Federated Averaging Langevin Dynamics: Toward a unified theory and new algorithms

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

Alain Durmus
CMAP, École Polytechnique
Institut Polytechnique de Paris

Éric Moulines
CMAP, École Polytechnique
Institut Polytechnique de Paris

Abstract

This paper focuses on Bayesian inference in a federated learning context (FL). While several distributed MCMC algorithms have been proposed, few consider the specific limitations of FL such as communication bottlenecks and statistical heterogeneity. Recently, Federated Averaging Langevin Dynamics (FALD) was introduced, which extends the Federated Averaging algorithm to Bayesian inference. We obtain a novel tight non-asymptotic upper bound on the Wasserstein distance to the global posterior for FALD. This bound highlights the effects of statistical heterogeneity, which causes a drift in the local updates that negatively impacts convergence. We propose a new algorithm VR-FALD* that uses control variates to correct the client drift. We establish non-asymptotic bounds showing that VR-FALD* is not affected by statistical heterogeneity. Finally, we illustrate our results on several FL benchmarks for Bayesian inference.

1 Introduction

The paradigm of fully centralized machine learning is increasingly at odds with real-world use cases. Centralized machine learning leads to (a) data processing bottlenecks, (b) inefficient use of communication resources and (c) risks exposing individuals' private data. As storage and computational capacity increases at the agent level, it becomes increasingly attractive to decentralize computational tasks whenever possible. The term *federated* learning (FL) was recently coined to capture some aspects of this grand challenge (McMahan et al., 2017; Kairouz et al., 2021; Yang et al., 2019; Alistarh et al., 2017; Horváth et al., 2022; Wang et al., 2021).

Reducing communication costs has been identified as one of the major challenges of FL (Kairouz et al., 2021). Two main approaches have been proposed to achieve this goal. In the former, agents perform multiple local optimization steps before sending a model update to the central node (McMahan et al., 2017). The latter consists in compressing the messages exchanged (Alistarh et al., 2017; Horváth et al., 2022). In this paper, we focus on the first approach which is widely used in practice. However, due to statistical heterogeneity, performing multiple steps can hinder convergence, as model updates target each agent's local minimizer (Li et al., 2019; Ro et al., 2021). This results in a trade-off between communication cost and convergence (Wang et al., 2020) and a need for algorithms that mitigate *client drift* (Karimireddy et al., 2020).

Most of existing FL algorithms minimize a training loss. However, their results do not provide reliable uncertainty quantification, a strong requirement in safety-critical applications (Coglianese and Lehr, 2016; Fatima et al., 2017). We address this problem by considering the federated version of Bayesian inference (Welling and Teh, 2011; Yurochkin et al., 2019; Chen and Chao, 2021; Izmailov et al., 2021; Wilson et al., 2021). The objective is to compute the predictive distribution, highest posterior density regions (HPD). To this end, it is required to sample the posterior distribution $\pi \propto \exp(-U)$ associated with the model at hand. This target posterior decomposes into the product of local posteriors $\pi = \prod_{i\in[b]} \pi^i$. It is well known that sampling according to product distributions (Neiswanger et al., 2014; Hoffman et al., 2013; Minsker et al., 2014; Wang et al., 2015; Al-Shedivat et al., 2021; Dai et al., 2021) raises serious computational challenges even when sampling from each local posterior π^i is reasonably easy. We tackle this question in our contributions which can be summarized as follows.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

Contributions.

- We study a random loop version of the FALD algorithm proposed in Deng et al. (2021), and we establish non-asymptotic upper bounds in Wasserstein distance for strongly convex potentials U. An analysis of FALD was conducted in (Deng et al., 2021, Theorem 5.7). However, the proof is plagued by an error; see Section 7.1.
- We give matching lower bounds to show that even with full batch gradients, FALD can be slower than Stochastic Gradient Langevin Dynamics (SGLD) due to client-drift.
- We propose a new method VR-FALD* that circumvents the shortcomings of FALD. This algorithm extends the Shifted Local-SVRG of Gorbunov et al. (2021) to the Bayesian context. It combines Stochastic Variance Reduced Gradient (SVRG) Langevin Dynamics (LD) (Dubey et al., 2016) and adapts the bias reduction techniques from SCAFFOLD (Karimireddy et al., 2020).
- We derive theoretical guarantees for VR-FALD* which highlight its gradient variance reduction effect and its ability to deal with data heterogeneity.
- The results are based on a general framework developed in the supplement, that encompasses a broad family of federated Bayes algorithms based on Langevin dynamics. This is the first unifying study among existing works on federated Bayesian inference.
- Finally, in Section 4 we illustrate our results using classical FL benchmarks and provide a thorough comparison with existing FL Bayesian methods.

Related works. Many distributed MCMC algorithms have been proposed in the last decade and it is difficult to credit all the references. The first significant contributions in this direction are the Consensus Monte Carlo (CMC) approach and "embarrassingly parallel" MCMC algorithms; see, e.g. Neiswanger et al. (2014); Wang and Dunson (2013); Scott et al. (2016). These methods require running separate MCMC chains on each client/computational node, with each chain targeting the local posterior π^i. In the final stage, the algorithms recombine the samples from these chains to generate samples from the desired global posterior π (Minsker et al., 2014). The local posteriors may differ significantly from each other due to statistical heterogeneity, data imbalance, and / or inaccurate approximation. The effectiveness of the final combinations is either based on stringent assumptions on the local likelihoods (Liu and Ihler, 2014; Nemeth and Sherlock, 2018; Mesquita et al., 2020; Chittoor and Simeone, 2021) or on "fusion" algorithms that are exact but scale badly with the dimension; see, e.g. Dai et al. (2021); De Souza et al. (2022). Vono et al. (2020); Rendell et al. (2020); Plassier et al. (2021); Vono et al. (2022a) introduced hierarchical Bayesian models to simulate separate MCMC chains on each machine. Inspired by the alternating direction method of multipliers (Boyd et al., 2011), each client is assigned an auxiliary parameter that is conditionally independent given the server parameter. These authors developed MCMC schemes which alternate between sampling the clients parameters given the server parameter, and sampling the server parameter given the clients parameters. However, these approaches require tuning an additional hyperparameter to control the dispersion of the "local parameters". This parameter characterizes the trade-off between computational tractability and closeness to the original target distribution.

A competing approach to Federated Averaging, the quantized-SGD scheme, has been proposed in (Alistarh et al., 2017) for non Bayesian FL. In this framework, the agents do not adapt parameters locally but a random subset of the agents compute at each iteration a new gradient estimator and transmit a compressed form—see Haddadpour et al. (2021), among many others, (Bernstein et al., 2018; Tang et al., 2021) for scalar quantization or (Shlezinger et al., 2020), for vector quantization. These approaches have been extended to the Bayesian inference context in Lee et al. (2020); Zhang et al. (2022); Vono et al. (2022b). Performance analysis is given in Vono et al. (2022b); Sun et al. (2022).

The Federated Gradient Stochastic Langevin Dynamics (FSGLD algorithm introduced by El Mekkaoui et al. (2021) extends the distributed-SGLD (DSGLD) (Ahn et al., 2014) to the FL setting. Specifically, FSGLD operates passing a Markov chain between computing nodes and using only local data to estimate gradients at each step.

Methods with multiple local steps have been considered by several authors. Deng et al. (2021) designed FALD as a Bayesian version of FEDAVG. Al-Shedivat et al. (2021) proposed FEDPA as a generalization of FEDAVG. This method performs several local steps to infer Gaussian approximations of the clients local parameters. These local parameters are then reweighted using the estimated local means and covariance matrices before being aggregated on the central server.

Notation and Convention. The Euclidean norm on $\mathbb{R}^d$ is denoted by $\|\cdot\|$, and we set $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$. For $n \in \mathbb{N}^\star$, we refer to $\{1, \ldots, n\}$ with the notation $[n]$. We denote by $\mathcal{P}_2(\mathbb{R}^d)$ the set of probability measures on $\mathbb{R}^d$ with finite 2-moment. For any random variable ξ with values in $\mathbb{R}^d$, we define $\mathrm{Var}(\xi) = \mathbb{E}[\|\xi - \mathbb{E}\xi\|^2]$. Let μ, ν be in $\mathcal{P}_2(\mathbb{R}^d)$, we define the Wasserstein distance of order 2 by $\mathbf{W}_2(\mu,\nu) = (\inf_{\zeta\in\mathbf{\Pi}(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - x'\|^2 \mathrm{d}\zeta(x,x'))^{1/2}$, where $\mathbf{\Pi}(\mu,\nu)$ is the set of transference plans of μ and ν.

