma} + \gamma^2 A_{\circledast,\bar\gamma}(1-\mathtt{m}\gamma/2)^{k-1} k \cdot \int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2 \mu(\mathrm{d}\theta)\,,$$

where $\theta^\star$ is defined in (5).

Compared to QLSD and QLSD*, QLSD^{++} only defines an inhomogeneous Markov chain, see Section S3.3 in the Supplementary Material for more details. For a step-size $\gamma \in (0,\bar\gamma]$ and an iteration $k \in \mathbb{N}$, we denote by $\mu Q_{\oplus,\gamma}^{(k)}$ the distribution of θ_k defined by QLSD^{++} starting from θ_0 with distribution μ.

Theorem 3. *Assume **H1**, **H2**, **H4** and **H5**, and let $l \in \mathbb{N}^*$ and $\alpha \in (0, 1/(\omega+1)]$. Then, there exists $\bar\gamma_{\oplus,\infty}$ such that for $\bar\gamma < \bar\gamma_{\oplus,\infty}$, there exist $A_{\oplus,\bar\gamma}, B_{\oplus,\bar\gamma}, C_{\oplus,\bar\gamma} > 0$ (explicitly given in Section S3 in the Supplementary Material and independent of α) satisfying for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, any step size $\gamma \in (0,\bar\gamma]$ and $k \in \mathbb{N}$,*

$$W_2^2(\mu Q_{\oplus,\gamma}^{(k)}, \pi) \le (1-\gamma\mathtt{m}/2)^k \cdot W_2^2(\mu,\pi) + \gamma B_{\oplus,\bar\gamma} + \gamma^2 A_{\oplus,\bar\gamma}(1-\gamma\mathtt{m}/2)^{\lfloor k/l\rfloor} \cdot \int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2 \mu(\mathrm{d}\theta) + \gamma C_{\oplus,\bar\gamma}[(1-\alpha)^k \wedge (1-\gamma\mathtt{m}/2)^{\lfloor k/l\rfloor}] \sum_{i=1}^b \|\nabla U_i(\theta^\star)\|^2\,,$$

where $\theta^\star$ is defined in (5).

Table 2 provides the dependencies of the asymptotic bias terms $B_{\circledast,\bar\gamma}, B_{\oplus,\bar\gamma}$ as $\bar\gamma \downarrow 0$ with respect to key quantities associated to the problem we consider. For comparison, we do the same regarding the specific instance of Algorithm 1, QLSD$^\#$. Remarkably, thanks to biased local stochastic gradients for QLSD* and the memory mechanism for QLSD^{++}, we can notice that their associated asymptotic biases do not depend on local posterior discrepancy in contrast to QLSD$^\#$. This is in line with non-asymptotic convergence results in risk-based FL which also show that the impact of data heterogeneity can be alleviated using such a memory mechanism (Philippenko and Dieuleveut, 2020). The impact of stochastic gradients is discussed in further details in the next paragraph.

Consistency Analysis in the Big Data Regime In Brosse et al. (2018), it was shown that ULA and SGLD define homogeneous Markov chains, each of which admits a unique stationary distribution. However, while the invariant distribution of ULA gets closer to π as N_i increases, conversely the invariant measure of SGLD never approaches π and is in fact very similar to the invariant measure of SGD. Moreover, the non-compressed counterpart of QLSD* has been shown not to suffer from this problem, and it has been theoretically proven to be a viable alternative to ULA in the Big Data environment. Since QLSD is a generalisation of SGLD, the conclusions of Brosse et al. (2018) hold. On the other hand, we show that the reduced-variance alternatives to QLSD that we introduced provide more accurate estimates of π as N_i increases, see the last column in Table 2. Detailed calculations are deferred to Section S4 in the Supplementary Material.

4 NUMERICAL EXPERIMENTS

This section illustrates our methodology with three numerical experiments that include both synthetic and real datasets. For all experiments, we consider the finite-sum setting and use the stochastic quantisation operator $\mathscr{C}^{(s)}$ for $s \ge 1$ defined in (4) to perform the compression step. In this case **H2**-(ii) is verified with $\omega = \min(d/s^2, \sqrt{d}/s)$. Further experimental results are given in Section S5 in the Supplementary Material.

Toy Gaussian Example This first experiment aims at illustrating the general behavior of Algorithm 1 with respect to the use of stochastic gradients and compression scheme. To this purpose, we set $b = 20$ and $d = 50$ and consider a Gaussian posterior distribution with density defined in (1) where, for any $i \in [b]$ and $\theta \in \mathbb{R}^d$, $U_i(\theta) = \sum_{j=1}^{N_i} \|\theta - y_{i,j}\|^2/2$, $\{y_{i,j}\}_{i\in[b],j\in[N_i]}$ being a set of synthetic independent but not identically distributed observations across clients and $N_i \in [10, 200]$, see Figure 1 (top row). Note that in this specific case, $\theta^\star$ admits a closed form expression. For all the

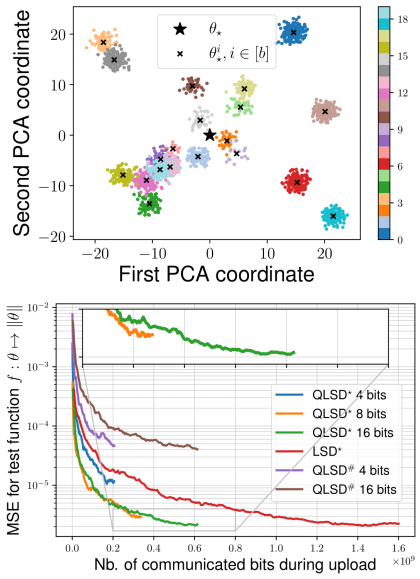

Figure 1: Toy Gaussian example. (top) 2D projection of the heterogeneous synthetic dataset where each color refers to a client and each dot is an observation $y_{i,j}$. (bottom) Estimation performances of the considered Bayesian FL algorithms.

algorithms, we choose the (optimised) step-size $\gamma = 4.9 \times 10^{-4}$ and choose a minibatch size $n_i = \lfloor N_i/10 \rfloor$. Instances of QLSD$^{\#}$ and QLSD* using $s = 2^p$ are referred to as p-bits instances of these MCMC algorithms. We compare these algorithms with the non-compressed counterpart of QLSD* referred to as LSD*, see Algorithm S2. Figure 1 shows the behavior of the mean squarred error (MSE) associated to the test function $f : \theta \mapsto \|\theta\|$, computed using 30 independent runs of each algorithm, with respect to the number of bits transmitted. We can notice that QLSD* always outperforms QLSD$^{\#}$and that decreasing the value of ω does not significantly reduce the bias associated to QLSD*. This illustrates the impact of the variance of the stochastic gradients and supports our theoretical analysis summarised in Table 2. On the other hand, QLSD* with $s = 2^{16}$ achieves a similar MSE as LSD* while requiring roughly 2.5 times less number of bits.

Bayesian Logistic Regression In this experiment, we compare the proposed methodology based on gradient compression with two existing FedAvg-type MCMC algorithms. Since $\theta^{\star}$ defined in (5) is not easily available, we implement QLSD^{++} detailed in Algorithm 2. We adopt a zero-mean Gaussian prior with covariance matrix $2 \cdot 10^{-2} \mathrm{I}_d$ and use the FEMNIST dataset (Caldas et al., 2018). We set $b = 50$, $l = 100$, $\alpha = 1/(\omega + 1)$ and $\gamma = 10^{-5}$. We launch QLSD^{++} for $s \in \{2^4, 2^8, 2^{16}\}$ and compare its performances with DG-SGLD (Plassier et al., 2021) and FSGLD (El Mekkaoui et al., 2021) which use multiple local steps to address the communication bottleneck. We are interested in performing uncertainty quantification by estimating highest posterior density (HPD) regions. For any $\alpha \in (0, 1)$, we define $\mathcal{C}_\alpha = \{\theta \in \mathbb{R}^d; -\log \pi(\theta|\mathrm{D}) \leq \eta_\alpha\}$ where $\eta_\alpha \in \mathbb{R}$ is chosen such that $\int_{\mathcal{C}_\alpha} \pi(\theta|\mathrm{D})\mathrm{d}\theta = 1 - \alpha$. We compute the relative HPD error based on the scalar summary η_α, *i.e.* $|\eta_\alpha - \eta_\alpha^{\mathrm{LSD}}|/\eta_\alpha^{\mathrm{LSD}}$ where $\eta_\alpha^{\mathrm{LSD}}$ has been estimated using the non-compressed counterpart of QLSD^{++}, referred to as LSD^{++} and standing for a serial variance-reduced SGLD, see Algorithm S3. Table 3 gives this relative HPD error for $\alpha = 0.01$ and provides the relative efficiency of QLSD^{++} and competitors corresponding to the savings in terms of transmitted bits per iteration. One can notice that the proposed approach provides similar results as its non-compressed counterpart while being 3 to 7 times more efficient. In addition, we show that QLSD^{++} provides similar performances as DG-SGLD and FSGLD which highlight that gradient compression and periodic communication are competing approaches.

Table 3: Bayesian Logistic Regression.

Algorithm	99% HPD error	Rel. efficiency
FSGLD	5.4e-3	6.2
DG-SGLD	5.2e-3	6.4
QLSD^{++} 4 bits	6.1e-3	7.6
QLSD^{++} 8 bits	4.3e-3	6.7
QLSD^{++} 16 bits	6.9e-4	3.1

Bayesian Neural Networks In our third experiment, we go beyond the scope of our theoretical analysis by performing posterior inference in Bayesian neural networks. We use the ResNet-20 model (He et al., 2016), choose a zero-mean Gaussian prior distribution with variance 1/5 and consider the classification problem associated with the CIFAR-10 dataset (Krizhevsky et al., 2009). We run QLSD^{++} with $s = 2$, $l = 20$, $\alpha = 1/(\omega + 1)$, and with either $p = 1$ (full participation) or $p = 0.25$ (partial participation). We compare the proposed methodology with a long-run Hamiltonian Monte Carlo (HMC) considered as a "ground truth" (Izmailov et al., 2021) and SGLD. For completeness, we also implement four other distributed/federated approximate sampling approaches, namely two instances of FedBe (Chen and Chao, 2021), DG-SGLD and FSGLD. Following Wilson et al. (2021), we compare the aforementioned algorithms through three metrics: classification *accuracy* on the test dataset using the minimum mean-square estimator, *agreement* between the top-1 prediction given by each algorithm and the one given by HMC and *total variation* between approximate and "true" (associated with HMC) predictive distributions. More details about algorithms' hy-

Table 4: Performances of Bayesian FL algorithms on the considered Bayesian neural networks problem.

Method	HMC	SGLD	QLSD^{++}	QLSD^{++} PP	FedBe-Dirichlet	FedBe-Gauss.	DG-SGLD	FSGLD
Accuracy	89.6	88.8	88.1	86.6	90.7	90.2	92.2	87.5
Agreement	0.94	0.91	0.90	0.90	0.90	0.89	0.91	0.91
TV	0.07	0.11	0.12	0.12	0.16	0.16	0.13	0.13

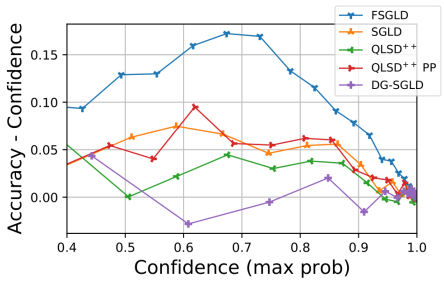

Figure 2: Bayesian Neural Networks.

perparameters and considered metrics are given in Section S5.3 in the Supplementary Material. The results we obtain are gathered in Table 4. In terms of agreement and total variation, QLSD^{++} (even with partial participation) gives similar results as SGLD and competes favorably with other existing federated approaches. Figure 2 complements this empirical analysis by showing calibration curves of posterior predictive distributions.

5 CONCLUSION

In this paper, we presented a general methodology based on Langevin stochastic dynamics for Bayesian FL. In particular, we addressed the challenges associated with this new ML paradigm by assuming that a subset of clients sends compressed versions of its local stochastic gradient oracles to the central server. Moreover, the proposed method was found to have favorable convergence properties, as evidenced by numerical illustrations. In particular, it compares favorably to FedAvg-type Bayesian FL algorithms. A limitation of this work is that the proposed method does not target the initial posterior distribution due to the use of a fixed discretisation time step. Therefore, this work paves the way for more advanced Bayesian FL approaches based, for example, on Metropolis-Hastings schemes to remove asymptotic biases. In addition, although the data ownership issue is implicitly tackled by the FL paradigm by not sharing data, stronger privacy guarantees can be ensured, typically by combining differential privacy, secure multi-party computation and homomorphic encryption methods. Proposing a differentially private version of our methodology is a possible extension of our work, that is left for further work. This work has no direct societal impact.

Acknowledgements

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

References

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

Alham Fikri Aji and Kenneth Heafield. Sparse Communication for Distributed Gradient Descent. In *Conference on Empirical Methods in Natural Language Processing*, 2017.

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.

Christophe Andrieu, Nando de Freitas, Arnaud Doucet, and Michael I. Jordan. An introduction to MCMC for machine learning. *Machine Learning*, 50 (1–2):5–43, 2003.

Jack Baker, Paul Fearnhead, Emily B. Fox, and Christopher Nemeth. Control variates for stochastic gradient MCMC. *Statistics and Computing*, 29(3): 599–615, 2019.

Nicolas Brosse, Alain Durmus, and Eric Moulines. The promises and pitfalls of Stochastic Gradient Langevin Dynamics. In *Advances in Neural Information Processing Systems*, 2018.

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.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konecny, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A Benchmark for Federated Settings. *arXiv preprint arXiv:1812.01097*, 2018.

Hong-You Chen and Wei-Lun Chao. FedBE: Making Bayesian Model Ensemble Applicable to Federated

Learning. In *International Conference on Learning Representations*, 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.

Arnak S. Dalalyan and Avetik Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and Their Applications*, 129(12):5278–5311, 2019.

Alexander Philip Dawid and Monica Musio. Theory and applications of proper scoring rules. *Metron*, 72 (2):169–183, 2014.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Annals of Statistics*, 48(3):1348–1382, 06 2020.

Alain Durmus and Eric Moulines. High-dimensional Bayesian inference via the unadjusted Langevin algorithm. *Bernoulli*, 25(4A):2854–2882, 2019.

Alain Durmus, Szymon Majewski, and Blażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

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.

Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Severine Dubuisson, and Isabelle Bloch. Encoding the latent posterior of Bayesian Neural Networks for uncertainty quantification. *arXiv preprint arXiv:2012.02818*, 2020.

Antonious M Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of federated learning: Privacy, communication and accuracy trade-offs. *arXiv preprint arXiv:2008.07180*, 2020.

Ulf Grenander and Michael I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society, Series B*, 56(4):549–603, 1994.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated Learning with Compression: Unified Analysis and Sharp Guarantees. *arXiv preprint arXiv:2007.01154*, 2020.

Leonard Hasenclever, Stefan Webb, Thibaut Lienart, Sebastian Vollmer, Balaji Lakshminarayanan, Charles Blundell, and Yee Whye Teh. Distributed Bayesian Learning with Stochastic Natural Gradient Expectation Propagation and the Posterior Server. *Journal of Machine Learning Research*, 18(106):1–37, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14(4):1303–1347, 2013.

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.

David J. Hunter. Uncertainty in the Era of Precision Medicine. *New England Journal of Medicine*, 375 (8):711–713, 2016.

Pavel Izmailov, Wesley J Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for Bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pages 1169–1179, 2020.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? *arXiv preprint arXiv:2104.14421*, 2021.

Rie Johnson and Tong Zhang. Accelerating Stochastic Gradient Descent Using Predictive Variance Reduction. In *Advances in Neural Information Processing Systems*, page 315–323, 2013.

Michael I. Jordan, Jason D. Lee, and Yun Yang. Communication-Efficient Distributed Statistical Inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, K. A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G.L. D'Oliveira, 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 Konevcný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, 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.

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*, 2020.

Rahif Kassab and Osvaldo Simeone. Federated Generalized Bayesian Learning via Distributed Stein Variational Gradient Descent. *arXiv preprint arXiv:2009.06419*, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. Available at https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

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

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. In *Machine Learning and Systems*, 2020.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and Bill Dally. Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training. In *International Conference on Learning Representations*, 2018.

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.

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

Willie Neiswanger, Chong Wang, and Eric P. Xing. Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014.

Christopher Nemeth and Chris Sherlock. Merging MCMC Subposteriors through Gaussian-Process Approximations. *Bayesian Analysis*, 13(2):507–530, 06 2018.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science, 2003.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua V Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *arXiv preprint arXiv:1906.02530*, 2019.

G. Parisi. Correlation functions and computer simulations. *Nuclear Physics B*, 180(3):378–384, 1981.

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.

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*, 2021.

Rahul Rahaman and Alexandre H Thiery. Uncertainty quantification and deep ensembles. *arXiv preprint arXiv:2007.08792*, 2020.

L. J. Rendell, A. M. Johansen, A. Lee, and N. Whiteley. Global consensus Monte Carlo. *Journal of Computational and Graphical Statistics*, 30(2):249–259, 2021.

Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science, 2013.

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

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, Berlin, 2 edition, 2004.

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

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and Communication-Efficient Federated Learning From Non-i.i.d. Data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3413, 2020.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-Bit Stochastic Gradient Descent and Application to Data-Parallel Distributed Training of Speech DNNs. In *Interspeech*, 2014.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with Memory. In *Advances in Neural Information Processing Systems*, 2018.

Cedric Villani. *Optimal Transport: Old and New*. Springer Berlin Heidelberg, 2008.

Maxime Vono, Daniel Paulin, and Arnaud Doucet. Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting. *Journal of Machine Learning Research*, 23(25):1–69, 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.

Xiangyu Wang and David B. Dunson. Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Xiangyu Wang, Fangjian Guo, Katherine A. Heller, and David B. Dunson. Parallelizing MCMC with random partition trees. In *Advances in Neural Information Processing Systems*, 2015.

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

Andrew Gordon Wilson, Pavel Izmailov, Matthew D Hoffman, Yarin Gal, Yingzhen Li, Melanie F Pradier, Sharad Vikram, Andrew Foong, Sanae Lotfi, and Sebastian Farquhar. Evaluating Approximate Inference in Bayesian Deep Learning. 2021. Available at https://izmailovpavel.github.io/neurips_bdl_competition/files/BDL_NeurIPS_Competition.pdf.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Supplementary Material: QLSD: Quantised Langevin Stochastic Dynamics for Bayesian Federated Learning

Notations and conventions. We denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$, $\mathbb{M}(\mathbb{R}^d)$ the set of all Borel measurable functions f on $\mathbb{R}^d$ and $\|\cdot\|$ the Euclidean norm on $\mathbb{R}^d$. For μ a probability measure on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ and $f \in \mathbb{M}(\mathbb{R}^d)$ a μ-integrable function, denote by $\mu(f)$ the integral of f with respect to (w.r.t.) μ. Let μ and ν be two sigma-finite measures on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. Denote by $\mu \ll \nu$ if μ is absolutely continuous w.r.t. ν and $\mathrm{d}\mu/\mathrm{d}\nu$ the associated density. We say that ζ is a transference plan of μ and ν if it is a probability measure on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that for all measurable set A of $\mathbb{R}^d$, $\zeta(\mathsf{A} \times \mathbb{R}^d) = \mu(\mathsf{A})$ and $\zeta(\mathbb{R}^d \times \mathsf{A}) = \nu(\mathsf{A})$. We denote by $\mathcal{T}(\mu, \nu)$ the set of transference plans of μ and ν. In addition, we say that a couple of $\mathbb{R}^d$-random variables (X, Y) is a coupling of μ and ν if there exists $\zeta \in \mathcal{T}(\mu, \nu)$ such that (X, Y) are distributed according to ζ. We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures with finite 2-moment: for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $\int_{\mathbb{R}^d} \|x\|^2 \mathrm{d}\mu(x) < \infty$. We define the squared Wasserstein distance of order 2 associated with $\|\cdot\|$ for any probability measures $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$W_2^2(\mu, \nu) = \inf_{\zeta \in \mathcal{T}(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \mathrm{d}\zeta(x, y) \,.$$

By Villani (2008, Theorem 4.1), for all μ, ν probability measures on $\mathbb{R}^d$, there exists a transference plan $\zeta^\star \in \mathcal{T}(\mu, \nu)$ such that for any coupling (X, Y) distributed according to $\zeta^\star$, $W_2(\mu, \nu) = \mathbb{E}[\|x - y\|^2]^{1/2}$. This kind of transference plan (respectively coupling) will be called an optimal transference plan (respectively optimal coupling) associated with W_2. By Villani (2008, Theorem 6.16), $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the Wasserstein distance W_2 is a complete separable metric space. 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 $[n]$ and $\wp_n$ the power set of $[n]$. 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 $\mathrm{N}(\mu, \Sigma)$.

S1 PROOF OF Theorem 1

This section aims at proving Theorem 1 in the main paper.

S1.1 Generalised quantised Langevin stochastic dynamics

We show that `QLSD` defined in Algorithm 1 in the main paper can be cast into a more general framework that we refer to as generalised quantised Langevin stochastic dynamics. Then, the guarantees for `QLSD` will be a simple consequence of the ones that we will establish for generalised `QLSD`. For ease of reading, we recall first the setting and the assumptions that we consider all along the paper. Recall that the dataset D is assumed to be partitioned into b *shards* $\{\mathrm{D}_i\}_{i=1}^b$ such that $\sqcup_{i=1}^b \mathrm{D}_i = \mathrm{D}$ and the posterior distribution of interest is assumed to admit a density with respect to the d-dimensional Lebesgue measure which factorises across clients, i.e. for any $\theta \in \mathbb{R}^d$,

$$\pi(\theta) = \exp\{-U(\theta)\} / \int_{\mathbb{R}^d} \mathrm{e}^{-U(\theta)} \,\mathrm{d}\theta \,, \qquad U(\theta) = \sum_{i=1}^b U_i(\theta) \,.$$

We consider the following assumptions on the potential U.