2 Algorithm derivation

We aim to sample a target probability density function π defined for $x \in \mathbb{R}^d$ by

$$\pi\left(x\right) \propto \textstyle\prod_{i=1}^{b} \pi^{i}\left(x\right) , \quad \pi^{i}(x) \propto \exp(-U^{i}(x)) , \qquad (1)$$

where b is the number of clients and the potential U^i is a finite sum expressed by

$$U^{i}(x) = \varpi^{i} U^{0}(x) + \textstyle\sum_{j=1}^{N_i} U^{i,j}(x) ,$$

with $\{\varpi^i\}_{i\in[b]} \in [0,1]^b$ and $\sum_{i\in[b]} \varpi^i = 1$. This setting encompasses the Bayesian federated learning as a particular case, in which π stands for the global posterior distribution and $\{\pi^i\}_{i\in[b]}$ are referred to as local posteriors (Wu and Robert, 2017; Dai et al., 2021). In this case U^0 is the global negative log-prior, N_i denotes the number of observations of client i, $U^{i,j}$ is the negative log-likelihood of the j-th data of client i, and $\varpi^i U^0$ is the fraction of the negative log-prior allocated to this client (Rendell et al., 2020).

Federated Averaging Langevin Dynamics (FALD). FALD, proposed in Deng et al. (2021), is an extension to the Bayesian setting of FEDAVG (McMahan et al., 2017). The updates performed on the ith client define a sequence of local parameters $(X_k^i)_{k\in\mathbb{N}}$ which are transmitted according to some preset schedule (which is deterministic in Deng et al. (2021) and is random in this work) to a central server. The central server averages the local parameters to update the global parameter. This global parameter is finally transmitted back to each client, and is used as a starting point of a new round of local iterations. Hence, each iteration $k \geq 0$ of FALD can be decomposed into two steps:

(1) **Local iteration on each client.** Each client i performs one step of the Langevin Monte Carlo algorithm (Grenander and Miller, 1994; Roberts and Tweedie, 1996) with a stochastic gradient associated with its local potential:

$$\begin{aligned} G_{k+1}^{i} &= \widehat{\nabla} U_{k+1}^{i}(X_k^i) , \\ \tilde{X}_{k+1}^{i} &= X_k^i - \gamma G_{k+1}^{i} + \sqrt{2\gamma}\, Z_{k+1}^{i} , \end{aligned} \qquad (2)$$

where $\gamma > 0$ and for $x \in \mathbb{R}^d$, $\widehat{\nabla} U_{k+1}^i(x)$ is an unbiased estimator of $\nabla U^i(x)$ given by (see Welling and Teh (2011) – general updates are considered in the supplement)

$$\widehat{\nabla} U_{k+1}^{i} = \varpi^{i} \nabla U^{0} + (N_i/n_i) \textstyle\sum_{j\in S_{k+1}^{i}} \nabla U^{i,j} , \qquad (3)$$

where $(S_k^i)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. uniform random subsets of $[N_i]$ of cardinal number n_i. Moreover, $(Z_k^i)_{k\in\mathbb{N}^*}$, $i \in [b]$ are sequence of i.i.d Gaussian random variables which might be correlated across the agents and the central server. More precisely, given independent sequences, $(\tilde{Z}_k^i)_{k\in\mathbb{N}^*}$, $i \in [b]$ and $(\tilde{Z}_k)_{k\in\mathbb{N}^*}$ of i.i.d. d-dimensional standard Gaussian random variables, for $\tau \in [0,1]$ we set

$$Z_k^i = \sqrt{\tau}\, \tilde{Z}_k + \sqrt{1-\tau}\, \tilde{Z}_k^i . \qquad (4)$$

(2) **A local update.** With probability $p_c \in (0,1]$, the ith client communicates its parameter $\tilde{X}_{k+1}^i$, resulting from the first step, to the central server which in turns broadcasts the average $X_{k+1} = b^{-1} \sum_{i\in[b]} \tilde{X}_{k+1}^i$. Finally, each client updates its parameter as $X_{k+1}^i = X_{k+1}$. When no communication is performed, each client updates its parameter as $X_{k+1}^i = \tilde{X}_{k+1}^i$.

The local recursions defined by FALD can be written for $i \in [b]$ and $k \geq 0$ as

$$X_{k+1}^{i} = (1 - B_{k+1}) \tilde{X}_{k+1}^{i} + (B_{k+1}/b) \textstyle\sum_{j\in[b]} \tilde{X}_{k+1}^{j} , \qquad (5)$$

where $(B_k)_{k\in\mathbb{N}^*}$ is a sequence of i.i.d. Bernoulli random variables with parameter p_c.

For $k \geq 1$, denote by $\mu_k^{(\mathrm{F})}$ the distribution of the average parameter

$$X_k = (1/b) \textstyle\sum_{i\in[b]} X_k^i . \qquad (6)$$

Non-asymptotic Wasserstein bounds between $\mu_k^{(\mathrm{F})}$ and the target distribution π are established in Theorem 1 under the following assumptions.

A1. *For any $i \in [b]$, U^i is continuously differentiable. In addition, there exist $m, L > 0$ such that for any $i \in [b]$, the function U^i is L-smooth and m-strongly convex, i.e., for any $x, x' \in \mathbb{R}^d$,*

$$\begin{aligned} (m/2)\|x' - x\|^2 \leq U^i(x') - U^i(x) - \langle \nabla U^i(x), x' - x \rangle \\ \leq (L/2)\|x' - x\|^2 . \end{aligned}$$

A2. *For any $i \in [b]$, $(\{\widehat{\nabla} U_k^i\}_{i\in[b]})_{k\in\mathbb{N}}$ are i.i.d. unbiased estimates of $\{\nabla U^i\}_{i\in[b]}$. In addition, there exists $\hat{L} \geq 0$ such that for any $x, x' \in \mathbb{R}^d$ we have*

$$\mathbb{E}\left[\|\widehat{\nabla} U_k^i(x') - \widehat{\nabla} U_k^i(x)\|^2\right] \leq \hat{L}^2 \|x' - x\|^2 .$$

In the mini-batch scenario (3), **A**2 is satisfied if for $i \in [b]$, $j \in [N_i]$ there exists $L_j^i \geq 0$ such that for any $x, x' \in \mathbb{R}^d$, $\|\nabla U^{i,j}(x') - \nabla U^{i,j}(x)\| \leq L_j^i \|x' - x\|$.

Finally, we also consider the following optional smoothness condition on the potentials $\{U^i\}_{i\in[b]}$. This additional assumption, often satisfied in applications have been considered e.g. in Durmus and Moulines (2019); Dalalyan and Karagulyan (2019).

HX1. *There exists $\tilde{L} \geq 0$, such that for any $i \in [b]$, the function U^i is three times continuously differentiable and for any $x, x' \in \mathbb{R}^d$, $\|\nabla^2 U^i(x) - \nabla^2 U^i(x')\| \leq \tilde{L}\|x - x'\|$.*

We introduce some key quantities appearing in the theoretical derivations below. Denote by $x_\star$ the minimizer of $\sum_{i\in[b]} U^i$ which exists and is unique under **A**1. We define

$$\begin{aligned} \mathsf{V}_\pi &= \textstyle\int_{\mathbb{R}^d} \mathrm{Var}\{b^{-1} \sum_{i\in[b]} \widehat{\nabla} U_1^i(x)\} \pi(\mathrm{d}x) , \\ \mathsf{V}_\star &= \mathrm{Var}\{b^{-1} \textstyle\sum_{i\in[b]} \widehat{\nabla} U_1^i(x_\star)\} , \end{aligned} \qquad (7)$$

the average of the stochastic gradient variance under the stationary distribution π and at the minimum $x_\star$, respectively. Finally, the statistical heterogeneity between the clients is quantified by (see, e.g. Stich et al. (2018))

$$\mathsf{H} = b^{-1} \sum_{i \in [b]} \|\nabla U^i(x_\star)\|^2 \,.$$

For ease of presentation, for two sequences $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ we write $a_k \lesssim b_k$ if there exists $C > 0$ only depending on the constants introduced in **A**1, **A**2 and **HX**1 such that $a_k \leq C b_k$, for any $k \in \mathbb{N}$.

Theorem 1 (Simplified)**.** *Assume **A**1, **A**2 and suppose for any $i \in [b]$, $X_0^i = X_0$. Then, there exist $\bar\gamma > 0$, such that for any $\gamma \in (0, \bar\gamma]$, $k \in \mathbb{N}$, $X_0 \sim \mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, we have*

$$\mathbf{W}_2^2(\mu_k^{(\mathrm{F})}, \pi) \lesssim (1 - \gamma m/8)^k\, \mathsf{I}(\mu_0) + \frac{\gamma^e}{b}\mathsf{J} + \gamma \mathsf{V}_\pi + \frac{\gamma^2(1-p_{\mathrm{c}})}{p_{\mathrm{c}}^2}\left\{\mathsf{H} + p_{\mathrm{c}}\mathsf{V}_\star + \frac{d}{b}\right\} + \frac{\gamma(1-\tau)(1-b^{-1})d}{p_{\mathrm{c}}} \,,$$

*where $\mathsf{J} = d$, $e = 1$ and $\mathsf{I}(\mu_0) < \infty$ is a function of the initial condition μ_0. If **HX**1 holds, then $e = 2$ and $\mathsf{J} = d(1 + d/b)$.*

Elements of proof are provided in Section 3; a precise statement is given in Theorem 20 with detailed proofs. Note the step size upper bound $\bar\gamma$ is proportional to p_{c}. In the single user case ($b = p_{\mathrm{c}} = \tau = 1$), we recover up to numerical constants the results stated in Durmus and Moulines (2019); Dalalyan and Karagulyan (2019). Note that, under **HX**1 the leading term in the step size γ is proportional to the stochastic gradient variance V_π, in accordance with the bounds obtained for SGLD by e.g., Dalalyan and Karagulyan (2019). More discussions on these bounds are postponed after the statement of Theorem 3.

Lower bounding the effect of heterogeneity. Similar to FedAvg, the convergence of FALD is impaired by data heterogeneity. Multiple local SGLD steps described in (2) cause X_k^i to target the local posteriors $\pi^i \propto \exp(U^i)$. We now provide lower bound on the Wasserstein distance between the distribution of the samples generated by FALD and the target distribution π which is proportional to the heterogeneity $\gamma^2\mathsf{H}$.

Proposition 2. *There exist $\bar\gamma > 0$, potentials $\{U^i\}_{i=1}^2$ on $\mathbb{R}$ satisfying **A**1, **HX**1 and an instance of* FALD *satisfying **A**2 such that for any $\gamma \in (0, \bar\gamma]$, we have*

$$\liminf_{k\to+\infty} \mathbf{W}_2^2(\mu_k^{(\mathrm{F})}, \pi) \gtrsim \gamma^2 \mathsf{H} \,.$$

This proposition extends Karimireddy et al. (2020, Theorem II) to the Bayesian context and underlines the same limitation as FedAvg. To circumvent this, various bias reduction techniques have been suggested in the stochastic optimization literature (Horváth et al., 2022; Gorbunov et al., 2021). In the next section, we adapt similar mechanisms to derive an alternative to FALD satisfying better finite bounds.

FALD with control variates and bias reduction. To mitigate the impact of local stochastic gradients, we adapt variance-reduction techniques (Wang et al., 2013; Kovalev et al., 2020) and bias-reduction techniques (Horváth et al., 2022; Gorbunov et al., 2021). This new approach introduces a different recursion rule in step (1) of FALD, while keeping step (2) unchanged. The local update rule is based on a reference point $Y_k \in \mathbb{R}^d$ common to all clients. This common point is updated with probability $q_{\mathrm{c}} \in (0, 1]$ and allows the inclusion of a local shift C_k to recenter the local gradients. This mechanism eliminates the “infamous non-stationarity of the local methods” (paraphrasing Gorbunov et al. (2021)) and therefore avoids extra bias. At each iteration k, the first step of the VR-FALD* algorithm is divided into two parts:

(1.1) **Update of the reference parameter and control variate.** The variance reduced gradient requires a sporadic computation of the full local gradient. Let $(B_k^Y)_{k\in\mathbb{N}^*}$ be a sequence of i.i.d. Bernoulli random variables with parameter $q_{\mathrm{c}} \in (0, 1]$. If $B_{k+1}^Y = 1$, then the client reference point Y_k is updated: the clients transmit their local parameter $\{X_k^i\}_{i\in[b]}$ to the central server which computes their average $Y_{k+1} = b^{-1}\sum_{i\in[b]} X_k^i$; which is sent back to the clients. The clients then compute the full gradients $\{\nabla U^i(Y_{k+1})\}_{i\in[b]}$ and transmit them to the central server which updates the shift $C_{k+1} = b^{-1}\sum_{i\in[b]} \nabla U^i(Y_{k+1})$. To summarize, the reference point and the shift are updated according to

$$\begin{aligned} Y_{k+1} &= (1 - B_{k+1}^Y) Y_k + (B_{k+1}^Y / b) \textstyle\sum_{i\in[b]} X_k^i \,, \\ C_{k+1} &= (1 - B_{k+1}^Y) C_k + (B_{k+1}^Y / b) \textstyle\sum_{i\in[b]} \nabla U^i(Y_{k+1}) \,. \end{aligned} \tag{8}$$

(1.2) **Local iteration on each client.** This step is similar to FALD, upon replacing the local updates (2) by the variance-reduced version

$$G_{k+1}^i = \widehat{\nabla} U_{k+1}^i(X_k^i) - \widehat{\nabla} U_{k+1}^i(Y_k) + C_k \,, \tag{9}$$

$$\tilde{X}_{k+1}^i = X_k^i - \gamma G_{k+1}^i + \sqrt{2\gamma} Z_{k+1}^i \,. \tag{10}$$

The VR-FALD* analysis relies on the following additional assumption.

A3. *There exists $\omega \geq 0$ such that for any $i \in [b]$, $k \in \mathbb{N}^\star$ and $x, y \in \mathbb{R}^d$, the following inequality holds*

$$\mathbb{E}\left[\|\widehat{\nabla} U_k^i(x) - \widehat{\nabla} U_k^i(y) - \nabla U^i(x) + \nabla U^i(y)\|^2\right] \leq \omega \|x - y\|^2 \,.$$

Under **A**1 and **A**2, **A**3 is satisfied with $\omega = 2L^2 + 2\hat{L}^2$. However, using this result leads to some discrepancy in previous existing analysis, since $\omega = 0$ in the non-stochastic gradient case while $2L^2 + 2\hat{L}^2 \neq 0$ in general. Finally, in the mini-batch scenario (3), if $\{\nabla U^{i,j}\}_{j\in[N_i]}$ are L_i-Lipschitz, then **A**3 holds with $\omega = \max_{i\in[b]}\{N_i L_i^2 / n_i\}$; see Remark 15.

For $k \geq 0$, denote by $\mu_k^{(\mathrm{Vr}\star)}$ the distribution of the average $X_k = b^{-1} \sum_{i \in [b]} X_k^i$ where X_k^i is defined as in (5) with $\tilde{X}_k^i$ given in (10). With these notations, we obtain the following theoretical guarantee on VR-FALD*.

Theorem 3 (Simplified). *Assume **A**1, **A**2, **A**3 and suppose for $i \in [b]$, $X_0^i = Y_0 = X_0$. Then, there exist $\bar{\gamma}^{\mathrm{Vr}\star} > 0$, such that for any $q_{\mathrm{c}} \leq p_{\mathrm{c}}$, $\gamma \in \left(0, \bar{\gamma}^{\mathrm{Vr}\star}\right]$, $k \in \mathbb{N}$, $X_0 \sim \mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, we have*

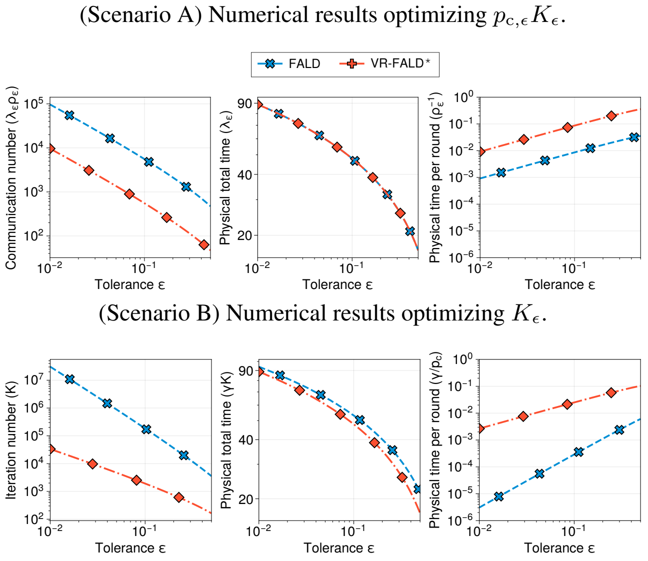
$$\mathbf{W}_2^2(\mu_k^{(\mathrm{Vr}\star)}, \pi) \lesssim (1 - \gamma m/8)^k \, \mathsf{I}^{\mathrm{Vr}\star}(\mu_0) + \frac{\gamma^e}{b} \mathsf{J} + \frac{\gamma^2 d}{b q_{\mathrm{c}}} \omega$$
$$+ \frac{\gamma(1-\tau)(1-b^{-1})d}{p_{\mathrm{c}}} + \frac{\gamma^2(1-p_{\mathrm{c}})}{p_{\mathrm{c}}^2} \left\{ \gamma \mathsf{V}_\star + \frac{d}{b} \right\},$$

*where $\mathsf{J} = d$, $e = 1$, $\mathsf{V}_\star$ is defined in (7), $\mathsf{I}^{\mathrm{Vr}\star}(\mu_0) < \infty$ is a function of the initial condition μ_0. If **HX**1 holds, then $e = 2$ and $\mathsf{J} = d(1 + d/b)$.*

The proof is postponed to Section 7.2. Compared to Theorem 1, the client-drift term does no longer appear, highlighting the advantage of VR-FALD* in dealing with data heterogeneity between agents.

Further, the variance of the stochastic gradients of VR-FALD* only appear in the factor $\gamma^2 \omega$. This result agrees with Chatterji et al. (2018) for SVRG-LD, which might be seen as a particular instance of VR-FALD* with $b = 1$, $p_{\mathrm{c}} = 1$. Nevertheless, a close inspection of the proof in Chatterji et al. (2018) reveals a gap—see Remark 31, which is corrected in the proof of Theorem 30.