HS1. *For any $i \in [b]$, U_i is continuously differentiable. In addition, suppose that the following conditions hold.*

(i) U is $\mathtt{m}$-strongly convex, i.e. *for any $\theta_1, \theta_2 \in \mathbb{R}^d$,*

$$U(\theta_1) \geq U(\theta_2) + \langle \theta_1 - \theta_2, \nabla U(\theta_2) \rangle + \mathtt{m} \|\theta_1 - \theta_2\|^2 / 2 \,.$$

(ii) U is $\mathtt{L}$-Lipschitz, i.e. *for any $\theta_1, \theta_2 \in \mathbb{R}^d$,*

$$\|\nabla U(\theta_1) - \nabla U(\theta_2)\| \leq \mathtt{L} \|\theta_1 - \theta_2\| \,.$$

Note that **H**S1-(i) implies that U admits a unique minimiser denoted by $\theta^\star \in \mathbb{R}^d$. Moreover, for any $(\theta_1, \theta_1) \in \mathbb{R}^d$, **H**S1-(i)-(ii) combined with Nesterov (2003, Equation 2.1.24) shows that

$$\langle \nabla U(\theta_2) - \nabla U(\theta_1), \theta_2 - \theta_1 \rangle \geq \frac{\mathtt{m}\mathtt{L}}{\mathtt{m}+\mathtt{L}} \|\theta_2 - \theta_1\|^2 + \frac{1}{\mathtt{m}+\mathtt{L}} \|\nabla U(\theta_2) - \nabla U(\theta_1)\|^2 \,. \tag{S1}$$

We consider the following assumptions on the family $\{H_i : \mathbb{R}^d \times \mathsf{X}_1 \to \mathbb{R}^d\}_{i\in[b]}$ and $\mathscr{C}$.

HS2. *There exists a probability measure ν_2 on a measurable space $(\mathsf{X}_2, \mathcal{X}_2)$ and a family of measurable functions $\{\mathscr{C}_i : \mathbb{R}^d \times \mathsf{X}_2 \to \mathbb{R}^d\}_{i\in[b]}$ such that the following conditions hold.*

(i) For any $\theta \in \mathbb{R}^d$ and any $i \in [b]$, $\int_{\mathsf{X}_2} \mathscr{C}_i(\theta, x^{(2)})\, \nu_2(\mathrm{d}x^{(2)}) = \theta$.

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

$$\int_{\mathsf{X}_2} \left\| \mathscr{C}_i(\theta, x^{(2)}) - \theta \right\|^2 \nu_2(\mathrm{d}x^{(2)}) \leq \omega_i \|\theta\|^2 \,.$$

HS3. *There exist a family of probability measures $\{\nu_1^{(i)}\}_{i\in[b]}$ defined on measurable spaces $\{(\mathsf{X}_1^{(i)}, \mathcal{X}_1^{(i)})\}_{i\in[b]}$ and a family of measurable functions $\{H_i : \mathbb{R}^d \times \mathsf{X}_1^{(i)} \to \mathbb{R}^d\}_{i\in[b]}$ such that the following conditions hold.*

(i) For any $\theta \in \mathbb{R}^d$,

$$\sum_{i=1}^b \int_{\mathsf{X}_1^{(i)}} H_i(\theta, x^{(1,i)}) \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) = \nabla U(\theta) \,.$$

(ii) There exist $\{\mathtt{M}_i > 0\}_{i\in[b]}$, such that for any $i \in [b]$, $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta_2, x^{(1,i)}) - H_i(\theta_1, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \leq \mathtt{M}_i \langle \theta_2 - \theta_1, \nabla U_i(\theta_2) - \nabla U_i(\theta_1) \rangle \,.$$

(iii) There exists $\sigma_\star, \mathtt{B}^\star \in \mathbb{R}_+$ such that for any $i \in [b]$, $\theta \in \mathbb{R}^d$, we have

$$\int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta^\star, x^{(1)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1)}) \leq \mathtt{B}^\star / b \,, \qquad \int_{\mathsf{X}_1^{(1)} \times \cdots \times \mathsf{X}_1^{(b)}} \left\| \sum_{i=1}^b H_i(\theta^\star, x^{(1,i)}) \right\|^2 \otimes_{i=1}^b \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \leq \sigma_\star^2 \,. \tag{S2}$$

We can notice that **H**S3-(ii) implies that ∇U_i is $\mathtt{M}_i$-Lipschitz continuous since by the Cauchy Schwarz inequality, for any $i \in [b]$ and any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\|\nabla U_i(\theta_1) - \nabla U_i(\theta_2)\|^2 \leq \mathtt{M}_i \langle \theta_1 - \theta_2, \nabla U_i(\theta_1) - \nabla U_i(\theta_2) \rangle \,.$$

In addition, it is worth mentioning that the first inequality in (S2) is also required for our derivation in the deterministic case where $H_i = \nabla U_i$ for any $i \in [b]$ due to the compression step. For $k \geq 1$, consider $(X_k^{(1,1)}, \ldots, X_k^{(1,b)})_{k\in\mathbb{N}}$ and $(X_k^{(2,1)}, \ldots, X_k^{(2,b)})_{k\in\mathbb{N}}$ two independent sequences of random variables distributed according to $\nu_1^{(1:b)} = \nu_1^{(1)} \otimes \cdots \otimes \nu_1^{(b)}$ and $\nu_2^{\otimes b}$, respectively.

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

H S4. *For any $k \in \mathbb{N}^*$, $\mathcal{A}_k = \{i \in [b] : B_{i,k} = 1\}$ 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 $p_i \in (0, 1]$.*

In other words, there exists a sequence $(X_k^{(3,1)}, \cdots, X_k^{(3,b)})_{k\in\mathbb{N}}$ of i.i.d. random variables distributed according $\nu_3 = \mathrm{Uniform}((0, 1])$, such that for any $k \geq 1$ and $i \in [b]$, client i is active at step k if $X_k^{(3,i)} \leq p_i$. We denote $\mathcal{A}_{k+1} = \{i \in [b]; X_{k+1}^{(3,i)} \leq p_i\}$ the set of active clients at round k. Given a step-size $\gamma \in (0, \bar{\gamma}]$ for some $\bar{\gamma} > 0$ and starting from $\theta_0 \in \mathbb{R}^d$, `QLSD` recursively defines $(\theta_k)_{k\in\mathbb{N}}$, for any $k \in \mathbb{N}$, as

$$\theta_{k+1} = \theta_k - \gamma \textstyle\sum_{i\in\mathcal{A}_{k+1}} (1/p_i) \mathscr{C}_i(H_i(\theta_k, X_{k+1}^{(1,i)}), X_{k+1}^{(2,i)}) + \sqrt{2\gamma} Z_{k+1} \,, \tag{S3}$$

where $(Z_{k+1})_{k\in\mathbb{N}}$ is a sequence of standard Gaussian random variables. Let $\mathsf{X}_3 = [0,1]$. For any $i \in [b]$, consider the unbiased partial participation operator $\mathscr{S}_i : \mathbb{R}^d \times \mathsf{X}_3 \to \mathbb{R}^d$ defined, for any $\theta \in \mathbb{R}^d$ and $x^{(3)} \in \mathsf{X}_3$ by

$$\mathscr{S}_i(\theta, x^{(3)}) = \mathbf{1}\{x^{(3)} \le p_i\}\theta/p_i \,. \tag{S4}$$

Then, (S3) can be written of the form

$$\theta_{k+1} = \theta_k - \gamma \sum_{i=1}^{b} \tilde{H}_i(\theta_k, X_{k+1}^{(i)}) + \sqrt{2\gamma} Z_{k+1} \,, \qquad k \in \mathbb{N} \,, \tag{S5}$$

where for any $i \in [b]$, we denote $X_{k+1}^{(i)} = (X_{k+1}^{(1,i)}, X_{k+1}^{(2,i)}, X_{k+1}^{(3,i)})$ and for any $\theta \in \mathbb{R}^d, x^{(1,i)} \in \mathsf{X}_1^{(i)}, x^{(2)} \in \mathsf{X}_2$ and $x^{(3)} \in \mathsf{X}_3$,

$$\tilde{H}_i\big(\theta, (x^{(1,i)}, x^{(2)}, x^{(3)})\big) = \mathscr{S}_i\left(\mathscr{C}_i\Big(H_i(\theta, x^{(1,i)}), x^{(2)}\Big), x^{(3)}\right) . \tag{S6}$$

With this notation and setting for any $i \in [b]$ $\tilde{\mathsf{X}}^{(i)} = \mathsf{X}_1^{(i)} \times \mathsf{X}_2 \times \mathsf{X}_3$ and $\tilde{\nu}^{(i)} = \nu_1^{(i)} \otimes \nu_2 \otimes \nu_3$, the Markov kernel associated with (S3) is given for any $(\theta, \mathsf{A}) \in \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ by

$$Q_\gamma(\theta, \mathsf{A}) = \int_{\mathsf{A}\times\tilde{\mathsf{X}}^{(1)}\times\cdots\times\tilde{\mathsf{X}}^{(b)}} \exp\left(-\|\tilde{\theta} - \theta + \gamma\textstyle\sum_{i=1}^b \tilde{H}_i(\theta, x^{(i)})\|^2/(4\gamma)\right) \frac{\mathrm{d}\tilde{\theta}\,\tilde{\nu}^{(1)}(\mathrm{d}x^{(1)}) \otimes \cdots \otimes \tilde{\nu}^{(b)}(\mathrm{d}x^{(b)})}{(4\pi\gamma)^{d/2}} \,. \tag{S7}$$

The following result establishes an essential property of $\{\tilde{H}_i\}_{i\in[b]}$ under **H**S2 and **H**S3.

Lemma S1. *Assume **H**S2, **H**S3 and **H**S4. Then, for any $\theta \in \mathbb{R}^d$, we have*

$$\textstyle\sum_{i=1}^b \int_{\tilde{\mathsf{X}}^{(i)}} \tilde{H}_i(\theta, x^{(i)})\,\mathrm{d}\tilde{\nu}^{(i)}(x^{(i)}) = \nabla U(\theta) \,, \tag{S8}$$

$$\begin{aligned}\int_{\tilde{\mathsf{X}}^{(1:b)}} \left\|\textstyle\sum_{i=1}^b \tilde{H}_i(\theta, x^{(i)}) - \nabla U(\theta)\right\|^2 \otimes_{i=1}^b \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)}) &\le 2 \max_{i\in[b]}\{\mathtt{M}_i(\omega_i+1)/p_i\}\,\langle\theta - \theta^\star, \nabla U(\theta)\rangle \\ &\quad + 2\left[\sigma_\star^2 + (\mathtt{B}^\star/b)\sum_{i=1}^b (1 - p_i + \omega_i)/p_i\right] ,\end{aligned} \tag{S9}$$

where for any $i \in [b]$, $\tilde{H}_i$ is defined in (S6).

Proof. The first identity (S8) is straightforward using **H**S3-(i) and **H**S2-(i). We now show the inequality (S9). Let $\theta \in \mathbb{R}^d$. Using **H**S2-(i) or **H**S3-(i), we get

$$\begin{aligned}&\int_{\tilde{\mathsf{X}}^{(1:b)}} \left\|\textstyle\sum_{i=1}^b \tilde{H}_i(\theta, x^{(i)}) - \nabla U(\theta)\right\|^2 \otimes_{i=1}^b \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)}) \\ &\qquad = \int_{\tilde{\mathsf{X}}^{(1:b)}} \left\|\sum_{i=1}^b \left[\tilde{H}_i(\theta, x^{(i)}) - \mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right)\right]\right\|^2 \otimes_{i=1}^b \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)}) \\ &\qquad + \int_{\mathsf{X}_1^{(1:b)}\times\mathsf{X}_2^b} \left\|\sum_{i=1}^b \mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right) - \nabla U(\theta)\right\|^2 \nu_2^{\otimes b}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^b \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \,.\end{aligned} \tag{S10}$$

In addition, by **H**S2-(i) and **H**S2-(ii), we obtain

$$\begin{aligned}&\int_{\tilde{\mathsf{X}}^{(1:b)}} \left\|\sum_{i=1}^b \left[\tilde{H}_i(\theta, x^{(i)}) - \mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right)\right]\right\|^2 \otimes_{i=1}^b \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)}) \\ &= \sum_{i=1}^b \int_{\tilde{\mathsf{X}}^{(i)}} \left\|\tilde{H}_i(\theta, x^{(i)}) - \mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right)\right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)})\nu_2(\mathrm{d}x^{(2,i)})\nu_3(\mathrm{d}x^{(3,i)}) \\ &\le \sum_{i=1}^b \left(\frac{1-p_i}{p_i}\right) \int_{\mathsf{X}_1^{(i)}\times\mathsf{X}_2} \left\|\mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right)\right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)})\nu_2(\mathrm{d}x^{(2,i)})\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{b} \left(\frac{1-p_i}{p_i}\right) \int_{\mathsf{X}_1^{(i)}\times\mathsf{X}_2} \left\| \mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right) - H_i(\theta, x^{(1,i)}) + H_i(\theta, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)})\nu_2(\mathrm{d}x^{(2,i)}) \\
&= \sum_{i=1}^{b} \left(\frac{1-p_i}{p_i}\right) \int_{\mathsf{X}_1^{(i)}\times\mathsf{X}_2} \left\| \mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right) - H_i(\theta, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)})\nu_2(\mathrm{d}x^{(2,i)}) \\
&+ \sum_{i=1}^{b} \left(\frac{1-p_i}{p_i}\right) \int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq \sum_{i=1}^{b} \left[\left(\frac{1-p_i}{p_i}\right)(\omega_i+1)\right] \int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \,. \qquad \text{(S11)}
\end{aligned}
$$

Using $\|a\|^2 \leq 2\|a-b\|^2 + 2\|b\|^2$ and **H**S3-(ii)-(iii), for any $i \in [b]$, we obtain

$$
\begin{aligned}
\int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) &\leq 2\mathtt{M}_i \left\langle \theta - \theta^\star, \nabla U_i(\theta) - \nabla U_i(\theta^\star) \right\rangle \\
&\quad + 2\int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta^\star, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq 2\mathtt{M}_i \left\langle \theta - \theta^\star, \nabla U_i(\theta) - \nabla U_i(\theta^\star) \right\rangle + 2\mathtt{B}^\star/b \,. \qquad \text{(S12)}
\end{aligned}
$$

Therefore, combining this result and (S11) gives

$$
\begin{aligned}
&\int_{\tilde{\mathsf{X}}^{(1:b)}} \left\| \sum_{i=1}^{b} \left[\tilde{H}_i(\theta, x^{(i)}) - \mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right)\right] \right\|^2 \otimes_{i=1}^{b} \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)}) \\
&\leq 2\sum_{i=1}^{b} \mathtt{M}_i \left(\frac{1-p_i}{p_i}\right)(\omega_i+1) \left\langle \theta - \theta^\star, \nabla U_i(\theta) - \nabla U_i(\theta^\star) \right\rangle + \frac{2\mathtt{B}^\star}{b} \sum_{i=1}^{b} \left(\frac{1-p_i}{p_i}\right)(\omega_i+1) \,. \qquad \text{(S13)}
\end{aligned}
$$

Similarly, by **H**S2-(i) and **H**S2-(ii), we have

$$
\begin{aligned}
&\int_{\mathsf{X}_1^{(1:b)}\times\mathsf{X}_2^b} \left\| \sum_{i=1}^{b} \mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes b}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^{b} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&= \int_{\mathsf{X}_1^{(1:b)}\times\mathsf{X}_2^b} \left\| \sum_{i=1}^{b} \left[\mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2,i)}\right) - H_i(\theta, x^{(1,i)})\right] \right. \\
&\left. + \sum_{i=1}^{b} \{H_i(\theta, x^{(1,i)})\} - \nabla U(\theta) \right\|^2 \nu_2^{\otimes b}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^{b} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&= \sum_{i=1}^{b} \int_{\mathsf{X}_1^{(i)}\times\mathsf{X}_2} \left\| \mathscr{C}_i\left(H_i(\theta, x^{(1,i)}), x^{(2)}\right) - H_i(\theta, x^{(1,i)}) \right\|^2 \nu_2(\mathrm{d}x^{(2)})\nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&+ \int_{\mathsf{X}_1^{(1:b)}} \left\| \sum_{i=1}^{b} H_i(\theta, x^{(1,i)}) - \nabla U(\theta) \right\|^2 \otimes_{i=1}^{b} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq \sum_{i=1}^{b} \omega_i \int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) + \int_{\mathsf{X}_1^{(1:b)}} \left\| \sum_{i=1}^{b} H_i(\theta, x^{(1,i)}) - \nabla U(\theta) \right\|^2 \otimes_{i=1}^{b} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \,. \qquad \text{(S14)}
\end{aligned}
$$

Since for any $a, b \in \mathbb{R}^d$, $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, we have by **H**S3-(i)

$$
\begin{aligned}
&\int_{\mathsf{X}_1^{(1:b)}} \left\| \sum_{i=1}^{b} H_i(\theta, x^{(1,i)}) - \nabla U(\theta) \right\|^2 \otimes_{i=1}^{b} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&= \int_{\mathsf{X}_1^{(1:b)}} \left\| \sum_{i=1}^{b} \left[H_i(\theta, x^{(1,i)}) - \int_{\mathsf{X}_1^{(i)}} H_i(\theta, x^{(1)}) \nu_1^{(i)}(\mathrm{d}x^{(1)}) \right] \right\|^2 \otimes_{i=1}^{b} \nu_1^{(i)}(\mathrm{d}x^{(1,i)})
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{b} \int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta, x^{(1,i)}) - \int_{\mathsf{X}_1^{(i)}} H_i(\theta, x^{(1)}) \nu_1^{(i)}(\mathrm{d}x^{(1)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq 2 \sum_{i=1}^{b} \int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta, x^{(1,i)}) - H_i(\theta^\star, x^{(1,i)}) - \left[\int_{\mathsf{X}_1^{(i)}} (H_i(\theta, x^{(1)}) - H_i(\theta^\star, x^{(1)})) \nu_1^{(i)}(\mathrm{d}x^{(1)}) \right] \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&+ 2 \sum_{i=1}^{b} \int_{\mathsf{X}_1^{(i)}} \left\| H_i(\theta_\star, x^{(1,i)}) - \int_{\mathsf{X}_1^{(i)}} H_i(\theta_\star, x^{(1)}) \nu_1^{(i)}(\mathrm{d}x^{(1)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq 2\sigma_\star^2 + 2 \sum_{i=1}^{b} \mathtt{M}_i \left\langle \nabla U_i(\theta) - \nabla U_i(\theta^\star), \theta - \theta^\star \right\rangle . \qquad \text{(S15)}
\end{aligned}
$$

By combining (S12), (S14) and (S15), we obtain

$$
\begin{aligned}
&\int_{\mathsf{X}_1^{(1:b)}} \left\| \sum_{i=1}^{b} \mathscr{C}_i \left(H_i(\theta, x^{(1,i)}), x^{(2,i)} \right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes b}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^{b} \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq 2 \sum_{i=1}^{b} \mathtt{M}_i(\omega_i + 1) \left\langle \nabla U_i(\theta) - \nabla U_i(\theta^\star), \theta - \theta^\star \right\rangle + 2 \left(\sigma_\star^2 + \frac{2\mathtt{B}^\star}{b} \sum_{i=1}^{b} \omega_i \right) .
\end{aligned}
$$

Finally, the last inequality combined with (S10) and (S13) completes the proof. □

In view of Lemma S1, it suffices to study the recursion specified in (S5) under the following assumption on $(\tilde{H}_i)_{i\in[b]}$ gathered in **H**S5. Indeed, Lemma S1 shows that Condition **H**S5 below holds with $\mathsf{X}^{(i)} = \tilde{\mathsf{X}}^{(i)} = \mathsf{X}_1^{(i)} \times \mathsf{X}_2 \times \mathsf{X}_3$, $\mathcal{X}^{(i)} = \tilde{\mathcal{X}}^{(i)} = \mathcal{X}_1^{(i)} \otimes \mathcal{X}_2 \otimes \mathcal{X}_3$, $\tilde{\nu}^{(i)} = \nu_1^{(i)} \otimes \nu_2 \otimes \nu_3$, $\{\tilde{H}_i\}_{i=1}^b = \{F_i\}_{i=1}^b$,

$$
\tilde{\mathtt{M}} = 2 \max_{i\in[b]} \{\mathtt{M}_i(1+\omega_i)/p_i\} ,
$$

$$
\tilde{\mathtt{B}}^\star = 2[\sigma_\star^2 + (\mathtt{B}^\star/b) \sum_{i=1}^{b} (1 - p_i + \omega_i)/p_i] .
$$

H S5. *There exists a family of probability measure $\{\nu^{(i)}\}_{i\in[b]}$ on a measurable spaces $\{(\tilde{\mathsf{X}}^{(i)}, \tilde{\mathcal{X}}^{(i)})\}_{i\in[b]}$ and a family of measurable functions $\{F_i : \mathbb{R}^d \times \mathsf{X}^{(i)} \to \mathbb{R}^d\}_{i\in[b]}$ such that the following conditions hold.*

(i) For any $\theta \in \mathbb{R}^d$, we have

$$
\sum_{i=1}^{b} \int_{\tilde{\mathsf{X}}^{(i)}} F_i(\theta, x^{(i)}) \nu^{(i)}(\mathrm{d}x^{(i)}) = \nabla U(\theta) .
$$

(ii) There exists $(\tilde{\mathtt{M}}, \tilde{\mathtt{B}}^\star) \in \mathbb{R}_+^2$ such that for any $\theta \in \mathbb{R}^d$, we have

$$
\int_{\tilde{\mathsf{X}}^{(1:b)}} \left\| \sum_{i=1}^{b} F_i(\theta, x^{(i)}) - \nabla U(\theta) \right\|^2 \otimes_{i=1}^{b} \nu^{(i)}(\mathrm{d}x^{(i)}) \leq \tilde{\mathtt{M}} \left\langle \theta - \theta^\star, \nabla U(\theta) - \nabla U(\theta^\star) \right\rangle + \tilde{\mathtt{B}}^\star .
$$

Then under **H**S5, consider $(X_k^{(1)}, \ldots, X_k^{(b)})_{k\in\mathbb{N}^*}$ an independent sequence distributed according to $\otimes_{i=1}^b \nu^{(i)}$. Define the general recursion

$$
\tilde{\theta}_{k+1} = \tilde{\theta}_k - \gamma \sum_{i=1}^{b} F_i(\tilde{\theta}_k, X_{k+1}^{(i)}) + \sqrt{2\gamma} Z_{k+1} , \qquad k \in \mathbb{N} . \qquad \text{(S16)}
$$

and the corresponding the Markov kernel given for any $\gamma \in \mathbb{R}_+^*$, $\theta \in \mathbb{R}^d, \mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$ by

$$
\tilde{Q}_\gamma(\theta, \mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}\times\tilde{\mathsf{X}}^{(1:b)}} \exp(-(4\gamma)^{-1} \|\bar{\theta} - \theta + \gamma \textstyle\sum_{i=1}^{b} F_i(\theta, x^{(i)})\|^2) \, \mathrm{d}\bar{\theta} \, \mathrm{d} \otimes_{i=1}^{b} \nu^{(i)}(x^{(i)}) . \qquad \text{(S17)}
$$

We refer to this Markov kernel as the generalised `QLSD` kernel. In our next section, we establish quantitative bounds between the iterates of this kernel and π in W_2. We then apply this result to `QLSD` and `QLSD`* as particular cases.