Complexity and Communication costs. We now discuss the complexity and communication costs of FALD and VR-FALD*. We study two extreme cases: (A) the local computation cost is negligible and only the communication cost matters, which is typical in cross-device applications. (B) the communication cost is negligible and only the local computation cost (complexity) matters. More general scenarios are discussed in the supplement Section 9. In this discussion, it is assumed that **HX**1 is satisfied and $\tau = 1$. In both cases, for a target precision $\epsilon > 0$, we optimize the hyperparameters (number of iterations K_ϵ, learning rate γ_ϵ, probability of communication $p_{\mathrm{c},\epsilon}$) to ensure $\mathbf{W}_2(\mu_{K_\epsilon}^{(\mathrm{F})}, \pi) \leq \epsilon$ (FALD) or $\mathbf{W}_2(\mu_{K_\epsilon}^{(\mathrm{Vr}\star)}, \pi) \leq \epsilon$ (VR-FALD*). The values of the parameters d, m, ω, H, J, V_π and $\mathsf{V}_\star$ are reported in Table 5.
(Scenario A) The objective is to minimize the number of communications $p_{\mathrm{c},\epsilon} K_\epsilon$. As γ can be arbitrarily small, we set $K_\epsilon = \gamma^{-1} \lambda_\epsilon$, $p_{\mathrm{c},\epsilon} = \rho_\epsilon \gamma$, where $\lambda_\epsilon, \rho_\epsilon > 0$. Hence, the optimization problem becomes $\min\{\lambda_\epsilon \rho_\epsilon\}$ subject to $\mathsf{I}(\mu_0) \exp(-\lambda_\epsilon m/8) + \rho_\epsilon^{-2}(\mathsf{H} + d/b) \leq \epsilon^2$. As $\epsilon \downarrow 0^+$, the minimum number of communications $p_{\mathrm{c},\epsilon} K_\epsilon$ scales as $\tilde{O}(\epsilon^{-1}\sqrt{\mathsf{H} + b^{-1}d})$ for FALD and $\tilde{O}(\epsilon^{-1}\sqrt{b^{-1}d})$ for VR-FALD*.
(Scenario B) We take $p_{\mathrm{c},\epsilon} = 1$ and seek to minimize the total number of iterations K_ϵ. As $\epsilon \downarrow 0^+$, K_ϵ scales as $\tilde{O}(\epsilon^{-2}(\mathsf{V}_\pi + \epsilon\sqrt{b^{-1}\mathsf{J}}))$ for FALD and $\tilde{O}(\epsilon^{-1}\sqrt{b^{-1}\mathsf{J} + b^{-1}\omega d})$ for VR-FALD*.


(Scenario A) Numerical results optimizing $p_{\mathrm{c},\epsilon} K_\epsilon$.

(Scenario B) Numerical results optimizing K_ϵ.

Figure 1: Complexity and Communication costs.

In Figures 1a-1b, we display the optimal number of communications $p_{\mathrm{c},\epsilon} K_\epsilon$ as a function of ϵ (left panels Figures 1a-1b). We also exhibit the *physical time* which corresponds to the time of the Langevin diffusion. The total physical times – λ_ϵ for (A) and $\gamma_\epsilon K_\epsilon$ for (B) – are displayed in the middle panels Figures 1a-1b. Finally, the right panels Figures 1a-1b represent the average physical time between two consecutive communications — ρ_ϵ^{-1} for (A) and $\gamma / p_{\mathrm{c},\epsilon}$ for (B). Note that, the total physical time is (almost) the same for FALD, VR-FALD*, in scenarios (A) and (B). VR-FALD* significantly reduces the number of communications $p_{\mathrm{c},\epsilon} K_\epsilon$ in scenario (A) (top panel) and number of rounds K_ϵ (B) (bottom panel) w.r.t. FALD.

Figures 1a-1b also illustrate that the "embarrassingly parallel" approach of (Neiswanger et al., 2014) is far from optimal. Indeed, our results show the importance of making multiple interactions (rather than a single consensus step) and using correlated noises between clients. In scenario (A), the optimal number of communications scales inversely proportional to $1/\epsilon$ which improve the bounds $\tilde{O}(1/\epsilon^2)$ derived in Deng et al. (2021, Section 5.3.1). For scenario (B), FALD has the same complexity as QLSD Vono et al. (2022b) under similar assumptions; see also Sun et al. (2022). VR-FALD* has the lowest complexity ($\tilde{O}(1/\epsilon)$) among the Bayesian Federated algorithms reported earlier. This bound matches the one obtained by Chatterji et al. (2018) for the fully centralized SVRG-LD (corresponding to $b = 1$).

3 Proofs outline

We briefly outline the main steps of the proof of Theorems 1 and 3. Details of the proofs can be found in the supplementary paper, where we analyze the two algorithms under a common unifying framework. For both algorithms, the local parameters $(X_k^i)_{i \in [b]}$, $k \geq 0$, are given by (5), where $(\tilde{X}_k^i)_{i \in [b]}$ stands for local iterations, which are given in (2) for FALD and (9) for VR-FALD*. Then, we bound the Wasserstein distance between the target distribution π and the distribution of $X_k = b^{-1} \sum_{i \in [b]} X_k^i$ which is denoted by $(\mu_k^{(\gamma)})_{k \in \mathbb{N}}$. The Wasserstein distance is defined as the infimum over the coupling. We use below the synchronous coupling construction used in (Durmus and Moulines, 2019; Dalalyan and Karagulyan, 2019) for the analysis of Stochastic Gradient Langevin algorithms.

Synchronous coupling. We first construct a Brownian motion $(\mathsf{W}_t)_{t \geq 0}$ by $\mathsf{W}_t = \sqrt{\tau}\, \tilde{\mathsf{W}}_t + \sqrt{(1-\tau)/b} \sum_{i \in [b]} \tilde{\mathsf{W}}_t^i$, starting from $b+1$ independent d-dimensional standard Brownian motions $(\tilde{\mathsf{W}}_t^i)_{t \geq 0}$, $i \in [b]$, and $(\tilde{\mathsf{W}}_t)_{t \geq 0}$. Second, we define the following standard Gaussian random variables $\tilde{Z}_{k+1}^i = \gamma^{-1/2}(\tilde{\mathsf{W}}_{(k+1)\gamma}^i - \tilde{\mathsf{W}}_{k\gamma}^i)$, $\tilde{Z}_{k+1} = \gamma^{-1/2}(\tilde{\mathsf{W}}_{(k+1)\gamma} - \tilde{\mathsf{W}}_{k\gamma})$, and we set Z_k^i as in (4). For $k \in \mathbb{N}$, it holds that $\sqrt{\gamma} \sum_{i \in [b]} Z_{k+1}^i = \sqrt{b}(\mathsf{W}_{(k+1)\gamma} - \mathsf{W}_{k\gamma})$. Finally, we consider $(\mathsf{X}_t)_{t \geq 0}$ the strong solution of the Langevin diffusion associated with π and starting from $\mathsf{X}_0 \sim \pi$ (see (1)) and driven by $(\mathsf{W}_t)_{t \geq 0}$:

$$\mathrm{d}\mathsf{X}_t = -(1/b) \sum_{i \in [b]} \nabla U^i(\mathsf{X}_t)\, \mathrm{d}t + \sqrt{2/b}\, \mathrm{d}\mathsf{W}_t \,. \quad (11)$$

Under **A**1 and **A**2, π is the unique stationary distribution for the Langevin diffusion, hence the distribution of X_t is π for all $t \geq 0$; see e.g. Roberts and Tweedie (1996). Hence, $(X_k, \mathsf{X}_{k\gamma})$ defines a coupling between $\mu_k^{(\gamma)}$ and π, thus for any $k \in \mathbb{N}$ we get

$$\mathbf{W}_2^2(\mu_k^{(\gamma)}, \pi) \leq \mathbb{E}\left[\|X_k - \mathsf{X}_{k\gamma}\|^2\right] \,.$$

The rest of the proof then consists in bounding the right-hand side. It is worth noting that in contrast to most analysis on Langevin dynamics, we consider a Langevin diffusion (11) we scale the gradient term by b^{-1} and the Brownian motion by $b^{-1/2}$. This scaling is adapted to the averaging procedure defining $(X_k)_{k \in \mathbb{N}}$.