S1.2 Quantitative bounds for the generalised QLSD kernel

Define

$$\bar{\gamma} = \bar{\gamma}_1 \wedge \bar{\gamma}_2 \wedge \bar{\gamma}_3 \,, \quad \bar{\gamma}_1 = 2/[5(\mathtt{m}+\mathtt{L})] \,, \quad \bar{\gamma}_2 = (\mathtt{m}+\mathtt{L}+\tilde{\mathtt{M}})^{-1} \,, \quad \bar{\gamma}_3 = (10\mathtt{m})^{-1} \,. \tag{S18}$$

Theorem S4. *Assume **HS1** and **HS5**. Then, for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, any step size $\gamma \in (0, \bar{\gamma}]$, any $k \in \mathbb{N}$, we have*

$$W_2^2(\mu \tilde{Q}_\gamma^k, \pi) \leq (1-\gamma\mathtt{m}/2)^k W_2^2(\mu, \pi) + \gamma \tilde{B}_{\bar{\gamma}} + \gamma^2 \tilde{A}_{\bar{\gamma}} (1-\mathtt{m}\gamma/2)^{k-1} k \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^2 \mu(\mathrm{d}\theta) \,,$$

where $\tilde{Q}_\gamma$ is defined in (S17) and

$$\tilde{B}_{\bar{\gamma}} = (2d\mathtt{L}^2/\mathtt{m})(1/\mathtt{m} + 5\bar{\gamma}) \left[1 + \bar{\gamma}\mathtt{L}^2/(2\mathtt{m}) + \bar{\gamma}^2\mathtt{L}^2/12\right] + 2\tilde{\mathtt{B}}^\star/\mathtt{m} + 2\mathtt{L}\tilde{\mathtt{M}}(2d + \bar{\gamma}\tilde{\mathtt{B}}^\star)/\mathtt{m}^2$$
$$\tilde{A}_{\bar{\gamma}} = \mathtt{L}\tilde{\mathtt{M}} \,.$$

Let $\xi \in \mathcal{P}_2(\mathbb{R}^{2d})$ be a probability measure on $(\mathbb{R}^{2d}, \mathcal{B}(\mathbb{R}^{2d}))$ with marginals ξ_1 and ξ_2, *i.e.* $\xi(\mathsf{A} \times \mathbb{R}^d) = \xi_1(\mathsf{A})$ and $\xi(\mathsf{A} \times \mathbb{R}^d) = \xi_2(\mathsf{A})$ for any $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$. Note that under **H**S1, the Langevin diffusion defines a Markov semigroup $(P_t)_{t\geq 0}$ satisfying $\pi P_t = \pi$ for any $t \geq 0$, see *e.g.* Roberts and Tweedie (1996, Theorem 2.1). We introduce a synchronous coupling $(\vartheta_{k\gamma}, \theta_k)$ between $\xi_1 P_{k\gamma}$ and $\xi_2 \tilde{Q}_\gamma^k$ for any $k \in \mathbb{N}$ based on a d-dimensional standard Brownian motion $(B_t)_{t\geq 0}$ and a couple of random variables (θ_0, ϑ_0) with distribution ξ independent of $(B_t)_{t\geq 0}$. Consider $(\vartheta_t)_{t\geq 0}$ the strong solution of the Langevin stochastic differential equation (SDE)

$$\mathrm{d}\vartheta_t = -\nabla U(\vartheta_t)\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t \,, \tag{S19}$$

starting from ϑ_0. Note that under **H**S1-(i), this SDE admits a unique strong solution (Revuz and Yor, 2013, Theorem (2.1) in Chapter IX). In addition, define $(\theta_k)_{k\in\mathbb{N}}$ starting from θ_0 and satisfying the recursion: for $k \geq 0$,

$$\theta_{k+1} = \theta_k - \gamma \sum_{i=1}^b F_i(\theta_k, x_{k+1}^{(i)}) + \sqrt{2}(B_{\gamma(k+1)} - B_{\gamma k}) \,, \tag{S20}$$

where $(x_j^{(1)}, \ldots, x_j^{(b)})_{j\in\mathbb{N}^*}$ is an independent sequence of random variables with distribution $\otimes_{i=1}^b \nu^{(i)}$. Then, by definition, $(\vartheta_{k\gamma}, \theta_k)$ is a coupling between $\xi_1 P_{k\gamma}$ and $\xi_2 \tilde{Q}_\gamma^k$ for any $k \in \mathbb{N}$ and therefore

$$W_2(\xi_1 P_{k\gamma}, \xi_2 \tilde{Q}_\gamma^k) \leq \mathbb{E}\left[\|\vartheta_{\gamma k} - \theta_k\|^2\right]^{1/2} \,. \tag{S21}$$

We can now give the proof of Theorem S4.

Proof. By Villani (2008, Theorem 4.1), for any couple of probability measures on $\mathbb{R}^d$, there exists an optimal transference plan $\xi^\star$ between ν and π since $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ by the strong convexity assumption **H**S1-(i). Let (ϑ_0, θ_0) be a corresponding coupling which therefore satisfies $W_2(\mu, \pi) = \mathbb{E}^{1/2}[\|\vartheta_0 - \theta_0\|^2]$. Consider then $(\vartheta_k)_{k\in\mathbb{N}}, (\theta_k)_{k\in\mathbb{N}}$ defined in (S19)-(S20) starting from (ϑ_0, θ_0). Note that since $\pi P_t = \pi$ by Roberts and Tweedie (1996, Theorem 2.1) for any $t \geq 0$ and θ_0 has distribution π, we get by Durmus and Moulines (2019, Proposition 1) that for any $k \in \mathbb{N}$, $\mathbb{E}[\|\vartheta_{k\gamma} - \theta^\star\|^2] \leq d/\mathtt{m}$ and then Lemma S3 below shows that for any $k \in \mathbb{N}$,

$$\mathbb{E}[\left\|\vartheta_{(k+1)\gamma} - \theta_{k+1}\right\|^2] \leq \kappa_\gamma \mathbb{E}[\|\vartheta_{k\gamma} - \theta_k\|^2] + \gamma^2 \mathtt{L}\tilde{\mathtt{M}}\, \mathbb{E}\left[\|\theta_0 - \theta^\star\|^2\right] \tilde{\kappa}_\gamma^k + \gamma^2 \mathrm{D}_\gamma \,,$$

where we have set

$$\kappa_\gamma = 1 - \gamma\mathtt{m}(1 - 5\gamma\mathtt{m}) \,, \qquad \tilde{\kappa}_\gamma = 1 - \gamma\mathtt{m}\left[2 - \gamma(\mathtt{m} + \tilde{\mathtt{M}})\right] \,, \qquad \mathrm{D}_\gamma = \mathrm{D}_{0,\gamma} + (1/\mathtt{m} + 5\gamma)(\gamma d \mathtt{L}^4/2\mathtt{m}) \,.$$

A straightforward induction shows that

$$\mathbb{E}[\|\vartheta_{k\gamma} - \theta_k\|^2] \leq \kappa_\gamma^k W_2^2(\mu, \pi) + \gamma^2 \mathtt{L}\tilde{\mathtt{M}}\, \mathbb{E}\left[\|\theta_0 - \theta^\star\|^2\right] \sum_{l=0}^{k-1} \kappa_\gamma^l \tilde{\kappa}_\gamma^{k-1-l} + \gamma^2 \mathrm{D}_\gamma/(1-\kappa_\gamma) \,.$$

Using $\kappa_\gamma \wedge \tilde{\kappa}_\gamma \leq 1 - \mathtt{m}\gamma/2$ since $\gamma \leq \bar{\gamma}$, (S21) and $\pi P_t = \pi$ for any $t \geq 0$ completes the proof. □

S1.2.1 Supporting Lemmata

In this subsection, we derived two lemmas. Taking $(\theta_k)_{k\in\mathbb{N}}$ defined by the recursion (S20), Lemma S2 aims to upper bound the squared deviation between θ_k and the minimiser of U denoted $\theta^\star$, for any $k \in \mathbb{N}$.

Lemma S2. *Assume **HS1** and **HS5**. Let $\gamma \in \left(0, 2/(\mathtt{m}+\mathtt{L}+\tilde{\mathtt{M}})\right]$. Then, for any $k \in \mathbb{N}, \theta_0 \in \mathbb{R}^d$, we have*

$$\int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2 \, \tilde{Q}_\gamma^k(\theta_0, \mathrm{d}\theta) \leq (1-\gamma\mathtt{m}\left[2-\gamma(\mathtt{m}+\tilde{\mathtt{M}})\right])^k \|\theta_0-\theta^\star\|^2 + \frac{2d+\gamma\tilde{\mathtt{B}}^\star}{\mathtt{m}\left[2-\gamma(\mathtt{m}+\tilde{\mathtt{M}})\right]} \,,$$

where $\tilde{Q}_\gamma$ is defined in (S17).

Proof. For any $\theta_0 \in \mathbb{R}^d$, by definition (S17) of $\tilde{Q}_\gamma$ and using **HS5**-(i), we obtain

$$\begin{aligned}\int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2 \, \tilde{Q}_\gamma(\theta_0, \mathrm{d}\theta) = \|\theta_0-\theta^\star\|^2 - 2\gamma \left\langle \theta_0-\theta^\star, \nabla U(\theta_0)\right\rangle \\ + \gamma^2 \int_{\tilde{\mathsf{X}}^{(1:b)}} \left\|\textstyle\sum_{i=1}^b F_i(\theta_0, x^{(i)})\right\|^2 \otimes_{i=1}^b \nu^{(i)}(\mathrm{d}x^{(i)}) + 2\gamma d \,. \end{aligned} \tag{S22}$$

Moreover, using **HS1**, **HS5** and (S1), it follows that

$$\begin{aligned}\int_{\tilde{\mathsf{X}}^{(1:b)}} \left\|\textstyle\sum_{i=1}^b F_i(\theta_0, x^{(i)})\right\|^2 \otimes_{i=1}^b \nu^{(i)}(\mathrm{d}x^{(i)}) &= \int_{\tilde{\mathsf{X}}^{(1:b)}} \left\|\textstyle\sum_{i=1}^b F_i(\theta_0, x^{(i)}) - \nabla U(\theta_0)\right\|^2 \otimes_{i=1}^b \nu^{(i)}(\mathrm{d}x^{(i)}) \\ &\quad + \|\nabla U(\theta_0)\|^2 \\ &\leq \tilde{\mathtt{M}} \left\langle \theta_0-\theta^\star, \nabla U(\theta_0)\right\rangle + \tilde{\mathtt{B}}^\star + \|\nabla U(\theta_0)-\nabla U(\theta^\star)\|^2 \\ &\leq [\mathtt{m}+\mathtt{L}+\tilde{\mathtt{M}}] \left\langle \theta_0-\theta^\star, \nabla U(\theta_0)\right\rangle + \tilde{\mathtt{B}}^\star - \mathtt{L}\mathtt{m} \|\theta_0-\theta^\star\|^2 \,. \end{aligned} \tag{S23}$$

Plugging (S23) in (S22) implies

$$\int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2 \, \tilde{Q}_\gamma(\theta_0, \mathrm{d}\theta) \;\leq\; (1-\gamma^2\mathtt{m}\mathtt{L})\|\theta_0-\theta^\star\|^2 - \gamma\{2-\gamma[\mathtt{m}+\mathtt{L}+\tilde{\mathtt{M}}]\} \left\langle \theta_0-\theta^\star, \nabla U(\theta_0)\right\rangle + \gamma^2\tilde{\mathtt{B}}^\star + 2\gamma d \,.$$

Using **HS1**-(i), we have $\langle \theta_0-\theta^\star, \nabla U(\theta_0)\rangle \geq \mathtt{m}\|\theta_0-\theta^\star\|^2$ which, combined with the condition $\gamma \leq 1/(\mathtt{m}+\mathtt{L}+\tilde{\mathtt{M}})$, gives

$$\int_{\mathbb{R}^d} \|\theta-\theta^\star\|^2 \, \tilde{Q}_\gamma(\theta_0, \mathrm{d}\theta) \leq (1-\gamma\mathtt{m}[2-\gamma(\mathtt{m}+\tilde{\mathtt{M}})])\|\theta_0-\theta^\star\|^2 + \gamma(2d+\gamma\tilde{\mathtt{B}}^\star) \,.$$

Using $0 < \gamma < 2/(\mathtt{m}+\tilde{\mathtt{M}})$ and the Markov property combined with a straightforward induction completes the proof. □

For any $k \in \mathbb{N}$, the following lemma gives an explicit upper bound on the expected squared norm between ϑ_{k+1} and θ_{k+1} in function of ϑ_k, θ_k. The purpose of this lemma is to derive a contraction property involving a contracting term and a bias term which is easy to control.

Lemma S3. *Assume **HS1** and **HS5**. Consider $(\vartheta_t)_{t\geq 0}$ and $(\theta_k)_{k\in\mathbb{N}}$ defined in (S19) and (S20), respectively, for some initial distribution $\xi \in \mathcal{P}_2(\mathbb{R}^{2d})$. For any $k \in \mathbb{N}$ and $\gamma \in \left(0, 2/[(5(\mathtt{m}+\mathtt{L})) \vee (\mathtt{m}+\tilde{\mathtt{M}}+\mathtt{L})]\right)$, we have*

$$\begin{aligned}\mathbb{E}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] \leq \{1-\gamma\mathtt{m}(1-5\gamma\mathtt{m})\}\mathbb{E}\left[\|\vartheta_k-\theta_k\|^2\right] + \gamma^2 \mathrm{D}_{0,\gamma} \\ + \gamma^2\mathtt{L}\tilde{\mathtt{M}}(1-\gamma\mathtt{m}\left[2-\gamma(\mathtt{m}+\tilde{\mathtt{M}})\right])^k \mathbb{E}[\|\theta_0-\theta^\star\|^2] \\ + \gamma^3(1/\mathtt{m}+5\gamma)\mathtt{L}^4\mathbb{E}\left[\|\vartheta_{k\gamma}-\theta^\star\|^2\right]/2 \,, \end{aligned}$$

where

$$\mathrm{D}_{0,\gamma} = d\mathtt{L}^2(1/\mathtt{m}+5\gamma)\left[1+\gamma^2\mathtt{L}^2/12\right] + \tilde{\mathtt{B}}^\star + \frac{\mathtt{L}\tilde{\mathtt{M}}(2d+\gamma\tilde{\mathtt{B}}^\star)}{\mathtt{m}\left[2-\gamma(\mathtt{m}+\tilde{\mathtt{M}})\right]} \,.$$

Proof. Let $k \in \mathbb{N}$. By (S19) and (S20), we have

$$\vartheta_{\gamma(k+1)} - \theta_{k+1} = \vartheta_{\gamma k} - \theta_k - \gamma \left[\nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right] - \int_0^\gamma \left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right] \mathrm{d}s + \gamma \sum_{i=1}^b \left[F_i(\theta_k, X_{k+1}^{(i)}) - \nabla U_i(\theta_k)\right] .$$

Define the filtration $(\mathcal{F}_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$ as $\mathcal{F}_0 = \sigma(\vartheta_0, \theta_0)$ and for $\tilde{k} \in \mathbb{N}^*$,

$$\mathcal{F}_{\tilde{k}} = \sigma(\vartheta_0, \theta_0, (X_l^{(1)}, \dots, X_l^{(b)})_{1\le l\le \tilde{k}}, (B_t)_{0\le t\le \gamma\tilde{k}}) .$$

Note that since $(\vartheta_t)_{t\ge 0}$ is a strong solution of (S19), then is easy to see that $(\vartheta_{\gamma\tilde{k}}, \theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$ is $(\mathcal{F}_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$-adapted. Taking the squared norm and the conditional expectation with respect to $\mathcal{F}_k$, we obtain using **H**S5-(i) that

$$\begin{aligned}
\mathbb{E}^{\mathcal{F}_k}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] &= \left\|\vartheta_{\gamma k} - \theta_k\right\|^2 - 2\gamma \left\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\rangle \\
&\quad + 2\gamma \int_0^\gamma \left\langle \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k), \mathbb{E}^{\mathcal{F}_k}\left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right]\right\rangle \mathrm{d}s \\
&\quad - 2\int_0^\gamma \left\langle \vartheta_{\gamma k} - \theta_k, \mathbb{E}^{\mathcal{F}_k}\left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right]\right\rangle \mathrm{d}s \\
&\quad + \gamma^2 \left\|\nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\|^2 \\
&\quad + \mathbb{E}^{\mathcal{F}_k}\left[\left\|\int_0^\gamma \left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right] \mathrm{d}s\right\|^2\right] \\
&\quad + \gamma^2 \mathbb{E}^{\mathcal{F}_k}\left[\left\|\sum_{i=1}^b F_i(\theta_k, X_{k+1}^{(i)}) - \nabla U(\theta_k)\right\|^2\right] .
\end{aligned} \tag{S24}$$

First, using Jensen inequality and the fact that for any $a, b \in \mathbb{R}^d$, $|\langle a, b\rangle| \le 2\left\|a\right\|^2 + 2\left\|b\right\|^2$, we get

$$\begin{aligned}
&\int_0^\gamma \left\langle \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k), \mathbb{E}^{\mathcal{F}_k}\left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right]\right\rangle \mathrm{d}s \\
&\qquad \le 2\gamma \left\|\nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\|^2 + 2\int_0^\gamma \mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \mathrm{d}s ,
\end{aligned} \tag{S25}$$

$$\mathbb{E}^{\mathcal{F}_k}\left[\left\|\int_0^\gamma \left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right] \mathrm{d}s\right\|^2\right] \le \gamma \int_0^\gamma \mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \mathrm{d}s .$$

In addition, given that for any $\varepsilon > 0, a, b \in \mathbb{R}^d$, $|\langle a, b\rangle| \le \varepsilon \left\|a\right\|^2 + (4\varepsilon)^{-1}\left\|b\right\|^2$, we get

$$\begin{aligned}
\left|\int_0^\gamma \left\langle \theta_k - \vartheta_{\gamma k}, \mathbb{E}^{\mathcal{F}_k}\left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right]\right\rangle \mathrm{d}s\right| &\le \gamma\varepsilon \left\|\vartheta_{\gamma k} - \theta_k\right\|^2 \\
&\quad + (4\varepsilon)^{-1} \int_0^\gamma \mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \mathrm{d}s .
\end{aligned} \tag{S26}$$

By **H**S1, for $k \in \mathbb{N}$ we get by (S1)

$$\left\|\nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\|^2 \le (\mathtt{m} + \mathtt{L}) \left\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\rangle - \mathtt{mL} \left\|\vartheta_{\gamma k} - \theta_k\right\|^2 . \tag{S27}$$

Lastly, **H**S5-(ii) yields

$$\mathbb{E}^{\mathcal{F}_k}\left[\left\|\sum_{i=1}^b F_i(\theta_k, X_{k+1}^{(i)}) - \nabla U(\theta_k)\right\|^2\right] \le \tilde{\mathtt{M}} \left\langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star)\right\rangle + \tilde{\mathtt{B}}^\star . \tag{S28}$$

Combining (S25), (S26), (S27) and (S28) into (S24), for $k \in \mathbb{N}$ we get for any $\varepsilon > 0$,

$$\mathbb{E}^{\mathcal{F}_k}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] \le (1 + 2\gamma\varepsilon - 5\gamma^2\mathtt{mL}) \left\|\vartheta_{\gamma k} - \theta_k\right\|^2$$

$$
\begin{aligned}
&- \gamma \left[2 - 5\gamma(\mathtt{m} + \mathtt{L})\right] \langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k) \rangle \\
&+ (5\gamma + (2\varepsilon)^{-1}) \int_0^\gamma \mathbb{E}^{\mathcal{F}_k} \left[\|\nabla U(\vartheta_{\gamma k + s}) - \nabla U(\vartheta_{\gamma k})\|^2 \right] \mathrm{d}s \\
&+ \gamma^2 \tilde{\mathtt{M}} \langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star) \rangle + \gamma^2 \tilde{\mathtt{B}}^\star \,. \qquad \text{(S29)}
\end{aligned}
$$

Next, we use that under **H**S1, $\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k) \rangle \geq \mathtt{m} \|\vartheta_{\gamma k} - \theta_k\|^2$ and $|\langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star) \rangle| \leq \mathtt{L} \|\theta_k - \theta^\star\|^2$, which implies taking $\varepsilon = \mathtt{m}/2$ and since $2 - 5\gamma(\mathtt{m} + \mathtt{L}) \geq 0$,

$$
\begin{aligned}
\mathbb{E}^{\mathcal{F}_k} \left[\left\| \vartheta_{\gamma(k+1)} - \theta_{k+1} \right\|^2 \right] &\leq (1 - \gamma\mathtt{m}(1 - 5\gamma\mathtt{m})) \left\| \vartheta_{\gamma k} - \theta_k \right\|^2 \\
&+ (5\gamma + \mathtt{m}^{-1}) \int_0^\gamma \mathbb{E}^{\mathcal{F}_k} \left[\|\nabla U(\vartheta_{\gamma k + s}) - \nabla U(\vartheta_{\gamma k})\|^2 \right] \mathrm{d}s \\
&+ \gamma^2 \tilde{\mathtt{M}} \mathtt{L} \left\| \theta_k - \theta^\star \right\|^2 + \gamma^2 \tilde{\mathtt{B}}^\star \,. \qquad \text{(S30)}
\end{aligned}
$$

Further, for any $s \in \mathbb{R}_+$, using Durmus and Moulines (2019, Lemma 21) we have

$$
\mathtt{L}^{-2} \, \mathbb{E}^{\mathcal{F}_k} \left[\|\nabla U(\vartheta_{\gamma k + s}) - \nabla U(\vartheta_{\gamma k})\|^2 \right] \leq ds \left(2 + s^2 \mathtt{L}^2/3\right) + 3s^2 \mathtt{L}^2/2 \left\| \vartheta_{\gamma k} - \theta^\star \right\|^2 .
$$

Integrating the previous inequality on $[0, \gamma]$, for $k \geq 0$ we obtain

$$
\mathtt{L}^{-2} \int_0^\gamma \mathbb{E}^{\mathcal{F}_k} \left[\|\nabla U(\vartheta_{\gamma k + s}) - \nabla U(\vartheta_{\gamma k})\|^2 \right] \mathrm{d}s \leq d\gamma^2 + d\gamma^4 \mathtt{L}^2/12 + \gamma^3 \mathtt{L}^2/2 \left\| \vartheta_{\gamma k} - \theta^\star \right\|^2 .
$$

Plugging this bounds in (S30) and taking the expectation combined with Lemma S2 conclude the proof. □

S1.3 Proof of Theorem 1

Based on Theorem S4, the next corollary explicits an upper bound in Wasserstein distance between π and μQ_γ^k, where we consider $(\theta_k)_{k \in \mathbb{N}}$ defined in (S3) and starting from $\tilde{\theta}$ following $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.

Theorem S5. *Assume **H**S1, **H**S2, **H**S3 and **H**S4. Then, for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, any step size $\gamma \in (0, \bar{\gamma}]$ where $\bar{\gamma}$ is defined in (S18), any $k \in \mathbb{N}$, we have*

$$
W_2^2(\mu Q_\gamma^k, \pi) \leq (1 - \gamma\mathtt{m}/2)^k W_2^2(\mu, \pi) + \gamma B_{\bar{\gamma}} + \gamma^2 A_{\bar{\gamma}} (1 - \mathtt{m}\gamma/2)^{k-1} k \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^2 \mu(\mathrm{d}\theta) \,,
$$

where Q_γ is defined in (S7) and

$$
\begin{aligned}
B_{\bar{\gamma}} &= (2d\mathtt{L}^2/\mathtt{m}) \left(1/\mathtt{m} + 5\bar{\gamma}\right) \left[1 + \bar{\gamma}\mathtt{L}^2/(2\mathtt{m}) + \bar{\gamma}^2 \mathtt{L}^2/12\right] + 4[\sigma_\star^2 + (\mathtt{B}^\star/b) \sum_{i=1}^b (1 - p_i + \omega_i)/p_i]/\mathtt{m} \\
&\quad + 8\mathtt{L} \max_{i \in [b]} \{\mathtt{M}_i (1 + \omega_i)/p_i\}) \left[d + \bar{\gamma}[\sigma_\star^2 + (\mathtt{B}^\star/b) \sum_{i=1}^b (1 - p_i + \omega_i)/p_i] \right] /\mathtt{m}^2 \qquad \text{(S31)} \\
A_{\bar{\gamma}} &= 2\mathtt{L} \max_{i \in [b]} \{\mathtt{M}_i (1 + \omega_i)/p_i\}) \,.
\end{aligned}
$$

Proof. By Lemma S1, the assumption **H**S5 is satified for a choice of $\tilde{\mathtt{M}} = 2 \max_{i \in [b]} \{\mathtt{M}_i (1 + \omega_i)/p_i\})$ and $\tilde{\mathtt{B}}^\star = 2[\sigma_\star^2 + (\mathtt{B}^\star/b) \sum_{i=1}^b (1 - p_i + \omega_i)/p_i]$. Therefore, applying Theorem S4 completes the proof. □

S2 PROOF OF Theorem 2

We assume here that $\{U_i\}_{i \in [b]}$ are defined, for any $i \in [b]$ and $\theta \in \mathbb{R}^d$, by

$$
U_i(\theta) = \sum_{j=1}^{N_i} U_{i,j}(\theta) \,, \quad N_i \in \mathbb{N}^* \,.
$$

We consider the following set of assumptions on $\{U_i\}_{i \in [b]}$ and $\{U_{i,j} : j \in [N_i]\}_{i \in [b]}$.

HS6. *For any $i \in [b], j \in [N_i]$, $U_{i,j}$ is continuously differentiable and the following conditions hold.*

(i) There exist $\{\mathtt{M}_i > 0\}_{i\in[b]}$, such that for any $i \in [b]$, $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\|\nabla U_i(\theta_2) - \nabla U_i(\theta_2)\|^2 \leq \mathtt{M}_i \langle \theta_2 - \theta_1, \nabla U_i(\theta_2) - \nabla U_i(\theta_1)\rangle \ .$$

(ii) There exists $\bar{\mathtt{M}} \geq 0$ such that, for any $\theta_1, \theta_2 \in \mathbb{R}^d$,

$$\|\nabla U_{i,j}(\theta_2) - \nabla U_{i,j}(\theta_1)\|^2 \leq \bar{\mathtt{M}} \langle \nabla U_{i,j}(\theta_2) - \nabla U_{i,j}(\theta_1), \theta_2 - \theta_1\rangle \ .$$

In all this section, we assume for any $i \in [b]$ that $n_i \in \mathbb{N}^*$, $n_i \leq N_i$ is fixed. For any $i \in [b]$, recall that $\wp_{N_i}$ denotes the power set of $[N_i]$ and

$$\wp_{N_i,n_i} = \{x \in \wp_{N_i} \ : \ \mathrm{card}(x) = n_i\} \, .$$

We set in this section $\nu_1^{(i)}$ as the uniform distribution on $\wp_{N_i,n_i}$. We consider the family of measurable functions $\{H_i^\star : \mathbb{R}^d \times \mathbb{R}^d \times \wp_{N_i} \to \mathbb{R}^d\}_{i\in[b]}$, defined for any $i \in [b]$, $\theta \in \mathbb{R}^d$, $x \in \wp_{N_i,n_i}$ by

$$H_i^\star(\theta, x) = \frac{N_i}{n_i} \sum_{j=1}^{N_i} \mathbf{1}_x(j) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right] \ . \tag{S32}$$

Using this specific family of gradient estimators boils down to the $\mathtt{QLSD}^\star$ algorithm detailed in Algorithm S1.

Algorithm S1 Variance-reduced Quantised Langevin Stochastic Dynamics ($\mathtt{QLSD}^\star$)

Input: minibatch sizes $\{n_i\}_{i\in[b]}$, number of iterations K, compression operators $\{\mathscr{C}_{k+1}\}_{k\in\mathbb{N}^*}$, step-size $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$ and initial point θ_0.
for $k = 0$ **to** $K - 1$ **do**
  for $i \in \mathcal{A}_{k+1}$ // On active clients **do**
    Draw $\mathcal{S}_{k+1}^{(i)} \sim \mathrm{Uniform}\,(\wp_{N_i,n_i})$.
    Set $H_{k+1}^{(i)}(\theta_k) = (N_i/n_i) \sum_{j\in\mathcal{S}_{k+1}^{(i)}} [\nabla U_{i,j}(\theta_k) - \nabla U_{i,j}(\theta^\star)]$.
    Compute $g_{i,k+1} = \mathscr{C}_{k+1}\left(H_{k+1}^{(i)}(\theta_k)\right)$.
    Send $g_{i,k+1}$ to the central server.
  end for
  // On the central server
  Compute $g_{k+1} = \frac{b}{|\mathcal{A}_{k+1}|} \sum_{i\in\mathcal{A}_{k+1}} g_{i,k+1}$.
  Draw $Z_{k+1} \sim \mathrm{N}(0_d, \mathrm{I}_d)$.
  Compute $\theta_{k+1} = \theta_k - \gamma g_{k+1} + \sqrt{2\gamma} Z_{k+1}$.
  Send θ_{k+1} to the b clients.
end for
Output: samples $\{\theta_k\}_{k=0}^K$.

Let $(X_k^{(1,1)}, \ldots, X_k^{(1,b)})_{k\in\mathbb{N}^*}$ and $(X_k^{(2,1)}, \ldots, X_k^{(2,b)})_{k\in\mathbb{N}^*}$ be two independent i.i.d. sequences with distribution $\otimes_{i=1}^b \nu_1^{(i)}$ and $\nu_2^{\otimes b}$. Let $(Z_k)_{k\in\mathbb{N}^*}$ be an i.i.d. sequence of d-dimensional standard Gaussian random variables independent of $(X_k^{(1,1)}, \ldots, X_k^{(1,b)})_{k\in\mathbb{N}^*}$ and $(X_k^{(2,1)}, \ldots, X_k^{(2,b)})_{k\in\mathbb{N}^*}$. Similarly as before, we consider the partial device participation context where at each communication round $k \geq 1$, each client has a probability $p_i \in (0, 1]$ of participating, independently from other clients. In other words, there exists a sequence $(X_k^{(3,1)}, \cdots, X_k^{(3,b)})_{k\in\mathbb{N}^*}$ of i.i.d. random variables distributed according $\nu_3 = \mathrm{Uniform}((0, 1])$, such that for any $k \geq 1$ and $i \in [b]$, client i is active at step k if $X_k^{(3,i)} \leq p_i$. We denote $\mathcal{A}_{k+1} = \{i \in [b]; X_{k+1}^{(3,i)} \leq p_i\}$ the set of active clients at round k. For ease of notation, denote for any $k \in \mathbb{N}^*$, $X_k^{(1)} = (X_k^{(1,1)}, \ldots, X_k^{(1,b)})$, $X_k^{(2)} = (X_k^{(2,1)}, \ldots, X_k^{(2,b)})$, $X_k^{(3)} = (X_k^{(3,1)}, \ldots, X_k^{(3,b)})$ and $X_k = (X_k^{(1)}, X_k^{(2)}, X_k^{(3)})$.

Note that with this notation and under **HS2**, $\mathtt{QLSD}^\star$ can be cast into the framework of the generalised $\mathtt{QLSD}$ scheme defined in (S3) since the recursion associated to $\mathtt{QLSD}^\star$ can be written as

$$\tilde{\theta}_{k+1} = \tilde{\theta}_k - \gamma \sum_{i=1}^b \mathscr{S}_i \left[\mathscr{C}_i\left(H_i^\star(\tilde{\theta}_k, X_{k+1}^{(1,i)}), X_{k+1}^{(2,i)}\right), X_{k+1}^{(3,i)}\right] + \sqrt{2\gamma} Z_{k+1} \, , \qquad k \in \mathbb{N} \, , \tag{S33}$$

where, for any $i \in [b]$, $\mathscr{S}_i$ is defined in (S4). Therefore, we only need to verify that **H**S5 is satisfied with $\mathsf{X}^{(i)} = \tilde{\mathsf{X}}^{(i)} = \mathsf{X}_1^{(i)} \times \mathsf{X}_2 \times \mathsf{X}_3$, $\mathcal{X}^{(i)} = \tilde{\mathcal{X}}^{(i)} = \mathcal{X}_1^{(i)} \otimes \mathcal{X}_2 \otimes \mathcal{X}_3$, $\tilde{\nu}^{(i)} = \nu_1^{(i)} \otimes \nu_2 \otimes \nu_3$ for $i \in [b]$ and $\{F_i\}_{i=1}^b = \{F_i^\star\}_{i=1}^b = \{\mathscr{S}_i \circ \mathscr{C}_i \circ H_i^\star\}_{i=1}^b$. This is done in Appendix S2.2.

S2.1 Proof of Theorem 2

The Markov kernel associated with (S33) is given for any $(\theta, \mathsf{A}) \in \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$ by

$$Q_{\circledast,\gamma}(\theta, \mathsf{A}) = (4\pi\gamma)^{-d/2} \int_{\mathsf{A}\times\tilde{\mathsf{X}}^b} \exp\left(-\|\tilde{\theta} - \theta + \gamma \textstyle\sum_{i=1}^b F_i^\star(\theta, x^{(i)})\|^2/(4\gamma)\right) \mathrm{d}\tilde{\theta} \otimes_{i=1}^b \tilde{\nu}^{(i)}(\mathrm{d}x^{(i)}) \,. \tag{S34}$$

Then, the following non-asymptotic convergence result holds for QLSD*.

Theorem S6. *Assume **H**S1, **H**S2, **H**S4 and **H**S6. Then, for any probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, any step size $\gamma \in (0, \bar{\gamma}]$ where $\bar{\gamma}$ is defined in (S18), any $k \in \mathbb{N}$, we have*

$$W_2^2(\mu Q_{\circledast,\gamma}^k, \pi) \le (1 - \gamma\mathtt{m}/2)^k W_2^2(\mu, \pi) + \gamma B_{\circledast,\bar{\gamma}} + \gamma^2 A_{\circledast,\bar{\gamma}} (1 - \mathtt{m}\gamma/2)^{k-1} k \int_{\mathbb{R}^d} \|\theta - \theta^\star\|^2 \mu(\mathrm{d}\theta) \,,$$

where $Q_{\circledast,\gamma}$ is defined in (S34) and

$$\begin{aligned} B_{\circledast,\bar{\gamma}} &= (2d\mathtt{L}^2/\mathtt{m})\,(1/\mathtt{m} + 5\bar{\gamma})\left[1 + \bar{\gamma}\mathtt{L}^2/(2\mathtt{m}) + \bar{\gamma}^2\mathtt{L}^2/12\right] \\ &\quad + 4\mathtt{L}d\bar{\mathtt{M}} \max_{i\in[b]} \{\omega_i N_i + (\omega_i + 1)(N_i[1 - p_i]/p_i + A_{n_i,N_i})\}\,/\mathtt{m}^2 \\ A_{\circledast,\bar{\gamma}} &= \mathtt{L}\bar{\mathtt{M}} \max_{i\in[b]} \{\omega_i N_i + (\omega_i + 1)(N_i[1 - p_i]/p_i + A_{n_i,N_i})\} \,, \end{aligned} \tag{S35}$$

A_{n_i,N_i} being defined in (S36) for any $i \in [b]$.

Proof. Using Lemma S5, **H**S5 is satisfied and applying Theorem S4 completes the proof. □

S2.2 Supporting Lemmata

In this subsection, we derive two key lemmata in order to prove Theorem S6.

Lemma S4. *For any $i \in [b]$ and any sequence $\{a_j\}_{j=1}^{N_i} \in (\mathbb{R}^d)^{\otimes N_i}$ where $N_i \ge 2$, we have*

$$\int_{\mathsf{X}_1^{(i)}} \left\| \sum_{j=1}^{N_i} \left[\mathbf{1}_{x^{(1)}}(j) - \frac{n_i}{N_i}\right] a_j \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1)}) \le \frac{n_i(N_i - n_i)}{N_i(N_i - 1)} \sum_{j=1}^{N_i} \|a_j\|^2 \,.$$

Proof. Let $i \in [b]$ and $X^{(1,i)}$ distributed according to $\nu_1^{(i)}$. Since $\sum_{j=1}^{N_i} \mathbf{1}_{X^{(1,i)}}(j) = n_i$, we have

$$\sum_{l=1}^{N_i} \mathbf{1}_{X^{(1,i)}}(l) + \sum_{j\neq j'} \mathbf{1}_{X^{(1,i)}}(j)\mathbf{1}_{X^{(1,i)}}(j') = n_i^2 \,.$$

Integrating this equality over $\mathsf{X}_1^{(i)}$ gives

$$N_i \times \frac{n_i}{N_i} + N_i(N_i - 1) \times \int_{\mathsf{X}_1^{(i)}} [\mathbf{1}_{x^{(1,i)}}(1)\mathbf{1}_{x^{(1,i)}}(2)]\, \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) = n_i^2 \,.$$

Thus, we deduce that $\int_{\mathsf{X}_1^{(i)}} [\mathbf{1}_{x^{(1,i)}}(1)\mathbf{1}_{x^{(1,i)}}(2)] \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) = n_i(n_i - 1)\,[N_i(N_i - 1)]^{-1}$. In addition, using that

$$\int_{\mathsf{X}_1^{(i)}} \left(\mathbf{1}_{x^{(1,i)}}(j) - \frac{n_i}{N_i}\right)\left(\mathbf{1}_{x^{(1,i)}}(j') - \frac{n_i}{N_i}\right) \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) = \int_{\mathsf{X}_1^{(i)}} [\mathbf{1}_{x^{(1,i)}}(1)\mathbf{1}_{x^{(1,i)}}(2)] \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) - \frac{n_i^2}{N_i^2} \,,$$

we obtain

$$\int_{\mathsf{X}_1^{(i)}} \left\| \sum_{j=1}^{N_i} \left[\mathbf{1}_{x^{(1,i)}}(j) - \frac{n_i}{N_i}\right] a_j \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) = \frac{n_i(N_i - n_i)}{N_i^2} \left[\sum_{l=1}^{N_i} \|a_l\|^2 - \sum_{j\neq j'} \frac{\langle a_j, a_{j'}\rangle}{N_i - 1}\right]$$

$$= \frac{n_i(N_i - n_i)}{N_i^2(N_i - 1)} \left[N_i \sum_{l=1}^{N_i} \|a_l\|^2 - \left\| \sum_{l=1}^{N_i} a_l \right\|^2 \right] .$$

□

For any $i \in [b]$, denote

$$A_{n_i,N_i} = \frac{N_i(N_i - n_i)}{n_i(N_i - 1)} . \tag{S36}$$

The next lemma aims at controlling the variance of the global stochastic gradient considered in QLSD*, required to apply Theorem S4.

Lemma S5. *Assume **H**S2, **H**S4 and **H**S6. Then, for any $\theta \in \mathbb{R}^d$, we have*

$$\int_{\mathsf{X}^{(1:b)}} \left\| \sum_{i=1}^b \mathscr{S}_i \left[\mathscr{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right), x^{(3,i)} \right] - \nabla U(\theta) \right\|^2 \otimes_{i=1}^b \nu^{(i)}(\mathrm{d}x^{(i)})$$
$$\leq \bar{\mathtt{M}} \max_{i \in [b]} \{ \omega_i N_i + (\omega_i + 1)(N_i[1 - p_i]/p_i + A_{n_i,N_i}) \} \langle \theta - \theta^\star, \nabla U(\theta) - \nabla U(\theta^\star) \rangle ,$$

*where $\{H_i^\star\}_{i \in [b]}$ and $\{A_{n_i,N_i}\}_{i \in [b]}$ are defined in (S32) and (S36), respectively. Hence **H**S5 is satified with $\tilde{\mathtt{B}}^\star = 0$ and $\tilde{\mathtt{M}} = \bar{\mathtt{M}} \max_{i \in [b]} \{ \omega_i N_i + (\omega_i + 1)(N_i[1 - p_i]/p_i + A_{n_i,N_i}) \}$.*

Proof. Let $\theta \in \mathbb{R}^d$, using **H**S2 gives

$$\begin{aligned}
&\int_{\mathsf{X}^{(1:b)}} \left\| \sum_{i=1}^b \mathscr{S}_i \left[\mathscr{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right), x^{(3,i)} \right] - \nabla U(\theta) \right\|^2 \otimes_{i=1}^b \nu^{(i)}(\mathrm{d}x^{(i)}) \\
&= \int_{\mathsf{X}^{(1:b)}} \left\| \sum_{i=1}^b \mathscr{S}_i \left[\mathscr{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right), x^{(3,i)} \right] - \mathscr{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right) \right\|^2 \otimes_{i=1}^b \nu^{(i)}(\mathrm{d}x^{(i)}) \\
&+ \int_{\mathsf{X}_1^{(1:b)} \times \mathsf{X}_2^b} \left\| \sum_{i=1}^b \mathscr{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes b}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^b \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq \sum_{i=1}^b \left(\frac{1 - p_i}{p_i} \right) (\omega_i + 1) \int_{\mathsf{X}_1^{(i)}} \left\| H_i^\star(\theta, x^{(1,i)}) \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&+ \int_{\mathsf{X}_1^{(1:b)} \times \mathsf{X}_2^b} \left\| \sum_{i=1}^b \mathscr{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes b}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^b \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\leq \bar{\mathtt{M}} \sum_{i=1}^b \left(\frac{1 - p_i}{p_i} \right) (\omega_i + 1) N_i \langle \theta - \theta^\star, \nabla U_i(\theta) - \nabla U_i(\theta^\star) \rangle \\
&+ \int_{\mathsf{X}_1^{(1:b)} \times \mathsf{X}_2^b} \left\| \sum_{i=1}^b \mathscr{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes b}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^b \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) .
\end{aligned} \tag{S37}$$

Again using **H**S2, it follows that

$$\begin{aligned}
&\int_{\mathsf{X}_1^{(1:b)} \times \mathsf{X}_2^b} \left\| \sum_{i=1}^b \mathscr{C}_i \left(H_i^\star(\theta, x^{(1,i)}), x^{(2,i)} \right) - \nabla U(\theta) \right\|^2 \nu_2^{\otimes b}(\mathrm{d}x^{(2,1:b)}) \otimes_{i=1}^b \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&= \int_{\mathsf{X}_1^{(1:b)} \times \mathsf{X}_2^b} \left\| \sum_{i=1}^b \mathscr{C}_i \left(\frac{N_i}{n_i} \sum_{j=1}^{N_i} \mathbf{1}_{x^{(1,i)}}(j) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star) \right], x^{(2,i)} \right) \right. \\
&\left. - \sum_{i=1}^b \frac{N_i}{n_i} \sum_{j=1}^{N_i} \mathbf{1}_{x^{(1,i)}}(j) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star) \right] \right\|^2 \\
&+ \int_{\mathsf{X}_1^{(1:b)}} \left\| \sum_{i=1}^b \frac{N_i}{n_i} \sum_{j=1}^{N_i} \left(\mathbf{1}_{x^{(1,i)}}(j) - \frac{n_i}{N_i} \right) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star) \right] \right\|^2 \otimes_{i=1}^b \nu_1^{(i)}(\mathrm{d}x^{(1,i)})
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{b} \omega_i \left(\frac{N_i}{n_i}\right)^2 \int_{\mathsf{X}_1^{(i)}} \left\| \sum_{j=1}^{N_i} \mathbf{1}_{x^{(1,i)}}(j) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right] \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&+ \sum_{i=1}^{b} \left(\frac{N_i}{n_i}\right)^2 \int_{\mathsf{X}_1^{(i)}} \left\| \sum_{j=1}^{N} \left(\mathbf{1}_{x^{(1,i)}}(j) - \frac{n_i}{N_i}\right) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right] \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&= \sum_{i=1}^{b} \omega_i \left\|\nabla U_i(\theta) - \nabla U_i(\theta^\star)\right\|^2 \\
&+ \sum_{i=1}^{b} (\omega_i + 1) \left(\frac{N_i}{n_i}\right)^2 \int_{\mathsf{X}_1^{(i)}} \left\| \sum_{j=1}^{N_i} \left(\mathbf{1}_{x^{(1,i)}}(j) - \frac{n_i}{N_i}\right) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right] \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \,. \qquad \text{(S38)}
\end{aligned}
$$

Using Lemma S4 combined with **H**S6 yields, for any $i \in [b]$,

$$
\begin{aligned}
&\int_{\mathsf{X}_1^{(i)}} \left\| \textstyle\sum_{j=1}^{N_i} \left(\mathbf{1}_{x^{(1,i)}}(j) - n_i/N_i\right) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right] \right\|^2 \nu_1^{(i)}(\mathrm{d}x^{(1,i)}) \\
&\qquad\qquad \leq \frac{n_i(N_i - n_i)}{N_i(N_i - 1)} \bar{\mathtt{M}} \left\langle \theta - \theta^\star, \nabla U_i(\theta) - \nabla U_i(\theta^\star) \right\rangle \,. \qquad \text{(S39)}
\end{aligned}
$$

In addition, Jensen inequality implies, for any $i \in [b]$, that

$$
\left\|\nabla U_i(\theta) - \nabla U_i(\theta^\star)\right\|^2 \leq N_i \sum_{j=1}^{N_i} \left\|\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\theta^\star)\right\|^2 \,,
$$

and therefore, using **H**S6, we have for any $i \in [b]$,

$$
\left\|\nabla U_i(\theta) - \nabla U_i(\theta^\star)\right\|^2 \leq \bar{\mathtt{M}} N_i \left\langle \nabla U_i(\theta) - \nabla U_i(\theta^\star), \theta - \theta^\star \right\rangle \,. \qquad \text{(S40)}
$$

Injecting (S39) and (S40) into (S38) and using (S37) conclude the proof. □

S3 PROOF OF Theorem 3

S3.1 Problem formulation.