Decomposition of $\mathbb{E}[\|X_k - \mathsf{X}_{k\gamma}\|^2]$. Denote by $\mathcal{F}_k$ the filtration generated by X_0, $(\mathsf{W}_t)_{t \leq k\gamma}$ and $(\{X_l^i\}_{i=1}^b)_{l \leq k}$. Using the definition (6) of $(X_k)_{k \in \mathbb{N}}$ combined with **A**1, we show in Proposition 4 that for any $\gamma \lesssim 1$

$$\begin{aligned} \mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] \lesssim (1 - \gamma m/2)\, \|\mathsf{X}_{k\gamma} - X_k\|^2 \\ + E_k + \gamma^2 S_k + V_k \,, \quad (12) \end{aligned}$$

where $V_k = b^{-1} \sum_{i \in [b]} \|X_k^i - X_k\|^2$ and

$$\begin{aligned} S_k &= \mathrm{Var}^{\mathcal{F}_k}(b^{-1} \textstyle\sum_{i \in [b]} G_k^i) \,, \\ E_k &= \gamma^{-1} \|\mathbb{E}^{\mathcal{F}_k}[I_k]\|^2 + \mathbb{E}^{\mathcal{F}_k}\left[\|I_k\|^2\right] \,, \end{aligned}$$

with $I_k = b^{-1} \sum_{i \in [b]} \int_{k\gamma}^{(k+1)\gamma} (\nabla U^i(\mathsf{X}_s) - \nabla U^i(\mathsf{X}_{k\gamma}))\mathrm{d}s$.

Bounding E_k. The term E_k accounts for the difference between the diffusion and its discretization; the bound is the same for FALD and VR-FALD*. By adapting Durmus and Moulines (2019, Lemma 21), we establish in Lemma 7 that

$$\mathbb{E}\left[E_k\right] \lesssim \gamma^2 d/b \,. \quad (13)$$

Under **HX**1, for $\gamma \lesssim 1$ the bound can be sharpened in

$$\mathbb{E}\left[E_k\right] \lesssim (\gamma^3 d/b)(1 + d/b) \,. \quad (14)$$

The right-hand side of (13) has a higher order with respect to the step size γ in comparison to (14). This step is the reason why we consider the more restrictive assumption **HX**1, which leads to different guarantees depending on whether this condition is met or not.

Bounding S_k. S_k is the conditional variance of the stochastic gradient. This is the main difference between the two algorithms. For FALD, we show in Lemma 19 that

$$\mathbb{E}\left[S_k\right] \lesssim \mathbb{E}\left[\|X_k - \mathsf{X}_{k\gamma}\|^2\right] + \mathbb{E}\left[V_k\right] + \mathsf{V}_\pi \,. \quad (15)$$

On the other hand, under **A**3, we establish in Lemma 27 that for VR-FALD*, it holds that

$$\begin{aligned} \mathbb{E}\left[S_k\right] \lesssim \omega \mathbb{E}\left[\|X_k - \mathsf{X}_{k\gamma}\|^2\right] + \omega \mathbb{E}\left[V_k\right] + \frac{\gamma \omega d}{b q_{\mathrm{c}}} \\ + \omega q_{\mathrm{c}} \sum_{l=0}^{k-1} (1 - q_{\mathrm{c}})^{k-l-1} \mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right] \,. \end{aligned}$$

Compared to the inequality (15), which holds for FALD, the variance term V_π for VR-FALD* is replaced by $\gamma \omega d / b q_{\mathrm{c}}$, which can be made arbitrarily small with $\gamma \to 0$. Note that this term is inversely proportional to the update probability q_{c} of the control variate. Interestingly, the term S_k vanishes when $\omega = 0$, i.e., when each client uses its full local gradient at each iteration.

Bounding V_k. We show in Lemma 18 (FALD) and Lemma 26 (VR-FALD*), there exist $a_0, a_1 \geq 0$ satisfying

$$\mathbb{E}[V_k] \leq (1 - \gamma m/8)^k a_0 + a_1 \,. \quad (16)$$

To establish this result, we consider the sequence $(f_k)_{k \in \mathbb{N}}$ with general term given by

$$f_k = V_k + \alpha_d d_k^2 + \alpha_\sigma \sigma_k^2 \,,$$

where $\alpha_d, \alpha_\sigma \geq 0$ are given in (97); $d_k = \|X_k - x_\star\|$ denotes the distance between the average parameter X_k and the minimizer $x_\star$ of the global potential U; $\sigma_k = 0$ for FALD and $\sigma_k^2 = b^{-1} \sum_{i \in [b]} \mathbb{E}^{\mathcal{F}_k}[\|\widehat{\nabla} U_k^i(Y_k) - \widehat{\nabla} U_k^i(x_\star)\|^2]$ for VR-FALD* with Y_k defined in (8). The weights α_d, α_σ are tailored to prove a contraction; more precisely, we show the existence of $a_2 > 0$ whose expression is given in Lemma 12, such that

$$f_{k+1} \leq (1 - \gamma m/4)\, f_k + \gamma^2 a_2 + 2\gamma d\, (1 - \tau)\, (1 - b^{-1}) \,.$$

An immediate induction combines with $V_k \leq f_k$ yields a first bound for $\mathbb{E}[V_k]$ of the form (16) with a_1 of order γ. In a final step Lemma 10, we refine this bound to obtain a term a_1 of order γ^2.

Gathering all the bounds. The proof is concluded by plugging the upper bounds derived for E_k, S_k, V_k into (12).

4 Numerical experiments

To illustrate our findings, we perform three numerical experiments on both synthetic toy-examples and real datasets. We compare FALD, VR-FALD* with Bayesian federated learning benchmarks: DG-LMC (Plassier et al., 2021), the Federated Stochastic Langevin Dynamics FSGLD (El Mekkaoui et al., 2021), the Quantized Langevin Stochastic Dynamic QLSD and its variance-reduced version QLSDPP (Vono et al., 2022b). We also include in our benchmark state of the art (centralized MCMC) algorithms: HMC (Brooks et al., 2011), the Stochastic Gradient Langevin Dynamics (SGLD) (Welling and Teh, 2011) and the preconditioned SGLD (pSGLD) (Li et al., 2016).

Gaussian posterior. We consider $b = 100$ clients associated to local Gaussian potentials with mean $\{\mu_i\}_{i\in[b]}$ and covariance $\{\Sigma_i\}_{i\in[b]}$, i.e., $U^i(x) = (1/2)(x-\mu_i)^\top\Sigma_i^{-1}(x-\mu_i)$. For different values of the hyperparameters (p_c, γ, τ), we run 100 chains with $k_1 = 10^7$ iterations $(X_k)_{k=1}^{k_1}$ and discard 10% of the samples (more details are reported in Section 10.1). For each chain, we estimate the posterior variance $\sigma_\star^2 = \int \|x - x_\star\|^2 \mathrm{d}\pi(x)$ using FALD and VR-FALD*, where $\pi \propto \exp(-\sum_{i\in[b]} U^i)$ and $x_\star = \arg\max_{x\in\mathbb{R}^d} \pi(x)$. We compute a Monte-Carlo estimates (over 10^2 independent replications) of the Mean Squared Error (MSE) given by $\{(k_1-k_0)^{-1}\sum_{k=k_0+1}^{k_1}\|X_k - x_\star\|^2 - \sigma_\star^2\}^2$ where k_1 is the total number of samples and k_0 is the burn-in period. The values of the hyperparameters are reported in Section 10.1. From Table 1, VR-FALD* always outperforms FALD for any choices of p_c, γ. This illustrates the impact of the heterogeneity and supports the theoretical findings given in Theorems 1 and 3. Furthermore, the asymptotic bias for VR-FALD* improves when $\tau = 1$ as derived in the theoretical analysis.

Bayesian Logistic Regression. We assess the performance of FALD and VR-FALD* using calibration metrics—the expected calibration error (ECE), the Brier score (BS), and the negative log likelihood (nNLL); see Guo et al. (2017)—and predictive accuracy. We consider Bayesian logistic regression applied to the Titanic dataset, which consists of $p = 2$ classes with $N = 2201$ samples in dimension $d = 4$. This dataset is allocated between $b = 10$ clients in a very heterogeneous manner, as displayed in Figure 3. We use an isotropic Gaussian prior with a mean of zero and variance 1. We also report the total variation distance between the predictive distribution obtained for FALD and VR-FALD* to the predictive distribution approximated by 100 long runs of Langevin Stochastic Dynamics (LSD). These metrics are evaluated on a test data sets of 441 samples, and the mean and standard deviation are reported in Table 2. Moreover, we illustrate the quality improvement of VR-FALD* over FALD in Figure 4. We compared the Wasserstein distance using POT (Flamary et al., 2021) between the empirical distributions generated by FALD, VR-FALD* to the estimated target distribution. Based on the same samples, we compute the relative highest posterior density (HPD) error; see Section 10.2 for details.

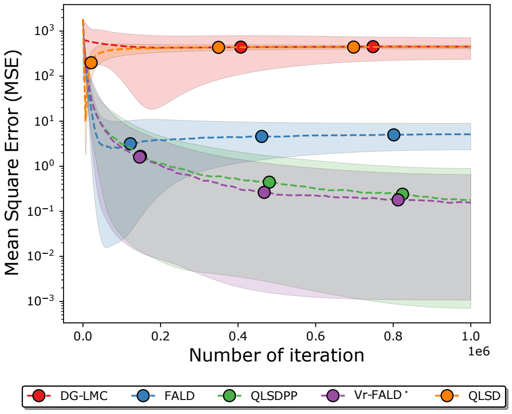


Figure 2: MSE comparison with $p_c = 1/5$ and $\gamma = \bar{\gamma}/3$.