We assume here that U is still of the form (1) and that there exist $\{N_i \in \mathbb{N}^*\}_{i\in[b]}$ such that for any $i \in [b]$, there exist N_i functions $\{U_{i,j} : \theta \in \mathbb{R}^d \to \mathbb{R}\}_{j\in[N_i]}$ such that for any $\theta \in \mathbb{R}^d$,

$$
U_i(\theta) = \sum_{j=1}^{N_i} U_{i,j}(\theta) \,.
$$

In all this section, we assume for any $i \in [b]$ that $n_i \in \mathbb{N}^*$, $n_i \leq N_i$ is fixed. Recall that $\wp_N$ denotes the power set of $[N]$ and

$$
\wp_{N,n} = \{x \in \wp_N \,:\, \mathrm{card}(x) = n\} \,.
$$

In addition, we set in this section $\nu_1^{(i)}$ as the uniform distribution on $\wp_{N_i,n_i}$. We consider the family of measurable functions $\{G_i : \mathbb{R}^d \times \mathbb{R}^d \times \wp_{N_i} \to \mathbb{R}^d\}_{i\in[b]}$, defined for any $i \in [b]$, $\theta \in \mathbb{R}^d$, $\zeta \in \mathbb{R}^d$, $x \in \wp_{N_i,n_i}$ by

$$
G_i(\theta, \zeta; x) = \frac{N_i}{n_i} \sum_{j=1}^{N_i} \mathbf{1}_x(j) \left[\nabla U_{i,j}(\theta) - \nabla U_{i,j}(\zeta)\right] + \nabla U_i(\zeta) \,. \qquad \text{(S41)}
$$

For ease of reading, we formalise more precisely the recursion associated with $\mathtt{QLSD}^{++}$ under **H**S2. Let $(X_k^{(1,1)}, \ldots, X_k^{(1,b)})_{k\in\mathbb{N}^*}$ and $(X_k^{(2,1)}, \ldots, X_k^{(2,b)})_{k\in\mathbb{N}^*}$ be two independent i.i.d. sequences with distribution

$\otimes_{i=1}^{b}\nu_1^{(i)}$ and $\nu_2^{\otimes b}$. Let $(Z_k)_{k\in\mathbb{N}^*}$ be an i.i.d. sequence of d-dimensional standard Gaussian random variables independent of $(X_k^{(1,1)},\dots,X_k^{(1,b)})_{k\in\mathbb{N}^*}$ and $(X_k^{(2,1)},\dots,X_k^{(2,b)})_{k\in\mathbb{N}^*}$. Similarly as before, we consider the partial device participation context where at each communication round $k\geq 1$, each client has a probability $p_i\in(0,1]$ of participating, independently from other clients. In other words, there exists a sequence $(X_k^{(3,1)},\cdots,X_k^{(3,b)})_{k\in\mathbb{N}^*}$ of i.i.d. random variables distributed according $\nu_3=\mathrm{Uniform}((0,1])$, such that for any $k\geq 1$ and $i\in[b]$, client i is active at step k if $X_k^{(3,i)}\leq p_i$. We denote $\mathcal{A}_{k+1}=\{i\in[b];X_{k+1}^{(3,i)}\leq p_i\}$ the set of active clients at round k. For ease of notation, denote for any $k\in\mathbb{N}^*$, $X_k^{(1)}=(X_k^{(1,1)},\dots,X_k^{(1,b)})$, $X_k^{(2)}=(X_k^{(2,1)},\dots,X_k^{(2,b)})$, $X_k^{(3)}=(X_k^{(3,1)},\dots,X_k^{(3,b)})$ and $X_k=(X_k^{(1)},X_k^{(2)},X_k^{(3)})$. Let $l\in\mathbb{N}^*$, $\gamma\in(0,\bar\gamma]$ and $\alpha\in(0,\bar\alpha]$ for $\bar\gamma,\bar\alpha>0$. Given $\Theta_0=(\theta_0,\zeta_0,\{\eta_0^{(i)}\}_{i\in[b]})\in\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^{db}$, with $\zeta_0=\theta_0$, we recursively define the sequence $(\Theta_k)_{k\in\mathbb{N}}=(\theta_k,\zeta_k,\{\eta_k^{(i)}\}_{i\in[b]})_{k\in\mathbb{N}}$, for any $k\in\mathbb{N}$ as

$$\theta_{k+1}=\theta_k-\gamma\tilde G(\Theta_k;X_{k+1})+\sqrt{2\gamma}Z_{k+1}\,, \tag{S42}$$

where

$$\tilde G(\Theta_k;X_{k+1})=\sum_{i=1}^{b}\left[\mathscr{S}_i\left(\mathscr{C}_i\left\{G_i\left(\theta_k,\zeta_k;X_{k+1}^{(1,i)}\right)-\eta_k^{(i)};X_{k+1}^{(2,i)}\right\},X_{k+1}^{(3,i)}\right)+\eta_k^{(i)}\right], \tag{S43}$$

$$\zeta_{k+1}=\begin{cases}\theta_{k+1}\,, & \text{if } k+1\equiv 0 \pmod l\,,\\ \zeta_k\,, & \text{otherwise}\,,\end{cases} \tag{S44}$$

and for any $i\in[b]$,

$$\eta_{k+1}^{(i)}=\eta_k^{(i)}+\alpha\mathscr{S}_i\left(\mathscr{C}_i\left\{G_i\left(\theta_k,\zeta_k;X_{k+1}^{(1,i)}\right)-\eta_k^{(i)};X_{k+1}^{(2,i)}\right\},X_{k+1}^{(3,i)}\right)\,. \tag{S45}$$

Since $\texttt{QLSD}^{++}$ involves auxiliary variables gathered with $(\theta_k)_{k\in\mathbb{N}}$ in $(\Theta_k)_{k\in\mathbb{N}}$, we cannot follow the same proof as for $\texttt{QLSD}^\star$ by verifying **H**S5 and then applying Theorem S4. Instead, we will adapt the proof Theorem S4 and in particular Lemma S2 and bound the variance associated to the stochastic gradient defined in (S43). Once this variance term will be tackled, the proof of Theorem 3 will follow the same lines as the proof of Theorem S4 upon using specific moment estimates for $\texttt{QLSD}^{++}$. In the next section, we focus on these two goals: we provide uniform bounds in the number of iterations k on the variance of the sequence of stochastic gradients associated with $\texttt{QLSD}^{++}$, $(\mathbb{E}[\|\tilde G_i(\Theta_k,X_{k+1})-\nabla U(\theta_k)\|^2])_{k\in\mathbb{N}}$ for any $i\in[b]$, and $(\mathbb{E}[\|\theta_k-\theta^\star\|^2])_{k\in\mathbb{N}}$, see Proposition S8 and Corollary S7. To this end, a key ingredient is the design of an appropriate Lyapunov function defined in (S57).

S3.2 Uniform bounds on the stochastic gradients and moment estimates for $\texttt{QLSD}^{++}$

Consider the filtration associated with $(\Theta_k)_{k\in\mathbb{N}}$ defined by $\mathcal{G}_0=\sigma(\Theta_0)$ and for $k\in\mathbb{N}^*$,

$$\mathcal{G}_k=\sigma(\Theta_0,(X_{\tilde k})_{\tilde k\leq k},(Z_{\tilde k})_{\tilde k\leq k})\,.$$

We denote for any $i\in[b]$, $\theta,\zeta\in\mathbb{R}^d$,

$$\Delta_i(\theta,\zeta)=\nabla U_i(\theta)-\nabla U_i(\zeta)\,. \tag{S46}$$

Similarly, we consider, for any $i\in[b]$, $j\in[N]$, $\theta,\zeta\in\mathbb{R}^d$,

$$\Delta_{i,j}(\theta,\zeta)=\nabla U_{i,j}(\theta)-\nabla U_{i,j}(\zeta)\,. \tag{S47}$$

The following lemma provides a first upper bound on the variance of the stochastic gradients used in $\texttt{QLSD}^{++}$.

Lemma S6. *Assume **H**S1, **H**S2, **H**S4 and **H**S6 and let $\gamma\in(0,\bar\gamma]$, $\alpha\in(0,\bar\alpha]$ for some $\bar\gamma,\bar\alpha>0$. Then, for any $s\in\mathbb{N}$, $r\in\{0,\dots,l-1\}$, we have*

$$\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\tilde G(\Theta_{sl+r};X_{sl+r+1})-\nabla U(\theta_{sl+r})\right\|^2\right]\leq\left[2\sum_{i=1}^{b}\frac{\mathtt{M}_i^2}{p_i}(\omega_i+1-p_i)+\left(\frac{\omega_i+1}{p_i}\right)A_{n_i,N_i}\bar{\mathtt{M}}\mathtt{M}_i\right]\|\theta_{sl+r}-\theta^\star\|^2$$
$$+\left[2\sum_{i=1}^{b}(\omega_i+1-p_i)/p_i\right]\left\|\nabla U_i(\theta^\star)-\eta_{sl+r}^{(i)}\right\|^2+2\bar{\mathtt{M}}\sum_{i=1}^{b}\left[\left(\frac{\omega_i+1}{p_i}\right)A_{n_i,N_i}\mathtt{M}_i\right]\|\theta_{sl}-\theta^\star\|^2\,,$$

where $(\Theta_{\tilde k})_{\tilde k\in\mathbb{N}}=(\theta_{\tilde k},\zeta_{\tilde k},\{\eta_{\tilde k}^{(i)}\}_{i\in[b]})_{\tilde k\in\mathbb{N}}$, $\tilde G$ and $A_{n,N}$ are defined in (S42), (S44), (S45), (S43) and (S36), respectively.

Proof. Let $s \in \mathbb{N}$ and $r \in \{0, \dots, l-1\}$. Using **H**S2, (S46) and (S47), we have

$$
\begin{aligned}
&\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\tilde{G}(\Theta_{sl+r}; X_{sl+r+1}) - \nabla U(\theta_{sl+r})\right\|^2\right] \\
&= \sum_{i=1}^{b} \mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\| \mathscr{S}_i\left(\mathscr{C}_i\left\{G_i\left(\theta_{sl+r}, \zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \eta_{sl+r}^{(i)}; X_{sl+r+1}^{(2,i)}\right\}, X_{sl+r+1}^{(3,i)}\right)\right.\right. \\
&\left.\left. - \mathscr{C}_i\left\{G_i\left(\theta_{sl+r}, \zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \eta_{sl+r}^{(i)}; X_{sl+r+1}^{(2,i)}\right\}\right\|^2\right] \\
&+ \sum_{i=1}^{b} \mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\mathscr{C}_i\left\{G_i\left(\theta_{sl+r}, \zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \eta_{sl+r}^{(i)}; X_{sl+r+1}^{(2,i)}\right\} + \eta_{sl+r}^{(i)} - \nabla U_i(\theta_{sl+r})\right\|^2\right] \\
&\leq \sum_{i=1}^{b}\left(\frac{1-p_i}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\mathscr{C}_i\left\{G_i\left(\theta_{sl+r}, \zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \eta_{sl+r}^{(i)}; X_{sl+r+1}^{(2,i)}\right\}\right\|^2\right] \\
&+ \sum_{i=1}^{b}\omega_i \mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|G_i\left(\theta_{sl+r}, \zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \eta_{sl+r}^{(i)}\right\|\right] + \sum_{i=1}^{b}\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|G_i\left(\theta_{sl+r}, \zeta_k; X_{sl+r+1}^{(1,i)}\right) - \nabla U_i(\theta_{sl+r})\right\|^2\right] \\
&\leq \sum_{i=1}^{b}\left(\frac{\omega_i + 1 - p_i}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|G_i\left(\theta_{sl+r}, \zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \eta_{sl+r}^{(i)}\right\|^2\right] \\
&+ \sum_{i=1}^{b}\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|G_i\left(\theta_{sl+r}, \zeta_{sl+r}; X_{sl+r+1}^{(1,i)}\right) - \nabla U_i(\theta_{sl+r})\right\|^2\right] \\
&\leq \sum_{i=1}^{b}\left(\frac{\omega_i + 1}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\frac{N_i}{n_i}\sum_{j=1}^{N_i}\left\{\mathbf{1}_{X_{sl+r+1}^{(1,i)}}(j)\Delta_{i,j}(\theta_{sl+r}, \zeta_{sl+r})\right\} - \Delta_i(\theta_{sl+r}, \zeta_{sl+r})\right\|^2\right] \\
&+ \sum_{i=1}^{b}\left(\frac{\omega_i + 1 - p_i}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\nabla U_i(\theta_{sl+r}) - \eta_{sl+r}^{(i)}\right\|^2\right] \\
&\leq \sum_{i=1}^{b}\left(\frac{\omega_i + 1}{p_i}\right)\frac{N_i(N_i - n_i)}{n_i(N_i - 1)}\bar{\mathtt{M}}\langle\theta_{sl+r} - \zeta_{sl+r}, \nabla U_i(\theta_{sl+r}) - \nabla U_i(\zeta_{sl+r})\rangle \\
&+ \sum_{i=1}^{b}\left(\frac{\omega_i + 1 - p_i}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r}}\left[\left\|\nabla U_i(\theta_{sl+r}) - \eta_{sl+r}^{(i)}\right\|^2\right],
\end{aligned}
$$

where the last line follows from **H**S6 and Lemma S4. The proof is concluded by using the Cauchy-Schwarz inequality, **H**S1 and $\zeta_{sl+r} = \theta_{sl}$. □

The two following lemmas aim at controlling the terms that appear in Lemma S6.

Lemma S7. *Assume **H**S1, **H**S2, **H**S4 and **H**S6, and let $\gamma \in (0, \bar{\gamma}]$, $\alpha \in (0, \bar{\alpha}]$ for some $\bar{\gamma}, \bar{\alpha} > 0$. Then, for any $s \in \mathbb{N}$ and $r \in [l]$, we have*

$$
\begin{aligned}
\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\theta_{sl+r} - \theta^\star\right\|^2\right] &\leq \left(1 - 2\gamma m + \gamma^2 B_{\mathbf{n},\mathbf{N}}\right)\left\|\theta_{sl+r-1} - \theta^\star\right\|^2 \\
&+ \gamma^2\left[2\sum_{i=1}^{b}(\omega_i + 1 - p_i)/p_i\right]\left\|\nabla U_i(\theta^\star) - \eta_{sl+r-1}^{(i)}\right\|^2 + 2\bar{\mathtt{M}}\gamma^2\sum_{i=1}^{b}\left[\left(\frac{\omega_i + 1}{p_i}\right)A_{n_i,N_i}\mathtt{M}_i\right]\left\|\theta_{sl} - \theta^\star\right\|^2 + 2\gamma d\,,
\end{aligned}
$$

where

$$
B_{\mathbf{n},\mathbf{N}} = 2\sum_{i=1}^{b}\left\{\frac{\mathtt{M}_i^2}{p_i}(\omega_i + 1 - p_i) + \left(\frac{\omega_i + 1}{p_i}\right)A_{n_i,N_i}\bar{\mathtt{M}}\mathtt{M}_i\right\} + \mathtt{L}^2\,, \tag{S48}
$$

$(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta_{\tilde{k}}^i\}_{i\in[b]})_{\tilde{k}\in\mathbb{N}}$ and $A_{n,N}$ are defined in (S42), (S44), (S45) and (S36) respectively.

Proof. Let $s \in \mathbb{N}$ and $r \in [l]$. Using (S42) and **H**S2, it follows

$$\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\|\theta_{sl+r}-\theta^\star\|^2\right] = \|\theta_{sl+r-1}-\theta^\star\|^2 + 2\gamma d - 2\gamma\langle\nabla U(\theta_{sl+r-1}), \theta_{sl+r-1}-\theta^\star\rangle + \gamma^2\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\tilde{G}(\Theta_{sl+r-1}; X_{sl+r})\right\|^2\right] . \quad \text{(S49)}$$

Using **H**S2 and (S41)-(S43), we have

$$\begin{aligned}
&\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\tilde{G}(\Theta_{sl+r-1}; X_{sl+r})\right\|^2\right] \\
&= \sum_{i=1}^{b}\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\mathscr{S}_i\left(\mathscr{C}_i\left\{G_i\left(\theta_{sl+r-1},\zeta_{sl+r-1}; X_{sl+r}^{(1,i)}\right) - \eta_{sl+r-1}^{(i)}; X_{sl+r}^{(2,i)}\right\}, X_{sl+r}^{(3,i)}\right)\right.\right. \\
&\left.\left. - \mathscr{C}_i\left\{G_i\left(\theta_{sl+r-1},\zeta_{sl+r-1}; X_{sl+r}^{(1,i)}\right) - \eta_{sl+r-1}^{(i)}; X_{sl+r}^{(2,i)}\right\}\right\|^2\right] \\
&+ \mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\sum_{i=1}^{b}\mathscr{C}_i\left\{G_i\left(\theta_{sl+r-1},\zeta_{sl+r-1}; X_{sl+r}^{(1,i)}\right) - \eta_{sl+r-1}^{(i)}; X_{sl+r}^{(2,i)}\right\} + \eta_{sl+r-1}^{(i)}\right\|^2\right] \\
&\leq \sum_{i=1}^{b}\left(\frac{\omega_i+1-p_i}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|G_i\left(\theta_{sl+r-1},\zeta_{sl+r-1}; X_{sl+r}^{(1,i)}\right) - \eta_{sl+r-1}^{(i)}\right\|^2\right] \\
&+ \sum_{i=1}^{b}\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\frac{N_i}{n_i}\sum_{j=1}^{N_i}\left\{\mathbf{1}_{X_{sl+r}^{(1,i)}}(j)\Delta_{i,j}(\theta_{sl+r-1},\zeta_{sl+r-1})\right\} - \Delta_i(\theta_{sl+r-1},\zeta_{sl+r-1})\right\|^2\right] + \|\nabla U(\theta_{sl+r-1})\|^2 \\
&= \sum_{i=1}^{b}\left(\frac{\omega_i+1}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\frac{N_i}{n_i}\sum_{j=1}^{N_i}\left\{\mathbf{1}_{X_{sl+r}^{(1,i)}}(j)\Delta_{i,j}(\theta_{sl+r-1},\zeta_{sl+r-1})\right\} - \Delta_i(\theta_{sl+r-1},\zeta_{sl+r-1})\right\|^2\right] \\
&+ \sum_{i=1}^{b}\left(\frac{\omega_i+1-p_i}{p_i}\right)\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\nabla U_i(\theta_{sl+r-1}) - \eta_{sl+r-1}^{(i)}\right\|^2\right] + \|\nabla U(\theta_{sl+r-1})\|^2 \\
&\leq \sum_{i=1}^{b}\left(\frac{\omega_i+1}{p_i}\right)\frac{N_i(N_i-n_i)}{n_i(N_i-1)}\bar{\mathtt{M}}\langle\theta_{sl+r-1}-\zeta_{sl+r-1}, \nabla U_i(\theta_{sl+r-1}) - \nabla U_i(\zeta_{sl+r-1})\rangle \\
&+ \sum_{i=1}^{b}\left(\frac{\omega_i+1-p_i}{p_i}\right)\left\|\nabla U_i(\theta_{sl+r-1}) - \eta_{sl+r-1}^{(i)}\right\|^2 + \|\nabla U(\theta_{sl+r-1})\|^2 , \quad \text{(S50)}
\end{aligned}$$

where the last line follows from **H**S6 and Lemma S4. The proof is concluded by injecting (S50) into (S49), using the Cauchy-Schwarz inequality, $\nabla U(\theta^\star) = 0$, **H**S1 and $\zeta_{sl+r-1} = \theta_{sl}$. □

Lemma S8. *Assume **H**S1, **H**S2, **H**S4 and **H**S6. Let $\gamma \in (0, \bar{\gamma}]$ for some $\bar{\gamma} > 0$ and $\alpha \in (0, 1/(\max_{i\in[b]}\omega_i + 1)]$. Then, for any $s \in \mathbb{N}$ and $r \in [l]$, we have*

$$\begin{aligned}
\sum_{i=1}^{b}\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\left\|\nabla U_i(\theta^\star) - \eta_{sl+r}^{(i)}\right\|^2\right] &\leq (1-\alpha)\sum_{i=1}^{b}\left\|\nabla U_i(\theta^\star) - \eta_{sl+r-1}^{(i)}\right\|^2 \\
&+ \alpha C_{\mathbf{n},\mathbf{N}}\left\|\theta_{sl+r-1}-\theta^\star\right\|^2 + 2\alpha\left[\sum_{i=1}^{b}A_{n_i,N_i}\bar{\mathtt{M}}\mathtt{M}_i\right]\left\|\theta_{sl}-\theta^\star\right\|^2 ,
\end{aligned}$$

where

$$C_{\mathbf{n},\mathbf{N}} = 2\sum_{i=1}^{b}\left\{A_{n_i,N_i}\bar{\mathtt{M}}\mathtt{M}_i + \mathtt{M}_i^2\right\} , \quad \text{(S51)}$$

$(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta_{\tilde{k}}^i\}_{i\in[b]})_{\tilde{k}\in\mathbb{N}}$ *and* $A_{n,N}$ *are defined in* (S42), (S44), (S45) *and* (S36), *respectively.*

Proof. Let $s \in \mathbb{N}$ and $r \in [l]$. Then, it follows

$$\sum_{i=1}^{b} \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| \nabla U_i(\theta^\star) - \eta_{sl+r}^{(i)} \right\|^2 \right] = \sum_{i=1}^{b} \left\| \nabla U_i(\theta^\star) - \eta_{sl+r-1}^{(i)} \right\|^2 + \sum_{i=1}^{b} \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| \eta_{sl+r}^{(i)} - \eta_{sl+r-1}^{(i)} \right\|^2 \right] + 2 \sum_{i=1}^{b} \langle \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\eta_{sl+r}^{(i)} - \eta_{sl+r-1}^{(i)} \right], \eta_{sl+r-1}^{(i)} - \nabla U_i(\theta^\star) \rangle \,. \tag{S52}$$

Using (S45) and **H**S2, we have for any $i \in [b]$,

$$\mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| \eta_{sl+r}^{(i)} - \eta_{sl+r-1}^{(i)} \right\|^2 \right] \leq \alpha^2 (\omega_i + 1) \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| G_i \left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)} \right) - \eta_{sl+r-1}^{(i)} \right\|^2 \right] , \tag{S53}$$

$$\mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\eta_{sl+r}^{(i)} - \eta_{sl+r-1}^{(i)} \right] = \alpha \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[G_i \left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)} \right) - \eta_{sl+r-1}^{(i)} \right] . \tag{S54}$$

Plugging (S53) and (S54) into (S52) yields

$$\begin{aligned} \sum_{i=1}^{b} \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| \nabla U_i(\theta^\star) - \eta_{sl+r}^{(i)} \right\|^2 \right] &\leq \sum_{i=1}^{b} \left\| \nabla U_i(\theta^\star) - \eta_{sl+r-1}^{(i)} \right\|^2 \\ &+ \alpha^2 \sum_{i=1}^{b} (\omega_i + 1) \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| G_i \left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)} \right) - \eta_{sl+r-1}^{(i)} \right\|^2 \right] \\ &+ 2\alpha \sum_{i=1}^{b} \langle \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[G_i \left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)} \right) - \eta_{sl+r-1}^{(i)} \right], \eta_{sl+r-1}^{(i)} - \nabla U_i(\theta^\star) \rangle \,. \end{aligned}$$

Using for any $i \in [b]$ $\alpha(1 + \omega_i) \leq 1$ and the fact, for any $a, b, c \in \mathbb{R}^d$, that $\|a - c\|^2 + 2\langle (a - c), (c - b) \rangle = \|a - b\|^2 - \|c - b\|^2$, we have

$$\begin{aligned} \sum_{i=1}^{b} \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| \nabla U_i(\theta^\star) - \eta_{sl+r}^{(i)} \right\|^2 \right] &\leq (1 - \alpha) \sum_{i=1}^{b} \left\| \nabla U_i(\theta^\star) - \eta_{sl+r-1}^{(i)} \right\|^2 \\ &+ \alpha \sum_{i=1}^{b} \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| G_i \left(\theta_{sl+r-1}, \zeta_{sl+r}; X_{sl+r}^{(1,i)} \right) - \nabla U_i(\theta^\star) \right\|^2 \right] . \end{aligned} \tag{S55}$$

Using (S41), **H**S6 and Lemma S4, it follows

$$\begin{aligned} &\sum_{i=1}^{b} \mathbb{E}^{\mathcal{G}_{sl+r-1}} \left[\left\| G_i \left(\theta_{sl+r-1}, \zeta_{sl+r-1}; X_{sl+r}^{(1,i)} \right) - \nabla U_i(\theta^\star) \right\|^2 \right] \\ &\leq \sum_{i=1}^{b} \frac{N_i (N_i - n_i)}{n_i (N_i - 1)} \bar{\mathtt{M}} \langle \theta_{sl+r-1} - \zeta_{sl+r-1}, \nabla U_i(\theta_{sl+r-1}) - \nabla U_i(\zeta_{sl+r-1}) \rangle \\ &\quad + \sum_{i=1}^{b} \| \nabla U_i(\theta_{sl+r-1}) - \nabla U_i(\theta^\star) \|^2 \,. \end{aligned} \tag{S56}$$

The proof is concluded by plugging (S56) into (S55), using the Cauchy-Schwarz inequality, **H**S1 and $\zeta_{sl+r-1} = \theta_{sl}$. □

Lemma S7 and Lemma S8 involve two dependent terms which prevents us from using a straightforward induction. To cope with this issue, we consider a Lyapunov function $\psi : \mathbb{R}^d \times \mathbb{R}^{bd} \to \mathbb{R}$ defined, for any $\theta \in \mathbb{R}^d$ and $\eta = (\eta^{(1)}, \ldots, \eta^{(b)})^\top \in \mathbb{R}^{bd}$ by

$$\psi(\theta, \eta) = \|\theta - \theta^\star\|^2 + (3/\alpha) \max_{i \in [b]} \{(\omega_i + 1 - p_i)/p_i\} \gamma^2 \sum_{i=1}^{b} \left\| \nabla U_i(\theta^\star) - \eta^{(i)} \right\|^2 . \tag{S57}$$

The following lemma provides an upper bound on this Lyapunov function. Define for $\alpha > 0$,

$$\bar{\gamma}_{\alpha,1} = \mathtt{m}^{-1}[\{\mathtt{m}^2(B_{\mathbf{n},\mathbf{N}} + 3\omega C_{\mathbf{n},\mathbf{N}})^{-1}\} \wedge \{\alpha/3\}]\,, \tag{S58}$$

where $B_{\mathbf{n},\mathbf{N}}$ and $C_{\mathbf{n},\mathbf{N}}$ are defined in (S48) and (S51) respectively.

Lemma S9. *Assume **HS1**, **HS2**, **HS4** and **HS6**. Let $\alpha \in (0, 1/(1+\max_{i\in[b]}\omega_i)]$, $\gamma \in (0, \bar{\gamma}_{\alpha,1}]$. Then, for any $s \in \mathbb{N}$ and $r \in [l]$, we have*

$$\begin{aligned}\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\psi(\theta_{sl+r}, \eta_{sl+r})\right] &\leq (1-\gamma\mathtt{m})\,\psi(\theta_{sl+r-1}, \eta_{sl+r-1}) \\ &\quad + 8\bar{\mathtt{M}}\gamma^2 \max_{i\in[b]}\{(\omega_i+1)/p_i\}\sum_{i=1}^{b} A_{n_i,N_i}\mathtt{M}_i \left\|\theta_{sl} - \theta^\star\right\|^2 + 2\gamma d\,,\end{aligned}$$

where ψ is defined in (S57) and $(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta^i_{\tilde{k}}\}_{i\in[b]})_{\tilde{k}\in\mathbb{N}}$ and $A_{n,N}$ are defined in (S42), (S44), (S45) and (S36), respectively.

Proof. Let $s \in \mathbb{N}$ and $r \in [l]$. Using Lemma S7 and Lemma S8, we have

$$\begin{aligned}&\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\psi(\theta_{sl+r}, \eta_{sl+r})\right] \\ &\quad\leq \left(1 - 2\gamma\mathtt{m} + \gamma^2\left[B_{\mathbf{n},\mathbf{N}} + 3\omega C_{\mathbf{n},\mathbf{N}}\right]\right)\left\|\theta_{sl+r-1} - \theta^\star\right\|^2 \\ &\qquad + \left[(2/3)\alpha + (1-\alpha)\right](3\gamma^2/\alpha)\max_{i\in[b]}\{(\omega_i + 1 - p_i)/p_i\}\sum_{i=1}^{b}\left\|\nabla U_i(\theta^\star) - \eta^{(i)}_{sl+r-1}\right\|^2 \\ &\qquad + 8\bar{\mathtt{M}}\gamma^2\max_{i\in[b]}\{(\omega_i+1)/p_i\}\sum_{i=1}^{b} A_{n_i,N_i}\mathtt{M}_i\left\|\theta_{sl} - \theta^\star\right\|^2 + 2\gamma d\,.\end{aligned}$$

Since $\gamma \leq \bar{\gamma}_{\alpha,1}$ with $\bar{\gamma}_{\alpha,1}$ given in (S58), it follows that

$$\begin{aligned}1 - 2\gamma\mathtt{m} + \gamma^2\left[B_{\mathbf{n},\mathbf{N}} + 3\omega C_{\mathbf{n},\mathbf{N}}\right] &\leq 1 - \gamma\mathtt{m} \\ (2/3)\alpha + (1-\alpha) &\leq 1 - \gamma\mathtt{m}\,.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}^{\mathcal{G}_{sl+r-1}}\left[\psi(\theta_{sl+r}, \eta_{sl+r})\right] &\leq (1-\gamma\mathtt{m})\,\psi(\theta_{sl+r-1}, \eta_{sl+r-1}) \\ &\quad + 8\bar{\mathtt{M}}\gamma^2 \max_{i\in[b]}\{(\omega_i+1)/p_i\}\sum_{i=1}^{b} A_{n_i,N_i}\mathtt{M}_i \left\|\theta_{sl} - \theta^\star\right\|^2 + 2\gamma d\,.\end{aligned}$$

□

Lemma S10. *Let $j \in \mathbb{N}^*$ and fix $\gamma > 0$ such that*

$$\gamma \leq \frac{\mathtt{m}}{16 j\bar{\mathtt{M}}\gamma^2 \max_{i\in[b]}\{(\omega_i+1)/p_i\}\sum_{i=1}^{b} A_{n_i,N_i}\mathtt{M}_i} \wedge \frac{1}{\mathtt{m}}\,.$$

Then,

$$(1-\gamma\mathtt{m})^j + 8j\gamma^2\bar{\mathtt{M}}\max_{i\in[b]}\{(\omega_i+1)/p_i\}\sum_{i=1}^{b} A_{n_i,N_i}\mathtt{M}_i \leq 1 - \gamma\mathtt{m}/2\,,$$

where $A_{n,N}$ is defined in (S36).

Proof. The proof is straightforward using $(1-\gamma\mathtt{m})^j \leq 1 - \gamma\mathtt{m}$. □

We have the following corollary regarding the Lyapunov function defined in (S57).

Denote for $\alpha > 0$,

$$\bar{\gamma}_{\alpha,2} = \bar{\gamma}_{\alpha,1} \wedge \left[\mathtt{m}/\{16 l\bar{\mathtt{M}}\max_{i\in[b]}\{(\omega_i+1)/p_i\}\textstyle\sum_{i=1}^{b} A_{n_i,N_i}\mathtt{M}_i\}\right]^{1/3}, \tag{S59}$$

where $\bar{\gamma}_{\alpha,1}$ is given in (S58).

Corollary S7. *Assume **HS1**, **HS2**, **HS4** and **HS6**. Let $\alpha \in (0, 1/(1+\max_{i\in[b]}\omega_i)]$ and $\gamma \in (0, \bar{\gamma}_{\alpha,2}]$. Then, for any $s \in \mathbb{N}$ and $r \in \{0,\dots,l-1\}$ we have*

$$\mathbb{E}^{\mathcal{G}_{sl}}\left[\psi(\theta_{(s+1)l-r}, \eta_{(s+1)l-r}\right] \leq (1-\gamma\mathtt{m}/2)\,\psi(\theta_{sl},\eta_{sl}) + 2\gamma(l-r)d\,,$$

where ψ is defined in (S57) and $(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta^i_{\tilde{k}}\}_{i\in[b]})_{\tilde{k}\in\mathbb{N}}$ is defined in (S42), (S44), (S45).

Proof. The proof follows from a straightforward induction of Lemma S9 combined with Lemma S10. □

We are now ready to control explicitly the variance of the stochastic gradient defined in (S43).

Proposition S8. *Assume **HS1**, **HS2**, **HS4** and **HS6**. Let $\alpha \in (0, 1/(1+\max_{i\in[b]}\omega_i)]$ and $\gamma \in (0, \bar{\gamma}_{\alpha,2}]$, where $\bar{\gamma}_{\alpha,2}$ is defined in (S59). Then, for any $k = sl + r$ with $s \in \mathbb{N}$, $r \in \{0,\dots,l-1\}$, $\theta_0 \in \mathbb{R}^d$ and $\eta_0 = (\eta_0^{(1)},\dots,\eta_0^{(b)})^\top \in \mathbb{R}^{db}$, we have*

$$\mathbb{E}\left[\left\|\tilde{G}(\Theta_{sl+r}; X_{sl+r+1}) - \nabla U(\theta_k)\right\|^2\right] \leq (1-\gamma\mathtt{m}/2)^s D_{\mathbf{n},\mathbf{N}}\psi(\theta_0,\eta_0) + 4ldD_{\mathbf{n},\mathbf{N}}/\mathtt{m} + \left[2\sum_{i=1}^b(\omega_i+1-p_i)/p_i\right](1-\alpha)^k \sum_{i=1}^b \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2\right],$$

where

$$D_{\mathbf{n},\mathbf{N}} = \left[2\sum_{i=1}^b \frac{\mathtt{M}_i^2}{p_i}(\omega_i+1-p_i) + \left(\frac{\omega_i+1}{p_i}\right)A_{n_i,N_i}\bar{\mathtt{M}}\mathtt{M}_i\right] + 2\bar{\mathtt{M}}\sum_{i=1}^b\left[\left(\frac{\omega_i+1}{p_i}\right)A_{n_i,N_i}\mathtt{M}_i\right] + 4C_{\mathbf{n},\mathbf{N}}\sum_{i=1}^b(\omega_i+1-p_i)/p_i\,, \tag{S60}$$

$A_{n,N}$ and $C_{n,N}$ are defined in (S36) and (S51) respectively, ψ is defined in (S57), and $(\Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}} = (\theta_{\tilde{k}}, \zeta_{\tilde{k}}, \{\eta^i_{\tilde{k}}\}_{i\in[b]})_{\tilde{k}\in\mathbb{N}}$ is defined in (S42), (S44), (S45).

Proof. Let $k \in \mathbb{N}$ and write $k = sl + r$ with $s \in \mathbb{N}$, $r \in \{0,\dots,l-1\}$ Then, using Lemma S6, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\tilde{G}(\Theta_{sl+r}; X_{sl+r+1}) - \nabla U(\theta_k)\right\|^2\right] \\
&\leq \left[2\sum_{i=1}^b \frac{\mathtt{M}_i^2}{p_i}(\omega_i+1-p_i) + \left(\frac{\omega_i+1}{p_i}\right)A_{n_i,N_i}\bar{\mathtt{M}}\mathtt{M}_i\right]\mathbb{E}\left[\|\theta_k-\theta^\star\|^2\right] \\
&+ \left[2\sum_{i=1}^b(\omega_i+1-p_i)/p_i\right]\mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_k^{(i)}\right\|^2\right] + 2\bar{\mathtt{M}}\sum_{i=1}^b\left[\left(\frac{\omega_i+1}{p_i}\right)A_{n_i,N_i}\mathtt{M}_i\right]\mathbb{E}\left[\|\theta_{sl}-\theta^\star\|^2\right].
\end{aligned} \tag{S61}$$

We now use our previous results to upper bound the three expectations at the right-hand side of (S61). First, using Corollary S7 and a straightforward induction gives

$$\begin{aligned}
\mathbb{E}\left[\|\theta_{sl}-\theta^\star\|^2\right] &\leq (1-\gamma\mathtt{m}/2)^s\psi(\theta_0,\eta_0) + 2\gamma ld\sum_{j=0}^{s-1}(1-\gamma\mathtt{m}/2)^j \\
&\leq (1-\gamma\mathtt{m}/2)^s\psi(\theta_0,\eta_0) + 4ld/\mathtt{m}\,.
\end{aligned} \tag{S62}$$

Similarly, we have

$$\begin{aligned}
\mathbb{E}\left[\|\theta_k-\theta^\star\|^2\right] &\leq (1-\gamma\mathtt{m}/2)^{s+1}\psi(\theta_0,\eta_0) + 2\gamma ld\sum_{j=0}^{s}(1-\gamma\mathtt{m}/2)^j \\
&\leq (1-\gamma\mathtt{m}/2)^s\psi(\theta_0,\eta_0) + 4ld/\mathtt{m}\,.
\end{aligned} \tag{S63}$$

Finally, using Lemma S8 combined with (S62) and (S63), we obtain

$$\sum_{i=1}^b \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_k^{(i)}\right\|^2\right] \leq (1-\alpha)\sum_{i=1}^b \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_{k-1}^{(i)}\right\|^2\right]$$

$$+ 2\alpha C_{\mathbf{n},\mathbf{N}}(1-\gamma\mathfrak{m}/2)^s\psi(\theta_0,\eta_0) + 8ld\alpha C_{\mathbf{n},\mathbf{N}}/\mathfrak{m}\,.$$

Then, a straightforward induction leads to

$$\sum_{i=1}^b \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_k^{(i)}\right\|^2\right] \le (1-\alpha)^k \sum_{i=1}^b \left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2 + 2C_{\mathbf{n},\mathbf{N}}(1-\gamma\mathfrak{m}/2)^s\psi(\theta_0,\eta_0) + 8ldC_{\mathbf{n},\mathbf{N}}/\mathfrak{m}\,. \tag{S64}$$

Combining (S62), (S63) and (S64) in (S61) concludes the proof. □

S3.3 Proof of Theorem 3

Note that $\gamma \in (0,\bar\gamma]$, $\alpha \in (0,\bar\alpha]$ and $l \in \mathbb{N}^*$, $(\Theta_{\tilde k})_{\tilde k\in\mathbb{N}} = (\theta_{\tilde k},\zeta_{\tilde k},\{\eta_{\tilde k}^{(i)}\}_{i\in[b]})_{\tilde k\in\mathbb{N}}$ defined in (S42), (S44), (S45) is a inhomogeneous Markov chain associated with the sequence of Markov kernel $(Q^{(k)}_{\gamma,\alpha,l})_{k\in\mathbb{N}}$ defined by as follows. Define for any $(\theta,\zeta,\eta)\in\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d$, and $x^{(1)}\in\wp_{N_i,n_i}$, $x^{(2)}\in\mathsf{X}_2$ and $x^{(3)}\in\mathsf{X}_3$,

$$\mathscr{F}_i((\theta,\zeta,\eta);(x^{(1)},x^{(2)},x^{(3)})) = \mathscr{S}_i\left(\mathscr{C}_i\left\{G_i\left(\theta,\zeta;x^{(1)}\right)-\eta;x^{(2)}\right\};x^{(3)}\right)$$
$$\mathscr{G}_i((\theta,\zeta,\eta);(x^{(1)},x^{(2)},x^{(3)})) = \eta + \alpha\mathscr{F}_i((\theta,\zeta,\eta);(x^{(1)},x^{(2)},x^{(3)}))\,.$$

and for $\tilde\theta\in\mathbb{R}^d$, $\{\eta^{(i)}\}_{i=1}^b\in\mathbb{R}^{db}$, $\{x^{(1,i)}\}_{i=1}^b\in\otimes_{i=1}^b\wp_{N_i,n_i}$, $\{x^{(2,i)}\}_{i=1}^b\in\mathsf{X}_2^b$, $\{x^{(3,i)}\}_{i=1}^b\in\mathsf{X}_3^b$, setting $x^{(1:b)} = \{(x^{(1,i)},x^{(2,i)},x^{(3,i)})\}_{i=1}^b$,

$$\varphi_\gamma((\tilde\theta,\theta,\zeta,\{\eta^{(i)}\}_{i=1}^b);x^{(1:b)}) = (4\pi\gamma)^{-d/2}\exp\left(-\|\tilde\theta-\theta+\gamma\textstyle\sum_{i=1}^b\mathscr{F}_i((\theta,\zeta,\eta^{(i)});x^{(i)})\|^2/(4\gamma)\right)\,.$$

Denote $\tilde{\mathsf{X}}^{(i)} = \wp_{N_i,n_i}\times\mathsf{X}_2\times\mathsf{X}_3$ and $\tilde\nu^{(i)} = \nu_1^{(i)}\times\nu_2\times\nu_3$. Set $Q^{(0)}_{\gamma,\alpha,l} = \mathrm{Id}$ and for $k\ge 0$, $k = ls+r$, $s\in\mathbb{N}$, $r\in\{0,\dots,l-1\}$, $(\theta,\zeta,\eta)\in\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^{db}$ and $\mathsf{A}\in\mathcal{B}(\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^{db})$,

if $r=0$

$Q^{(k+1)}_{\gamma,\alpha,l}((\theta,\zeta,\eta),\mathsf{A}) =$

$\int_{\otimes_{i=1}^b\tilde{\mathsf{X}}^{(i)}}\mathbf{1}_{\mathsf{A}}(\tilde\theta,\tilde\zeta,\tilde\eta)\varphi_\gamma((\tilde\theta,\theta,\zeta,\{\eta^{(i)}\}_{i=1}^b);x^{(1:b)})\{\prod_{i=1}^b\delta_{\mathscr{G}_i((\theta,\zeta,\eta);x^{(i)})}(\mathrm{d}\tilde\eta^{(i)})\}\delta_\theta(\mathrm{d}\tilde\zeta)\,\mathrm{d}\tilde\theta\,\otimes_{i=1}^b\tilde\nu^{(i)}(\mathrm{d}x^{(i)})$

otherwise

$Q^{(k+1)}_{\gamma,\alpha,l}((\theta,\zeta,\eta),\mathsf{A}) =$

$\int_{\otimes_{i=1}^b\tilde{\mathsf{X}}^{(i)}}\mathbf{1}_{\mathsf{A}}(\tilde\theta,\tilde\zeta,\tilde\eta)\varphi_\gamma((\tilde\theta,\theta,\zeta,\{\eta^{(i)}\}_{i=1}^b);x^{(1:b)})\{\prod_{i=1}^b\delta_{\mathscr{G}_i((\theta,\zeta,\eta);x^{(i)})}(\mathrm{d}\tilde\eta^{(i)})\}\delta_\zeta(\mathrm{d}\tilde\zeta)\,\mathrm{d}\tilde\theta\,\otimes_{i=1}^b\tilde\nu^{(i)}(\mathrm{d}x^{(i)})\,.$

Consider then, the Markov kernel on $\mathbb{R}^d\times\mathcal{B}(\mathbb{R}^d)$,

$$R^{(k)}_{\gamma,\alpha,l,\eta_0}(\theta_0,\mathsf{A}) = Q^{(k)}_{\gamma,\alpha,l}((\theta_0,\theta_0,\eta_0),\mathsf{A}\times\mathbb{R}^d\times\mathbb{R}^{db})\,. \tag{S65}$$

Define

$$\bar\gamma_\alpha = \bar\gamma_{\alpha,2}\wedge\bar\gamma_4\,,\qquad \bar\gamma_4 = 1/(10\mathfrak{m})\,, \tag{S66}$$

where $\bar\gamma_{\alpha,2}$ is defined in (S59). The following theorem provides a non-asymptotic convergence bound for the $\texttt{QLSD}^{++}$ kernel.