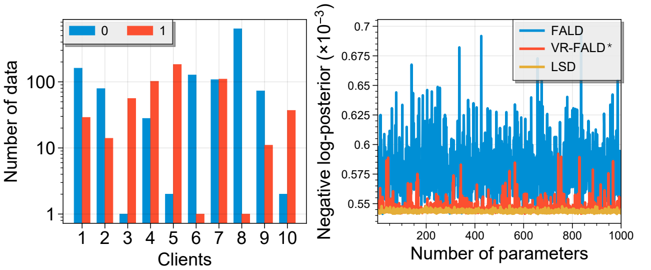


Figure 3: Logistic regression – dataset distribution (Log Scale) and negative log-posterior (right).

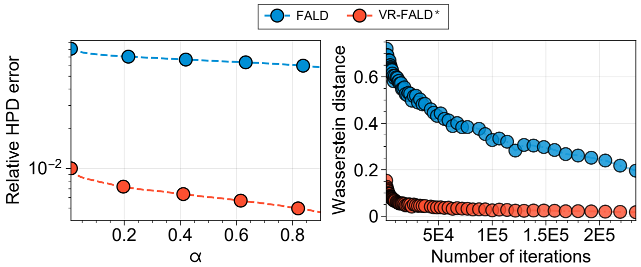


Figure 4: Logistic regression – HPD relative error (left) and Wasserstein distance (right).

Probability p_c	$p_c = 1/5$			$p_c = 1/10$			$p_c = 1/20$		
Stepsize γ	$\frac{1}{2}p_c\bar{\gamma}$	$\frac{1}{5}p_c\bar{\gamma}$	$\frac{1}{10}p_c\bar{\gamma}$	$\frac{1}{2}p_c\bar{\gamma}$	$\frac{1}{5}p_c\bar{\gamma}$	$\frac{1}{10}p_c\bar{\gamma}$	$\frac{1}{2}p_c\bar{\gamma}$	$\frac{1}{5}p_c\bar{\gamma}$	$\frac{1}{10}p_c\bar{\gamma}$
FALD ($\tau = 0$)	2.5E+01	9.5E-01	3.9E-02	3.6E+01	1.1E+00	8.2E-02	4.2E+01	2.0E+00	1.1E-01
VR-FALD* ($\tau = 0$)	4.8E-02	2.6E-02	1.4E-02	5.0E-02	4.9E-02	3.7E-02	9.8E-02	5.3E-02	3.9E-02
VR-FALD* ($\tau = 1$)	2.8E-02	2.0E-02	1.3E-02	4.1E-02	3.7E-02	1.4E-02	8.6E-02	4.3E-02	2.1E-02

Table 1: Asymptotic bias in function of τ, p_c and γ.

Method	Accuracy	Agreement	$10^4\times$ TV	10×ECE	10×BS	10×nNLL
LSD	72.4 ± 0.1	99.9 ± 0.1	5.53 ± 2.00	1.20 ± 0.01	3.44 ± 0.00	5.30 ± 0.00
FALD	77.0 ± 0.8	91.3 ± 0.9	533.32 ± 8.13	1.05 ± 0.09	3.37 ± 0.01	5.19 ± 0.00
VR-FALD*	74.9 ± 0.1	93.6 ± 0.1	287.81 ± 2.04	1.00 ± 0.05	3.51 ± 0.00	5.35 ± 0.00

Table 2: Bayesian Logistic Regression on Titanic.

Bayesian Neural Network: MNIST. To illustrate the behavior of FALD and VR-FALD* in a non-convex setting, we perform Bayesian Neural Network (BNN) inference on the MNIST dataset (Deng, 2012). To this end, we distribute the dataset to $b = 20$ clients as follows: 80% of the data labeled $\mathrm{y} \in \{0, \dots, 9\}$ are equally allocated to clients $i = \mathrm{y} + 1$ and $i = \mathrm{y} + 10$; the remaining data are evenly distributed among the b clients. The likelihood of the observations is computed using LeNet5 neural network (LeCun et al., 1998) with an isotropic Gaussian prior. Finally, we implement FALD and its variants with $p_c = 1/b$ and $q_c = N_b/N_d$, where N_b is the batch size used in the experiments and N_d is the total number of data. All standard deviations and the values of the other parameters are reported in Section 10.3.

In Table 3 we can observe that the best results are obtained by VR-FALD*: it achieves similar performance to the (fully centralized) SGLD and pSGLD. Alleviating client drift using control variates is still effective even in the highly non-convex BNN setting.

Method	SGLD	pSGLD	FALD	VR-FALD*	FSGLD
Accuracy	99.1	99.2	99.1	99.2	98.5
$10^3\times$ECE	6.88	21.6	4.07	4.34	6.34
$10^2\times$BS	1.66	1.45	1.47	1.39	2.39
$10^2\times$nNLL	3.53	4.24	3.06	3.43	4.87

Table 3: Performance of Bayesian FL algorithms on MNIST.

Bayesian Neural Network: CIFAR10. We consider the CIFAR10 dataset (Krizhevsky, 2009) and the ResNet-20 model (He et al., 2016). We split the data across 20 clients, similar to the previous example. Denote by $\mathrm{Y} = \{\mathrm{y}_1, \dots, \mathrm{y}_{10}\}$ the set of labels. Then 80% of the data associated with a label $\mathrm{y}_j \in \mathrm{Y}$, $j \in [10]$, is distributed among clients j and $j + 10$, while the rest of the data is evenly distributed among clients. We assess the performance of FALD and VR-FALD* against HMC, Deep Ensemble, and SGLD. We follow Izmailov et al. (2021) by computing the *accuracy*, *agreement*, and total deviation distance between the predictive distribution. All of these quantities are defined in the Appendix; see Section 10.4. We also report the calibration results and all resulting scores in Table 7; the results for HMC and SGLD are from Izmailov et al. (2021, Table 6). Details on the implementation and choice of hyperparameters can be found in Section 10.4. We can see that VR-FALD* gives very similar results to SGLD and performs favorably in terms of agreement. Finally, FALD and VR-FALD* outperform Deep Ensembles.

Method	HMC	SGD	Deep Ens.	SGLD	FALD	VR-FALD*
Accuracy	89.6	91.57	91.68	89.96	**92.54**	92.03
Agreement	94.0	90.99	91.03	**92.43**	91.53	91.12
10× TV	0.74	1.45	1.49	**1.03**	1.42	1.39
$10^2\times$ECE	5.9	4.71	5.44	4.41	3.79	**3.26**
10×BS	1.4	1.69	1.45	1.53	**1.16**	1.20
10×nNLL	3.07	3.35	3.81	3.15	2.75	**2.63**

Table 4: Performance of Bayesian FL algo. on CIFAR10.

5 Conclusion

In this work, we propose VR-FALD* which extends the FALD Deng et al. (2021) algorithm by introducing control variates to mitigate client drift and reducing stochastic gradient variance. We develop a unifying framework for Bayesian FL combining ideas from Langevin Monte Carlo and Federated Averaging schemes. The theory covers a wide range of local stochastic gradient algorithms; connections can even be made with the global consensus Monte Carlo method (Rendell et al., 2020; Vono et al., 2022a). Using this theoretical framework, we develop non-asymptotic bounds for the algorithms FALD and VR-FALD*, and discuss the choice of hyperparameters (learning rate, communication probability, control variate update probability) to obtain optimal trade-offs. Our analysis allows to correct some errors in the results obtained previously for FALD. The results we obtain on both toy examples and applications to BNNs clearly show the importance of variance reduction and heterogeneity, even when the potential is non-convex.

Acknowledgements. Alain Durmus and Eric Moulines gratefully acknowledge support from the Lagrange Mathematics and Computing Research Center.

References

Ahn, S., Shahbaba, B., and Welling, M. (2014). Distributed Stochastic Gradient MCMC. In *International Conference on Machine Learning*.

Al-Shedivat, M., Gillenwater, J., Xing, E., and Rostamizadeh, A. (2021). Federated Learning via posterior inference: A new perspective and practical algorithms. In *ICLR 2021*.

Alistarh, D., Grubic, D., Li, J., Tomioka, R., and Vojnovic, M. (2017). QSGD: Communication-efficient SGD via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. (2018). signSGD: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Chatterji, N., Flammarion, N., Ma, Y., Bartlett, P., and Jordan, M. (2018). On the theory of variance reduction for stochastic gradient Monte Carlo. In *International Conference on Machine Learning*, pages 764–773. PMLR.

Chen, H.-Y. and Chao, W.-L. (2021). Fedbe: Making Bayesian model ensemble applicable to Federated Learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chittoor, H. H. S. and Simeone, O. (2021). Coded consensus Monte Carlo: Robust one-shot distributed Bayesian learning with stragglers. *arXiv preprint arXiv:2112.09794*.

Clark, D. S. (1987). Short proof of a discrete gronwall inequality. *Discrete applied mathematics*, 16(3):279–281.

Coglianese, C. and Lehr, D. (2016). Regulating by robot: Administrative decision making in the machine-learning era. *Geo. LJ*, 105:1147.

Dai, H., Pollock, M., and Roberts, G. (2021). Bayesian fusion: Scalable unification of distributed statistical analyses. *arXiv preprint arXiv:2102.02123*.