Theorem S9. *Assume **HS1**, **HS2**, **HS4** and **HS6**. and let $l\in\mathbb{N}^*$. Then, for any probability measure $\mu\in\mathcal{P}_2(\mathbb{R}^d)$, $\eta_0\in\mathbb{R}^{db}$, $\alpha\in\left(0,1/(1+\max_{i\in[b]}\omega_i)\right]$, $\gamma\in(0,\bar\gamma_\alpha]$, and $k = sl+r\in\mathbb{N}$ with $s\in\mathbb{N}$, $r\in\{0,\dots,l-1\}$, we have*

$$W_2^2(\mu R^{(k)}_{\gamma,\alpha,l,\eta_0},\pi) \le (1-\gamma\mathfrak{m}/2)^k W_2^2(\mu,\pi) + (2\gamma/\mathfrak{m})(1-\gamma\mathfrak{m}/2)^s D_{\mathbf{n},\mathbf{N}}\int_{\mathbb{R}^d}\psi(\theta_0,\eta_0)\mathrm{d}\mu(\theta_0) + (4\gamma/\mathfrak{m})\left[\sum_{i=1}^b(\omega_i+1-p_i)/p_i\right](1-\alpha)^k\sum_{i=1}^b\left\|\nabla U_i(\theta^\star)-\eta_0^{(i)}\right\|^2 + \gamma\mathrm{B}_{\oplus,\bar\gamma_\alpha}\,,$$

where $R^{(k)}_{\gamma,\alpha,l,\eta_0}$ *is defined in* (S65), ψ *is defined in* (S57), $D_{n,N}$ *in* (S60) *and*

$$\mathrm{B}_{\oplus,\bar{\gamma}_\alpha} = 2d\mathrm{L}^2(1/\mathtt{m} + 5\bar{\gamma}_\alpha)\left[1 + \bar{\gamma}_\alpha \mathrm{L}^2/(2\mathtt{m}) + \bar{\gamma}_\alpha^2 \mathrm{L}^2/12\right]/\mathtt{m} + 96ld\left(\sum_{i=1}^{b} \mathtt{M}_i(\omega_i+1)(\mathtt{M}_i + \bar{\mathtt{M}}A_{n_i,N_i})/p_i\right)/\mathtt{m}^2 . \quad \text{(S67)}$$

Proof. Let $k \in \mathbb{N}$. The proof follows from the same lines as Theorem S5. By (S19) and (S42), we have

$$\vartheta_{\gamma(k+1)} - \theta_{k+1} = \vartheta_{\gamma k} - \theta_k - \gamma\left[\nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right] - \int_0^\gamma \left[\nabla U(\vartheta_{\gamma k+s}) - \nabla U(\vartheta_{\gamma k})\right] \mathrm{d}s + \gamma\left[\tilde{G}(\Theta_k; X_{k+1}) - \nabla U(\theta_k)\right] .$$

Define the filtration $(\mathcal{H}_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$ as $\mathcal{H}_0 = \sigma(\vartheta_0, \Theta_0)$ and for $\tilde{k} \in \mathbb{N}^*$,

$$\mathcal{H}_{\tilde{k}} = \sigma(\vartheta_0, \Theta_0, (X_l^{(1)}, \dots, X_l^{(b)})_{1\le l\le \tilde{k}}, (B_t)_{0\le t\le \gamma\tilde{k}}) .$$

Note that since $(\vartheta_t)_{t\ge 0}$ is a strong solution of (S19), then is easy to see that $(\vartheta_{\gamma\tilde{k}}, \Theta_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$ is $(\mathcal{H}_{\tilde{k}})_{\tilde{k}\in\mathbb{N}}$-adapted. Taking the squared norm and the conditional expectation with respect to $\mathcal{H}_k$, we obtain using **H**S5-(i) that

$$\begin{aligned}
\mathbb{E}^{\mathcal{H}_k}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] &= \left\|\vartheta_{\gamma k} - \theta_k\right\|^2 - 2\gamma\left\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\rangle \\
&+ 2\gamma\int_0^\gamma \left\langle \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k), \mathbb{E}^{\mathcal{H}_k}\left[\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right]\right\rangle \mathrm{d}u \\
&- 2\int_0^\gamma \left\langle \vartheta_{\gamma k} - \theta_k, \mathbb{E}^{\mathcal{H}_k}\left[\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right]\right\rangle \mathrm{d}u \\
&+ \gamma^2\left\|\nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\|^2 \\
&+ \mathbb{E}^{\mathcal{H}_k}\left[\left\|\int_0^\gamma \left[\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right] \mathrm{d}u\right\|^2\right] \\
&+ \gamma^2\mathbb{E}^{\mathcal{H}_k}\left[\left\|\tilde{G}(\Theta_k; X_{k+1}) - \nabla U(\theta_k)\right\|^2\right] . \qquad \text{(S68)}
\end{aligned}$$

Using Proposition S8, we obtain

$$\begin{aligned}
\mathbb{E}\left[\left\|\tilde{G}(\Theta_k; X_{k+1}) - \nabla U(\theta_k)\right\|^2\right] &\le (1-\gamma\mathtt{m}/2)^{\lfloor k/l\rfloor} D_{\mathbf{n},\mathbf{N}}\psi(\theta_0,\eta_0) + 4ldD_{\mathbf{n},\mathbf{N}}/\mathtt{m} \\
&+ \left[2\sum_{i=1}^b (\omega_i + 1 - p_i)/p_i\right](1-\alpha)^k \sum_{i=1}^b \mathbb{E}\left[\left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2\right] . \qquad \text{(S69)}
\end{aligned}$$

Then, we control the remaining terms in (S68) using (S25), (S26) and (S27). Combining these bounds and (S69) into (S68), for any $\varepsilon > 0$, yields

$$\begin{aligned}
\mathbb{E}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] &\le (1 + 2\gamma\varepsilon - 5\gamma^2\mathtt{m}\mathrm{L})\mathbb{E}\left[\left\|\vartheta_{\gamma k} - \theta_k\right\|^2\right] \\
&- \gamma\left[2 - 5\gamma(\mathtt{m}+\mathrm{L})\right]\mathbb{E}\left[\left\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k)\right\rangle\right] \\
&+ (5\gamma + (2\varepsilon)^{-1})\int_0^\gamma \mathbb{E}\left[\left\|\nabla U(\vartheta_{\gamma k+u}) - \nabla U(\vartheta_{\gamma k})\right\|^2\right] \mathrm{d}u \\
&+ \gamma^2(1-\gamma\mathtt{m}/2)^{\lfloor k/l\rfloor} D_{\mathbf{n},\mathbf{N}}\mathbb{E}\left[\psi(\theta_0,\eta_0)\right] + 4ldD_{\mathbf{n},\mathbf{N}}/\mathtt{m} \\
&+ 2\gamma^2\left[\sum_{i=1}^b (\omega_i + 1 - p_i)/p_i\right](1-\alpha)^k \sum_{i=1}^b \left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2 .
\end{aligned}$$

Next, we use that under **H**S1, $\langle \vartheta_{\gamma k} - \theta_k, \nabla U(\vartheta_{\gamma k}) - \nabla U(\theta_k) \rangle \geq \mathtt{m} \|\vartheta_{\gamma k} - \theta_k\|^2$ and $|\langle \theta_k - \theta^\star, \nabla U(\theta_k) - \nabla U(\theta^\star) \rangle| \leq \mathtt{L} \|\theta_k - \theta^\star\|^2$, which implies taking $\varepsilon = \mathtt{m}/2$ and since $2 - 5\gamma(\mathtt{m} + \mathtt{L}) \geq 0$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\vartheta_{\gamma(k+1)} - \theta_{k+1}\right\|^2\right] &\leq (1 - \gamma\mathtt{m}(1 - 5\gamma\mathtt{m}))\mathbb{E}\left[\|\vartheta_{\gamma k} - \theta_k\|^2\right] \\
&\quad + (5\gamma + \mathtt{m}^{-1}) \int_0^\gamma \mathbb{E}\left[\|\nabla U(\vartheta_{\gamma k + u}) - \nabla U(\vartheta_{\gamma k})\|^2\right] \mathrm{d}u \\
&\quad + \gamma^2 (1 - \gamma\mathtt{m}/2)^{\lfloor k/l \rfloor} D_{\mathbf{n},\mathbf{N}} \mathbb{E}\left[\psi(\theta_0, \eta_0)\right] + 4ldD_{\mathbf{n},\mathbf{N}}/\mathtt{m} \\
&\quad + 2\gamma^2 \left[\sum_{i=1}^b (\omega_i + 1 - p_i)/p_i\right] (1 - \alpha)^k \sum_{i=1}^b \left\|\nabla U_i(\theta^\star) - \eta_0^{(i)}\right\|^2 . \qquad \text{(S70)}
\end{aligned}$$

Further, for any $u \in \mathbb{R}_+$, using Durmus and Moulines (2019, Lemma 21) we have

$$\mathtt{L}^{-2}\, \mathbb{E}\left[\|\nabla U(\vartheta_{\gamma k + u}) - \nabla U(\vartheta_{\gamma k})\|^2\right] \leq du\left(2 + u^2\mathtt{L}^2/3\right) + 3u^2\mathtt{L}^2/2\mathbb{E}\left[\|\vartheta_{\gamma k} - \theta^\star\|^2\right] .$$

Integrating the previous inequality on $[0, \gamma]$, we obtain

$$\mathtt{L}^{-2} \int_0^\gamma \mathbb{E}\left[\|\nabla U(\vartheta_{\gamma k + u}) - \nabla U(\vartheta_{\gamma k})\|^2\right] \mathrm{d}u \leq d\gamma^2 + d\gamma^4\mathtt{L}^2/12 + \gamma^3\mathtt{L}^2/2\mathbb{E}\left[\|\vartheta_{\gamma k} - \theta^\star\|^2\right] .$$

Plugging this bounds in (S70) and using Durmus and Moulines (2019, Proposition 1) complete the proof.

□

S4 CONSISTENCY ANALYSIS IN THE BIG DATA REGIME

In this section, we assume that the number of observations on each client $i \in [b]$ writes $N_i = \lfloor c_i N \rfloor$ where $\{c_i > 0\}_{i \in [b]}$, $N \in \mathbb{N}^*$, and provide upper bounds on the asymptotic bias associated to each algorithm when N tends towards infinity. For simplicity, we assume for any $i \in [b]$, that $n_i = \lfloor c_i n \rfloor$ with $n \in [N]$, $\mathtt{M}_i = \mathtt{M}$ with $\mathtt{M} > 0$, $p_i = 1$ and $\omega_i = \omega$ with $\omega > 0$ but note that our conclusions also hold for the general setting considered in this paper.

S4.1 Asymptotic analysis for Algorithm 1

The following corollary is associated with `QLSD` defined in Algorithm 1 in the main paper.

Corollary S10. *Assume* ***H****S1, **H**S2 **H**S3 and **H**S4. In addition, assume that* $\liminf_{N\to\infty} \mathtt{m}/N > 0$ *and* $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ *for* $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}, \mathtt{B}^\star, \sigma_\star\}$. *Then, we have* $\bar{\gamma} = \bar{\eta}/N$ *where* $\bar{\eta} > 0$ *and* $\bar{\gamma}$ *is defined in* (S18). *In addition,*

$$B_{\bar{\gamma}} = (\omega + 1)\, \mathrm{O}(N) ,$$

where $B_{\bar{\gamma}}$ *is defined in* (S31).

Proof. Since we assume that $\liminf_{N\to\infty} \mathtt{m}/N > 0$ and $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ for $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}, \mathtt{B}^\star, \sigma_\star\}$, there exist $C_\mathtt{m}$, $C_\mathtt{L}$, $C_\mathtt{M}$, $C_{\mathtt{B}^\star}$ and $C_{\sigma_\star} > 0$ such that $\mathtt{m} \geq C_\mathtt{m} N$, $\mathtt{L} \leq C_\mathtt{L} N$, $\mathtt{M} \leq C_\mathtt{M} N$, $\mathtt{B}^\star \leq C_{\mathtt{B}^\star} N$ and $\sigma_\star \leq C_{\sigma_\star} N$. Under these assumptions, it is straightforward from (S18) to see that there exists $\bar{\eta} > 0$ such that $\bar{\gamma} = \bar{\eta}/N$. In addition, it follows from (S31) that

$$B_{\bar{\gamma}} \leq \frac{2dC_\mathtt{L}^2}{C_\mathtt{m}} \left(\frac{1}{C_\mathtt{m}} + 5\bar{\eta}\right) \left[1 + \frac{\bar{\eta}C_\mathtt{L}^2}{2C_\mathtt{m}} + \frac{\bar{\eta}^2 C_\mathtt{L}^2}{12}\right] + \frac{4}{C_\mathtt{m}} (\omega C_{\mathtt{B}^\star} + C_{\sigma_\star}^2 N) + \frac{8(\omega + 1) C_\mathtt{L} C_\mathtt{M}}{C_\mathtt{m}^2} \left[d + \bar{\eta}\left(\omega C_{\mathtt{B}^\star} + C_{\sigma_\star}^2 N\right)\right] .$$

The proof is concluded by letting N tend towards infinity. □

Regarding the specific instance `QLSD`$^\#$ of Algorithm 1 in the main paper, a similar result holds. Indeed, by using Lemma S4, we can notice that **H**S3-(iii) is verified with $\sigma_\star = C_{\sigma_\star} N$ for some $C_{\sigma_\star} > 0$ and we can apply Corollary S10.

S4.2 Asymptotic analysis for Algorithm 2

The following corollary is associated with QLSD* defined in Algorithm 2 in the main paper.

Corollary S11. *Assume* ***HS1****,* ***HS2****,* ***HS4*** *and* ***HS6****. In addition, assume that* $\liminf_{N\to\infty} \mathtt{m}/N > 0$ *and* $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ *for* $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}\}$*. Then, we have* $\bar{\gamma} = \bar{\eta}/N$ *where* $\bar{\eta} > 0$ *and* $\bar{\gamma}$ *is defined in* (S18)*. In addition,*

$$B_{\circledast,\bar{\gamma}} = d(\omega+1)\,\mathrm{O}(1)\,,$$

where $B_{\circledast,\bar{\gamma}}$ *is defined in* (S35)*.*

Proof. Since we assume that $\liminf_{N\to\infty} \mathtt{m}/N > 0$ and $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ for $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}\}$, there exist $C_\mathtt{m}$, $C_\mathtt{L}$ and $C_\mathtt{M} > 0$ such that $\mathtt{m} \geq C_\mathtt{m} N$, $\mathtt{L} \leq C_\mathtt{L} N$ and $\mathtt{M} \leq C_\mathtt{M} N$. Under these assumptions, it is straightforward from (S18) to see that there exists $\bar{\eta} > 0$ such that $\bar{\gamma}_\alpha = \bar{\eta}/N$. In addition, it follows from (S10) that

$$B_{\circledast,\bar{\gamma}} \leq \frac{2dC_\mathtt{L}^2}{C_\mathtt{m}}\left(\frac{1}{C_\mathtt{m}} + 5\bar{\eta}\right)\left[1 + \frac{\bar{\eta}C_\mathtt{L}^2}{2C_\mathtt{m}} + \frac{\bar{\eta}^2 C_\mathtt{L}^2}{12}\right] + \frac{4d\bar{\mathtt{M}}C_\mathtt{L}}{C_\mathtt{m}^2}\max_{i\in[b]}\left\{c_i\omega + (\omega+1)\cdot\frac{N-n}{n(\lfloor c_i N\rfloor - 1)}\right\}.$$

The proof is concluded by letting N tend towards infinity. □

Lastly, we have the following asymptotic convergence result regarding QLSD^{++} defined in Algorithm 2 in the main paper.

Corollary S12. *Assume* ***HS1****,* ***HS2****,* ***HS4*** *and* ***HS6****. In addition, assume that* $\liminf_{N\to\infty} \mathtt{m}/N > 0$ *and* $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ *for* $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}\}$*. Then, we have* $\bar{\gamma}_\alpha = \bar{\eta}/N$ *where* $\bar{\eta} > 0$ *and* $\bar{\gamma}_\alpha$ *is defined in* (S66)*. In addition,*

$$B_{\oplus,\bar{\gamma}_\alpha} = d(\omega+1)\,\mathrm{O}(1)\,,$$

where $B_{\oplus,\bar{\gamma}_\alpha}$ *is defined in* (S67)*.*

Proof. Since we assume that $\liminf_{N\to\infty} \mathtt{m}/N > 0$ and $\limsup_{N\to\infty} \mathtt{A}/N < \infty$ for $\mathtt{A} \in \{\mathtt{L}, \mathtt{M}\}$, there exist $C_\mathtt{m}$, $C_\mathtt{L}$ and $C_\mathtt{M} > 0$ such that $\mathtt{m} \geq C_\mathtt{m} N$, $\mathtt{L} \leq C_\mathtt{L} N$ and $\mathtt{M} \leq C_\mathtt{M} N$. Under these assumptions, it is straightforward from (S66) to see that there exists $\bar{\eta} > 0$ such that $\bar{\gamma}_\alpha = \bar{\eta}/N$. In addition, it follows from (S67) that

$$\mathrm{B}_{\oplus,\bar{\gamma}_\alpha} \leq \frac{2dC_\mathtt{L}^2}{C_\mathtt{m}}\left(\frac{1}{C_\mathtt{m}} + 5\bar{\eta}\right)\left[1 + \frac{\bar{\eta}C_\mathtt{L}^2}{2C_\mathtt{m}} + \frac{\bar{\eta}^2 C_\mathtt{L}^2}{12}\right] + \frac{96(\omega+1)ldbC_\mathtt{M}}{C_\mathtt{m}^2}\left(\frac{(N-n)\bar{\mathtt{M}}}{n(\min_{i\in[b]}\{\lfloor c_i N\rfloor\} - 1)} + C_\mathtt{M}\right).$$

The proof is concluded by letting N tend towards infinity. □

S5 EXPERIMENTAL DETAILS

In this section, we provide additional details regarding our numerical experiments. The code, data and instructions to reproduce our experimental results can be found in the supplementary material (see folder ./code).

S5.1 Toy Gaussian example

Pseudo-code of LSD*. For completeness, we provide in Algorithm S2 the pseudo-code of the non-compressed counterpart of QLSD*, namely LSD*.

Additional experimental details. The code associated to this experiment can be found in the supplementary material (see ./code/notebook_toy_Gaussian-experiment.ipynb). As highlighted in Section 4 (*Toy Gaussian example* paragraph) in the main paper, the synthetic dataset has been generated so that each client owns a heterogeneous and unbalanced dataset. An illustration of the unbalancedness is given in Figure S1. The precise procedure to generate such a dataset can be found in the aforementioned notebook.

To obtain the figure at the bottom row of Figure 1 in the main paper, we launched all the MCMC algorithms with $K = 500,000$ outer iterations and considered a burn-in period of $450,000$ iterations. Hence, only the last $50,000$ samples have been used to compute the MSE associated to the test function $f : \theta \mapsto \|\theta\|$. In order to compute the expected number of bits transmitted during each upload period, we considered the Elias encoding scheme and used the upper-bounds given in Alistarh et al. (2017, Theorem 3.2 and Lemma A.2).

Algorithm S2 Variance-reduced Langevin Stochastic Dynamics (LSD*)

Input: minibatch sizes $\{n_i\}_{i\in[b]}$, number of iterations K, step-size $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$ and initial point θ_0.
for $k = 0$ **to** $K-1$ **do**
  for $i \in \mathcal{A}_{k+1}$ // On active clients **do**
    Draw $\mathcal{S}_{k+1}^{(i)} \sim \text{Uniform}\,(\wp_{N_i,n_i})$.
    Set $H_{k+1}^{(i)}(\theta_k) = (N_i/n_i)\sum_{j\in\mathcal{S}_{k+1}^{(i)}} [\nabla U_{i,j}(\theta_k) - \nabla U_{i,j}(\theta^\star)]$.
    Compute $g_{i,k+1} = H_{k+1}^{(i)}(\theta_k)$.
    Send $g_{i,k+1}$ to the central server.
  end for
  // On the central server
  Compute $g_{k+1} = \frac{b}{|\mathcal{A}_{k+1}|}\sum_{i\in\mathcal{A}_{k+1}} g_{i,k+1}$.
  Draw $Z_{k+1} \sim \mathrm{N}(0_d, \mathrm{I}_d)$.
  Compute $\theta_{k+1} = \theta_k - \gamma g_{k+1} + \sqrt{2\gamma} Z_{k+1}$.
  Send θ_{k+1} to the b clients.
end for
Output: samples $\{\theta_k\}_{k=0}^K$.

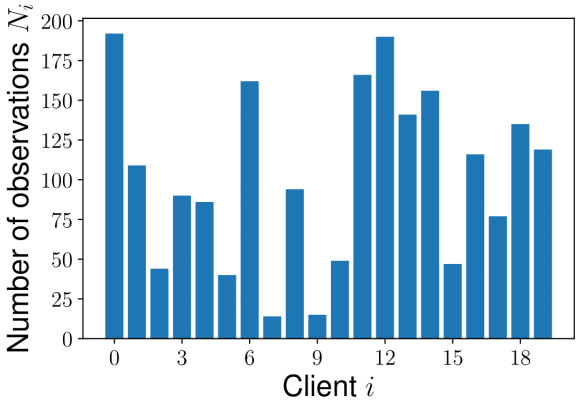

Figure S1: Illustration of the unbalancedness of the synthetic dataset used in the Toy Gaussian experiment.

- **License of the assets:** No existing asset has been used for this experiment.
- **Total amount of compute and type of resources used:** This experiment has been run on a laptop running Windows 10 and equipped with Intel(R) Core(TM) i7_8565U CPU 1.80GHz with 16Go of RAM. The total amount of compute is roughly 33 hours.
- **Training details:** All training details (here hyperparameters) are detailed in Section 4 in the main paper.