Dalalyan, A. (2017). Further and stronger analogy between sampling and optimization: Langevin Monte Carlo and gradient descent. In *Conference on Learning Theory*, pages 678–689. PMLR.

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

Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron*, 72(2):169–183.

De Souza, D. A., Mesquita, D., Kaski, S., and Acerbi, L. (2022). Parallel MCMC without embarrassing failures. In *International Conference on Artificial Intelligence and Statistics*, pages 1786–1804. PMLR.

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

Deng, W., Ma, Y.-A., Song, Z., Zhang, Q., and Lin, G. (2021). On convergence of federated averaging Langevin dynamics. *arXiv preprint arXiv:2112.05120*.

Douc, R., Moulines, E., Priouret, P., and Soulier, P. (2018). *Markov chains*. Springer.

Dubey, K. A., J Reddi, S., Williamson, S. A., Poczos, B., Smola, A. J., and Xing, E. P. (2016). Variance reduction in stochastic gradient Langevin dynamics. *Advances in neural information processing systems*, 29.

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

El Mekkaoui, K., Mesquita, D., Blomstedt, P., and Kaski, S. (2021). Federated stochastic gradient Langevin dynamics. In *Uncertainty in Artificial Intelligence*, pages 1703–1712. PMLR.

Fatima, M., Pasha, M., et al. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01):1.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8.

Gorbunov, E., Hanzely, F., and Richtárik, P. (2021). Local sgd: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 3556–3564. PMLR.

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

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Haddadpour, F., Kamani, M. M., Mokhtari, A., and Mahdavi, M. (2021). Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2350–2358. PMLR.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

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

Holte, J. M. (2009). Discrete gronwall lemma and applications. In *MAA-NCS meeting at the University of North Dakota*, volume 24, pages 1–7.

Horváth, S., Kovalev, D., Mishchenko, K., Richtárik, P., and Stich, S. (2022). Stochastic distributed learning with gradient quantization and double-variance reduction. *Optimization Methods and Software*, pages 1–16.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. G. (2021). What are Bayesian neural network posteriors really like? In *International Conference on Machine Learning*, pages 4629–4640. PMLR.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for Federated Learning. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR.

Kovalev, D., Horváth, S., and Richtárik, P. (2020). Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pages 451–467. PMLR.

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

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

Lee, S., Park, C., Hong, S.-N., Eldar, Y. C., and Lee, N. (2020). Bayesian Federated Learning over wireless networks. *IEEE Journal on Selected Areas in Communications*.

Li, C., Chen, C., Carlson, D., and Carin, L. (2016). Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. (2019). On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*.

Liu, Q. and Ihler, A. T. (2014). Distributed estimation, information loss and exponential families. *Advances in neural information processing systems*, 27.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32.

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

Mesquita, D., Blomstedt, P., and Kaski, S. (2020). Embarrassingly parallel MCMC using deep invertible transformations. In *Uncertainty in Artificial Intelligence*, pages 1244–1252. PMLR.

Minsker, S., Srivastava, S., Lin, L., and Dunson, D. (2014). Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning*.

Neiswanger, W., Wang, C., and Xing, E. P. (2014). Asymptotically exact, embarrassingly parallel mcmc. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 623–632.

Nemeth, C. and Sherlock, C. (2018). Merging MCMC subposteriors through Gaussian-process approximations. *Bayesian Analysis*, 13(2):507–530.

Nesterov, Y. (2003). *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Plassier, V., Vono, M., Durmus, A., and Moulines, E. (2021). DG-LMC: A turn-key and scalable synchronous distributed MCMC algorithm via Langevin Monte Carlo within gibbs. In *International Conference on Machine Learning*, pages 8577–8587. PMLR.

Rendell, L. J., Johansen, A. M., Lee, A., and Whiteley, N. (2020). Global consensus Monte Carlo. *Journal of Computational and Graphical Statistics*, 30(2):249–259.

Ro, J., Chen, M., Mathews, R., Mohri, M., and Suresh, A. T. (2021). Communication-efficient agnostic federated averaging. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, pages 1753–1757. International Speech Communication Association.

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

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). Bayes and big data: The consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88.

Shlezinger, N., Chen, M., Eldar, Y. C., Poor, H. V., and Cui, S. (2020). Uveqfed: Universal vector quantization for federated learning. *IEEE Transactions on Signal Processing*, 69:500–514.

Smith, L. N. and Topin, N. (2019). Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, page 1100612. International Society for Optics and Photonics.

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. (2018). Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, 31.

Sun, L., Salim, A., and Richtárik, P. (2022). Federated Learning with a sampling algorithm under isoperimetry. *arXiv preprint arXiv:2206.00920*.

Tang, H., Gan, S., Awan, A. A., Rajbhandari, S., Li, C., Lian, X., Liu, J., Zhang, C., and He, Y. (2021). 1-bit adam: Communication efficient large-scale training with adam's convergence speed. In *International Conference on Machine Learning*, pages 10118–10129. PMLR.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Vono, M., Dobigeon, N., and Chainais, P. (2020). Asymptotically exact data augmentation: Models, properties, and algorithms. *Journal of Computational and Graphical Statistics*, 30(2):335–348.

Vono, M., Paulin, D., and Doucet, A. (2022a). Efficient MCMC sampling with dimension-free convergence rate using ADMM-type splitting. *Journal of Machine Learning Research*, 23(25).

Vono, M., Plassier, V., Durmus, A., Dieuleveut, A., and Moulines, E. (2022b). Qlsd: Quantised Langevin Stochastic Dynamics for Bayesian federated learning. In *International Conference on Artificial Intelligence and Statistics*, pages 6459–6500. PMLR.

Wang, C., Chen, X., Smola, A. J., and Xing, E. P. (2013). Variance reduction for stochastic gradient optimization. *Advances in neural information processing systems*, 26.

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

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*.

Wang, X. and Dunson, D. B. (2013). Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*.

Wang, X., Guo, F., Heller, K. A., and Dunson, D. B. (2015). Parallelizing MCMC with random partition trees. In *Advances in Neural Information Processing Systems*.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on International Conference on Machine Learning*, page 681–688. Available at https://www.ics.uci.edu/~welling/publications/papers/stoclangevin_v6.pdf.

Wilson, A. G., Izmailov, P., Hoffman, M. D., Gal, Y., Li, Y., Pradier, M. F., Vikram, S., Foong, A., Lotfi, S., and Farquhar, S. (2021). Evaluating approximate inference in Bayesian deep learning.

Wu, C. and Robert, C. P. (2017). Average of recentered parallel mcmc for big data. *arXiv preprint arXiv:1706.04780*.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., and Khazaeni, Y. (2019). Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pages 7252–7261. PMLR.

Zhang, Y., Liu, D., and Simeone, O. (2022). Leveraging channel noise for sampling and privacy via quantized federated Langevin Monte Carlo.

Federated Averaging Langevin Dynamics: Toward a unified theory and new algorithms — Supplementary Materials

Contents

Notation and convention. The Euclidean norm and the scalar product on $\mathbb{R}^d$ are denoted by $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ respectively. We set $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$ and denote by $\mathbf{N}(m, \Sigma)$ the Gaussian distribution with mean vector m and covariance matrix Σ. Finally, for any $f : \mathbb{R}^d \to \mathbb{R}$ twice continuously differentiable, we define the Laplacian Δf, which for all $x \in \mathbb{R}^d$ is given by $\Delta f(x) = \{\sum_{l=1}^d (\partial^2 f_j)(x)/\partial x_l^2\}_{j=1}^d$.

Theoretical road map. The derivations leading to Theorem 1 and Theorem 3 are split in two sections:

- Section 6 consists of general results under mild assumptions. In this section, we derive an upper bound on V_k – see Section 6.3, and provide a Wasserstein upper bound holding for general federated averaging Langevin schemes in Theorem 8.
- Section 7 is subdivided between the results on FALD (Section 7.1) and VR-FALD* (Section 7.2). In each subsection, we prove intermediate results showing that results of Section 6.3 hold, and finally we apply Theorem 8 to derive the final theoretical guarantees on FALD and VR-FALD*.