Discretisation step-size and compression trade-off. We complement the analysis made in the main paper by showing on Figure S2 that the saving in terms of number of transmitted bits can be further improved by decreasing the value of γ. This numerical finding illustrates our theory which in particular shows that the asymptotic bias associated to QLSD* is of the order $\omega\,\mathrm{O}(\gamma)$, see Table 1 in the main paper.

S5.2 Bayesian logistic regression

Pseudo-code of LSD^{++}. For completeness, we provide in Algorithm S3 the pseudo-code of the non-compressed counterpart of QLSD^{++}, namely LSD^{++}.

Additional experimental details. The code associated to this experiment can be found in the supplementary material (see ./code/notebook_logistic_regression.ipynb). For the Bayesian logistic regression experiment

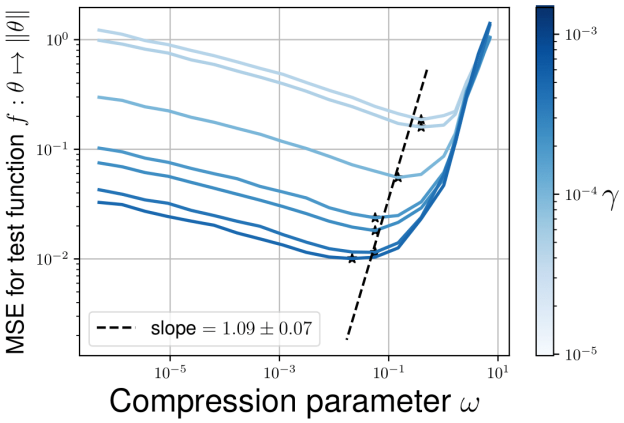

Figure S2: Toy Gaussian example. Trade-off between step-size and compression parameter values.

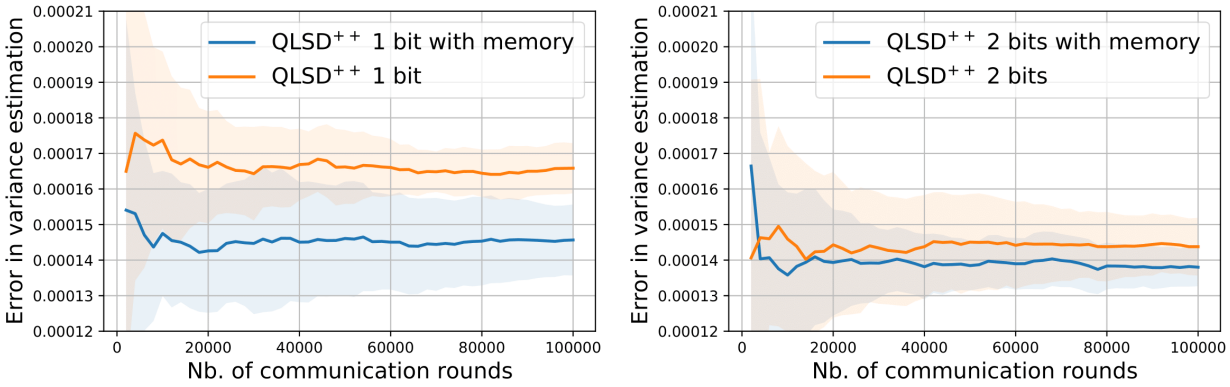

Figure S3: Bayesian logistic regression on synthetic data.

detailed in the main paper, we ran the MCMC algorithms with $K = 500,000$ outer iterations and considered a burn-in period of length $50,000$.

Benefits of the memory mechanism. We also run an additional experiment on a low-dimensional synthetic dataset to highlight the benefits brought by the memory mechanism involved in `QLSD`$^{++}$ when the dataset is highly heterogeneous. To this end, we consider the SYNTHETIC(α, β) dataset (Li et al., 2020) with $\alpha = \beta = 1$, $d = 2$ and $b = 50$. We run `QLSD`$^{++}$ with and without memory terms using $l = 100$, $\alpha = 1/(\omega + 1)$, $\gamma = 10^{-5}$ and for huge compression parameters, namely $s \in \{2^1, 2^2\}$. We use $K = 100,000$ outer iterations without considering a burn-in period. In order to have access to some ground truth, we also implement the Metropolis-adjusted Langevin algorithm (`MALA`) (Robert and Casella, 2004).

Figure S3 shows the Euclidean norm of the error between the true variance under π estimated with `MALA` and the empirical variance computed using samples generated by `QLSD`$^{++}$. As expected, we can notice that the memory mechanism reduces the impact of the compression on the asymptotic bias of `QLSD`$^{++}$ when ω is large.

Results on a non-image dataset. In order to complement our results on an image dataset (FEMNIST), we also implement our methodology and one competitor (`DG-SGLD`) on the *covtype*[1] dataset. Again, the ground truth has been obtained by implementing a long-run Metropolis-adjusted Langevin algorithm. The results we obtained are gathered in Table S1.

- **License of the assets:** We use the Synthetic dataset whose associated code is under the MIT license, and the FEMNIST dataset whose data are publicy available and associated code is under MIT license.

[1] https://archive.ics.uci.edu/ml/datasets/covertype

Algorithm S3 Variance-reduced Langevin Stochastic Dynamics (LSD^{++})

Input: minibatch sizes $\{n_i\}_{i\in[b]}$, number of iterations K, step-size $\gamma \in (0, \bar{\gamma}]$ with $\bar{\gamma} > 0$, initial point θ_0 and $\alpha \in (0, \bar{\alpha}]$ with $\bar{\alpha} > 0$.

// Memory mechanism initialisation

Initialise $\{\eta_0^{(1)}, \ldots, \eta_0^{(b)}\}$ and $\eta_0 = \sum_{i=1}^b \eta_0^{(i)}$.

for $k = 0$ **to** $K - 1$ **do**

// Update of the control variates

if $k \equiv 0 \pmod{l}$ **then**

Set $\zeta_k = \theta_k$.

else

Set $\zeta_k = \zeta_{k-1}$

end if

for $i \in \mathcal{A}_{k+1}$ // On active clients **do**

Draw $\mathcal{S}_{k+1}^{(i)} \sim \text{Uniform}\,(\wp_{N_i,n_i})$.

Set $H_{k+1}^{(i)}(\theta_k) = (N_i/n_i) \sum_{j\in\mathcal{S}_{k+1}^{(i)}} [\nabla U_{i,j}(\theta_k) - \nabla U_{i,j}(\zeta_k)] + \nabla U_i(\zeta_k)$.

Compute $g_{i,k+1} = H_{k+1}^{(i)}(\theta_k) - \eta_k^{(i)}$.

Send $g_{i,k+1}$ to the central server.

Set $\eta_{k+1}^{(i)} = \eta_k^{(i)} + \alpha g_{i,k+1}$.

end for

// On the central server

Compute $g_{k+1} = \eta_k + \frac{b}{|\mathcal{A}_{k+1}|} \sum_{i\in\mathcal{A}_{k+1}} g_{i,k+1}$.

Set $\eta_{k+1} = \eta_k + \alpha \sum_{i\in\mathcal{A}_{k+1}}^b g_{i,k+1}$.

Draw $Z_{k+1} \sim \mathrm{N}(0_d, \mathrm{I}_d)$.

Compute $\theta_{k+1} = \theta_k - \gamma g_{k+1} + \sqrt{2\gamma} Z_{k+1}$.

Send θ_{k+1} to the b clients.

end for

Output: samples $\{\theta_k\}_{k=0}^K$.

- **Total amount of compute and type of resources used:** This experiment has been run on a laptop running Windows 10 and equipped with Intel(R) Core(TM) i7_8565U CPU 1.80GHz with 16Go of RAM. The total amount of compute is roughly 30 hours.
- **Training details:** Hyperparameter values are detailed in Section 4 in the main paper. Regarding our experiment on real data, we use a random subset of the initial training data (for computational reasons).

S5.3 Bayesian neural networks

The code associated to this experiment can be found in the supplementary material (see ./code/experiments-bayesian_neural_network/).

- **License of the assets:** We use the MNIST, FMNIST, CIFAR10 and SVHN datasets which are publicly downloadable with the torchvision.datasets package.

Table S1: Bayesian Logistic Regression on *covtype* dataset.

Algorithm	99% HPD error
DG-SGLD	1.8e-2
QLSD^{++} 4 bits	2.2e-3
QLSD^{++} 8 bits	2.0e-2
QLSD^{++} 16 bits	1.9e-2

- **Total amount of compute and type of resources used:** The total computational cost depends on the dataset, but is roughly 40 hours in the worst case.
- **Training details:** We consider the same hyperparameter values detailed in Table S2 for both training on MNIST and CIFAR10 except for the initialisation and the sampling period. For the MNIST dataset, we use the default random weights given by pytorch whereas for CIFAR-10 we use the warm-start provided by the pytorchcv library and consider a burn-in period of half the sampling period ($K = 10^4$ iterations) with a thinning of 10.

In the following, we denote D_{test} the test dataset and for any data $(x, y) \in \mathrm{D}_{\text{test}}$, we define the preditive density by

$$p(y \mid x) = \int p(y \mid x, \theta)\ \pi(\theta \mid \mathrm{D})\ \mathrm{d}\theta\,, \tag{S71}$$

where $p(y \mid x, \theta)$ is the conditional likelihood. For any input x, the predicted label is denoted by $y_{\text{pred}}(x) = \arg\max_y p(y \mid x)$.

Metrics used for the Bayesian neural network experiment in the main paper. In the main paper, we consider three metrics to compare the different Bayesian FL algorithms, namely *Accuracy*, *Agreement* and *TV*. They are defined in the following.

- **Accuracy:** Based on samples from the approximate posterior distribution, we compute the minimum mean-square estimator (*i.e.* corresponding to the posterior mean) and use it to make predictions on the test dataset. The *Accuracy* metric corresponds to the percentage of well-predicted labels.
- **Agreement:** Let denote p_{ref} and p the predictive densities associated to `HMC` and an approximate simulation-based algorithm, respectively. Similar to Izmailov et al. (2021), we define the agreement between p_{ref} and p as the fraction of the test datapoints for which the top-1 predictions of p_{ref} and p, *i.e.*

$$\text{agreement}(p_{\text{ref}}, p) = \frac{1}{|\mathrm{D}_{\text{test}}|} \sum_{x \in \mathrm{D}_{\text{test}}} \mathbf{1}\left\{ \arg\max_{y'} p_{\text{ref}}(y' \mid x) = \arg\max_{y'} p(y' \mid x) \right\}.$$

- **Total variation (TV):** By denoting $\mathcal{Y}$ the set of possible labels, we consider the total variation metric between p_{ref} and p, *i.e.*

$$\mathrm{TV}(p_{\text{ref}}, p) = \frac{1}{2|\mathrm{D}_{\text{test}}|} \sum_{x \in \mathrm{D}_{\text{test}}} \sum_{y' \in \mathcal{Y}} |p_{\text{ref}}(y' \mid x) - p(y' \mid x)|\,.$$

Performance results on a highly heterogeneous dataset. We train LeNet5 (LeCun et al., 1998) architechture on the MNIST dataset (Deng, 2012) and we consider the FMNIST (Xiao et al., 2017) as the out-of-distribution dataset. To obtain a highly heterogeneous setting, we split the data among $b = 20$ clients so that each client has a dominant label representing 40% of the total amount in the training set and 1% of the other labels as described in Figure S4.

Inspired by the scores defined in Guo et al. (2017), we measure the performance of the different algorithms and report those results in Table S2. These statistics aim to better understand the predictions in order to calibrate the models (Rahaman and Thiery, 2020).

Expected Calibration Error (ECE). To measure the difference between the accuracy and confidence of the predictions, we group the data into $M \geq 1$ buckets defined for any $m \in [M]$ by $\mathrm{B}_{\text{m}} = \{(x, y) \in \mathrm{D}_{\text{test}} : p(y_{\text{pred}}(x)|x) \in\ ](m-1)/M, m/M]\}$. As in the previous work of Ovadia et al. (2019), we denote the model accuracy on B_{m} by

$$\text{acc}\,(\mathrm{B}_{\text{m}}) = \frac{1}{|\mathrm{B}_{\text{m}}|} \sum_{(x,y) \in \mathrm{B}_{\text{m}}} \mathbf{1}_{y_{\text{pred}}(x) = y}$$

and define the confidence on B_{m} by

$$\text{conf}\,(\mathrm{B}_{\text{m}}) = \frac{1}{|\mathrm{B}_{\text{m}}|} \sum_{(x,y) \in \mathrm{B}_{\text{m}}} p(y_{\text{pred}}(x)|x)\,.$$

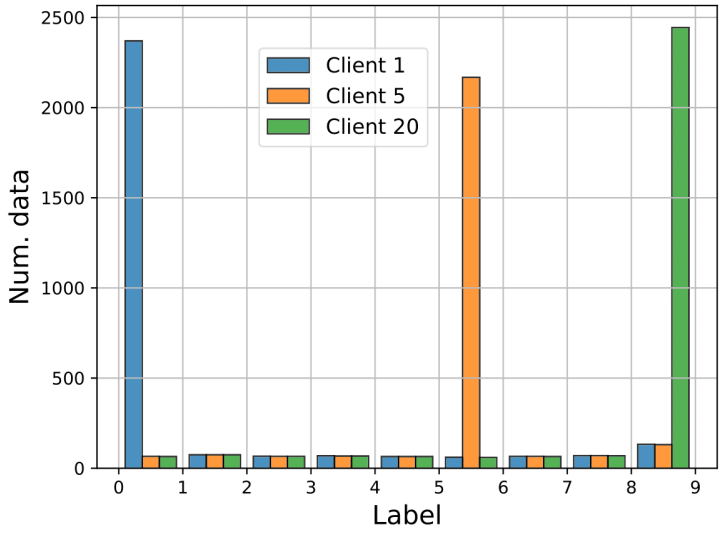

Figure S4: Number of labels owned by different clients.

Method	SGLD	pSGLD	QLSD	QLSD PP	QLSD^{++}	QLSD^{++} PP	FedBe-Gauss.	FedBe-Dirich.	FSGLD
Accuracy	99.1	99.2	98.8	98.3	98.8	98.7	43.5	79.3	98.5
$10^2\times$ ECE	0.577	1.25	0.916	1.57	0.692	0.930	7.51	21.3	2.65
$10^2\times$ BS	1.38	1.39	1.98	2.23	1.91	2.18	66.6	36.1	2.64
$10^2\times$ nNLL	2.86	3.16	4.15	4.82	4.11	4.65	139	78.0	6.19
Weight Decay	5	5	5	5	5	5	0	0	5
Batch Size	64	64	64	64	64	64	64	64	64
Learning rate	1e-07	1e-08	1e-07	1e-07	1e-07	1e-07	1e-02	1e-02	1e-07
Local steps	N/A	N/A	1	1	1	1	250	250	16
Burn-in	100epch.	100epch.	1e04	1e04	1e04	1e04	N/A	N/A	1e04
Thinning	1	1	500	500	500	500	N/A	N/A	500
Training	1e03epch.	1e03epch.	1e05it.	1e05it.	1e05it.	1e05it.	N/A	N/A	1e05it.

Table S2: Performance of Bayesian FL algorithms trained on the highly-heterogeneous dataset.

As stressed in Guo et al. (2017), for any $m \in [M]$ the accurcay $\mathrm{acc}\left(\mathrm{B_m}\right)$ is an unbiased and consistent estimator of $\mathbb{P}\left(y_{\mathrm{pred}}(x) = y \mid (m-1)/M < p(y_{\mathrm{pred}}(x)|x) \leq m/M\right)$. Therefore, the ECE defined by

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|\mathrm{B_m}|}{|\mathrm{D_{test}}|} \left|\mathrm{acc}\left(\mathrm{B_m}\right) - \mathrm{conf}\left(\mathrm{B_m}\right)\right|$$

is an estimator of

$$\mathbb{E}_{(x,y)}\Big[\left|\mathbb{P}\left(y_{\mathrm{pred}}(x) = y \mid p(y_{\mathrm{pred}}(x)|x)\right) - p(y_{\mathrm{pred}}(x)|x)\right|\Big].$$

Thus, ECE measures the absolute difference between the confidence level of a prediction and its accuracy.

Brier Score (BS). The BS is a proper scoring rule (see for example Dawid and Musio (2014)) that can only evaluate random variables taking a finite number of values. Denote by $\mathcal{Y}$ the finite set of possible labels, the BS measures the model's confidence in its predictions and is defined by

$$\mathrm{BS} = \frac{1}{|\mathrm{D_{test}}|} \sum_{(x,y)\in \mathrm{D_{test}}} \sum_{c\in\mathcal{Y}} \left(p(y=c|x) - \mathbf{1}_{y=c}\right)^2.$$

Normalised negative log-likelihood (nNLL). This classical score defined by

$$\mathrm{nNLL} = -\frac{1}{|\mathrm{D_{test}}|} \sum_{(x,y)\in \mathrm{D_{test}}} \log p(y|x)$$

measures the model ability to predict good labels with high probability.

Out of distribution detection. Here we study the behavior of our proposed algorithms in the out-of-distribution (OOD) framework, we consider the pairs MNIST/FMNIST and CIFAR10/SVHN, comparing the densities of the predictive entropies on the ID vs OOD data. These densities denoted by p_{in} and p_{out} respectively, are approximated using a kernel estimator based on of the histogram associated with $\{\text{Ent}(x) : x \in \text{D}_{\text{test}}^x\}$ for $\text{D}_{\text{test}} \in \{\text{MNIST}, \text{FMNIST}\}$ or $\{\text{CIFAR10}, \text{SVHN}\}$, where $\text{Ent}(x)$ is the predictive entropy defined by:

$$\text{Ent}(x) = \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \,,$$

and $p(y|x)$ is defined by (S71) and estimated by the different methods that we consider. The resulting densities from the different methods that we consider ared displayed in Figure S5.

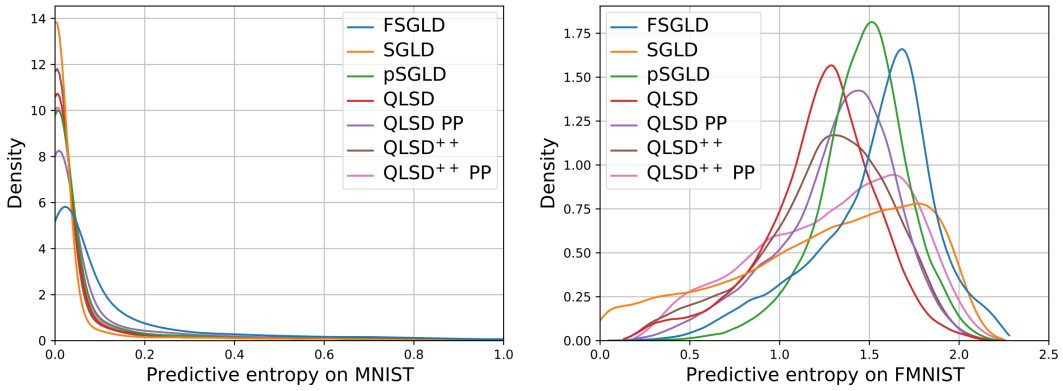

Figure S5: Predictive entropies comparison between MNIST and FMNIST.

A new data point x is then labeled in the original dataset (MNIST or CIFAR10) if $p_{\text{in}}(\text{Ent}(x)) > p_{\text{out}}(\text{Ent}(x))$ and out-of-distribution otherwise.

Calibration results. Interpreting the predicted outputs as probabilities is only correct for well a calibrated model. Indeed, when a model is calibrated, the confidence is closed to the accuracy of the predictions. In order to evaluate the calibration of the models, we display the reliability diagram on the left-hand side of Figure S6. It represents the evolution of $\text{acc}(\text{B}_\text{m}) - \text{conf}(\text{B}_\text{m})$ in function of $\text{conf}(\text{B}_\text{m})$, closer the values are to zero better the model is calibrated.

For the second sub-experiment, we consider for any $\tau \in [0, 1]$, the set $\text{D}_{\text{pred}}^{(\tau)} = \{x \in \text{D}_{\text{test}}^x : p(y|x) \geq \tau\}$ of classified data with credibility greater than τ. We define the test accuracy on $\text{D}_{\text{pred}}^{(\tau)}$ by

$$\text{Card}(\{x \in \text{D}_{\text{pred}}^{(\tau)} : y_{\text{true}}(x) = y_{\text{pred}}(x)\})/\text{Card}(\text{D}_{\text{pred}}^{(\tau)}) \,.$$

The right-hand side of Figure S6 shows the evolution of the test accuracy on $\text{D}_{\text{pred}}^{(\tau)}$ with respect to the credibility threshold τ. It can be noted that in both plots of Figure S6, the accuracy tends to 100% for confident predictions.

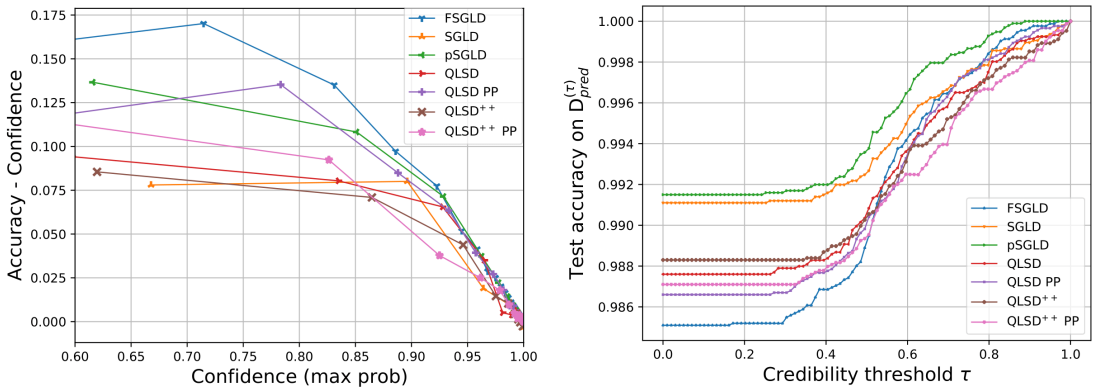

Figure S6: Left: Calibration test from reliability diagrams – Right: Test accuracy on $\mathrm{D}_{\mathrm{pred}}^{(\tau)}$ with respect to the threshold τ.