6 General scheme and technical results

Problem statement. We consider a general recursion that includes both FALD and VR-FALD*. This general scheme is based on i.i.d. random variables $\{\xi_k : k \in \mathbb{N}\}$ taking values in a measurable space $(\mathsf{E}, \mathcal{E})$ and whose joint distribution is denoted by ν_ξ. Moreover, we introduce a family of measurable functions $\{\mathscr{G}^i : \mathbb{R}^d \times \mathsf{Y}^2 \times \mathsf{C}^2 \times \mathsf{E} \to \mathbb{R}^d, \mathscr{Y}^i : \mathbb{R}^d \times \mathsf{Y}^2 \times \mathsf{E} \to \mathsf{Y}, \mathscr{C}^i : \mathbb{R}^d \times \mathsf{Y} \times \mathsf{C}^2 \times \mathsf{E} \to \mathsf{C}\}_{i=1}^b$, where $(\mathsf{Y}, \mathcal{Y})$ and $(\mathsf{C}, \mathcal{C})$ are measurable spaces. For each $i \in [b]$, the functions $(\mathscr{G}^i, \mathscr{Y}^i, \mathscr{C}^i)$ correspond to the update of the local parameter and control variate by the ith agent. To define the global control variate update, we consider the function $\mathscr{D} : \mathsf{Y} \times \mathsf{C}^{b+1} \times (\mathbb{R}^d)^{b+1} \times \mathsf{E} \to \mathsf{Y} \times \mathsf{C}$. Starting from $\{G_0^i\}_{i=1}^b, \{X_0^i\}_{i=1}^b \in (\mathbb{R}^d)^b$, $(C_0, \{C_0^i\}_{i=1}^b) \in \mathsf{C}^{b+1}$, $(Y_0, \{Y_0^i\}_{i=1}^b) \in \mathsf{Y}^{b+1}$ and set $X_0 = b^{-1} \sum_{i=1}^b X_0^i$. For each $k \in \mathbb{N}$ the random variables are updated according to

$$\begin{aligned}
G_{k+1}^i &= \mathscr{G}^i\left(X_k^i, Y_k^i, Y_k, C_k^i, C_k, \xi_{k+1}\right) , \qquad (17)\\
\tilde{X}_{k+1}^i &= X_k^i - \gamma G_{k+1}^i + \sqrt{2\gamma}\left(\sqrt{\tau/b}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,Z_{k+1}^i\right) ,\\
Y_{k+1}^i &= \mathscr{Y}^i\left(X_k^i, Y_k^i, Y_k, \xi_{k+1}\right) , \qquad (18)\\
C_{k+1}^i &= \mathscr{C}^i\left(X_k^i, Y_k^i, C_k^i, C_k, \xi_{k+1}\right) , \qquad (19)\\
X_{k+1}^i &= B_{k+1} \sum_{j=1}^b \tilde{X}_{k+1}^j + (1 - B_{k+1})\tilde{X}_{k+1}^i , \qquad (20)\\
(Y_{k+1}, C_{k+1}) &= \mathscr{D}(Y_k, C_k, \{C_k^i\}_{i=1}^b, \{X_k^i\}_{i=1}^b, \xi_{k+1}) , \qquad (21)
\end{aligned}$$

where $\tau \in [0, 1]$; $\gamma \in (0, \bar{\gamma}]$ is the stepsize; $\{(B_k, \xi_k, \tilde{Z}_k, Z_k^1, \ldots, Z_k^b) : k \in \mathbb{N}^\star\}$ is a set of independent sequences of i.i.d. random variables such that for any $k \in \mathbb{N}^*$ B_k, is a Bernoulli random variable with parameter $p_{\mathrm{c}} \in (0, 1]$; and $(\tilde{Z}_k, Z_k^1, \ldots, Z_k^b)$ are d-dimensional standard Gaussian random variables. Recall that $(\xi_k)_{k\geq 1}$ is a set of i.i.d. random variables distributed according to ν_ξ such that **H**1 holds to ensure that the combination of functions $\{\mathscr{G}^i\}_{i\in[b]}$ provides an unbiased estimate of ∇U.

In iteration $k \geq 0$, the local parameter of the ith client is denoted by X_k^i, and G_k^i stands for its local gradient. If $B_k = 1$ (communication round), the local parameter X_k^i is set to the value of the global server parameter X_k. If $B_k = 0$, X_k^i is set to the local update $\tilde{X}_k^i$. Moreover, we write Y_k^i the reference point used to compute the control variate C_k^i. The first step (17) corresponds to the computation of a stochastic estimate of ∇U^i by the ith client. Then, the client updates the reference point Y_k^i (18) at which the local control variate is computed. The client also update its own local control variate C_k^i in (19). If $B_{k+1} = 1$, then the server averages the parameter of each client, and broadcasts this average. If $B_{k+1} = 0$, then each client keeps $\tilde{X}_{k+1}^i$ as its new local parameter. Finally, the server updates the reference point Y_k and the global control variate C_k according to (21). Denote the filtration $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$ defined for any $k \geq 0$, by

$$\mathcal{F}_k = \sigma\left(\mathsf{X}_0, \left(B_l, C_l, Y_l, \tilde{Z}_l, \xi_l, \left(C_l^i, G_l^i, X_l^i, \tilde{X}_l^i, Y_l^i, Z_l^i\right)_{i=1,\ldots,n}\right)_{0\leq l\leq k}\right) \tag{22}$$

and consider the conditional expectation and variance denoted by $\mathbb{E}^{\mathcal{F}_k}$, $\mathrm{Var}^{\mathcal{F}_k}(\cdot) = \mathbb{E}^{\mathcal{F}_k}[\|\cdot - \mathbb{E}^{\mathcal{F}_k}[\cdot]\|^2]$ respectively. For $k \in \mathbb{N}$, we introduce X_k the average of the local parameters given by

$$X_k = \frac{1}{b}\sum_{i=1}^b X_k^i \tag{23}$$

and we set

$$V_k = \frac{1}{b}\sum_{i=1}^b \|X_k^i - X_k\|^2 . \tag{24}$$

Finally, to control the distance between the average parameter X_k and the minimizer $x_\star = \arg\min U$, we consider the parameter d_k, which for $k \geq 0$ is given by

$$d_k = \|X_k - x_\star\| . \tag{25}$$

For each $k \in \mathbb{N}$ and $\gamma \in (0, \bar{\gamma}]$, we denote by $\mu_k^{(\gamma)}$ the distribution of X_k defined by (23). To ensure the quality of the samples generated by Algorithm 1, we control the Wasserstein distance $\mathbf{W}_2(\pi, \mu_k^{(\gamma)})$. Recall that the Wasserstein distance

Algorithm 1 Stochastic Averaging Langevin Dynamics - FALD and its variants

```
Input: initial vectors $(X_0^i)_{i\in[b]}$, noise parameter $\tau \in [0,1]$, number of communication rounds $K$, probability $p_{\mathrm{c}} \in (0,1]$ of communication, probability $q_{\mathrm{c}} \in [0,1]$ to update the control variates, and step-size $\gamma$
Initialize: $Y_0 = (1/b)\sum_{i=1}^b X_0^i$ and $C_0 = (1/b)\nabla U(Y_0)$
for $k = 0$ to $K-1$ do
    Draw $B_{k+1} \sim \mathcal{B}(p_{\mathrm{c}}), \tilde{Z}_{k+1} \sim \mathbf{N}(0_d, \mathrm{I}_d)$            // On every client
    for $i = 1$ to $b$ do                                                    // In parallel on the $b$ clients
        Draw $\xi_{k+1}^i \sim \nu_\xi^i$, $\tilde{Z}_{k+1}^i \sim \mathbf{N}(0_d, \mathrm{I}_d)$
        Compute $G_k^i$ following (17)
        Set $\tilde{X}_{k+1}^i = X_k^i - \gamma G_k^i + \sqrt{2\gamma}\,(\sqrt{\tau/b}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i)$
        if $B_{k+1} = 1$ then
            Broadcast $\tilde{X}_{k+1}^i$ to the server                      // Communication round
        else
            Update $X_{k+1}^i \leftarrow \tilde{X}_{k+1}^i$                     // Local step
        if $\tilde{B}_{k+1} = 1$ then                                          // Control variate update round
            Broadcast the necessary information to the server in order to update $(Y_k^i, C_k^i, Y_k, C_k)$
        else
            Set $(Y_{k+1}^i, C_{k+1}^i, Y_{k+1}, C_{k+1}) \leftarrow (Y_k^i, C_k^i, Y_k, C_k)$        // No update
    if $B_{k+1} = 1$ then                                                     // During communication round
        Update then broadcast $X_{k+1} \leftarrow (1/b)\sum_{i=1}^b \tilde{X}_{k+1}^i$    // On the central server
        Update the local parameter $X_{k+1}^i \leftarrow X_{k+1}$               // On every client
    if $\tilde{B}_{k+1} = 1$ then                                             // During control variate update round
        If needed, update then broadcast $Y_{k+1} \leftarrow (1/b)\sum_{i=1}^b X_k^i$   // On the central server
        Update $(Y_k^i, C_k^i)$ using the parameters $(X_k^i, Y_k^i, Y_k, Y_{k+1}, C_k)$    // On every client
        Update then broadcast $C_{k+1} \leftarrow (1/b)\sum_{i=1}^b C_{k+1}^i$    // On the central server
Output: samples $\{X_\ell\}_{\{\ell\in[K]\,:\,B_\ell=1\}}$.
```

is the infimum of $\mathbb{E}[\|\mathsf{X}_{k\gamma} - X_k\|^2]$ over all couplings $(\mathsf{X}_{k\gamma}, X_k)$ such that $\mathsf{X}_{k\gamma}$ is distributed according to π. Thus, to study the convergence of $(\mu_k^{(\gamma)})_{k\in\mathbb{N}}$, we introduce a synchronous coupling $(\mathsf{X}_{k\gamma}, X_k)_{k\geq 0}$ with values in $(\mathbb{R}^d)^2$ between π and $\mu_k^{(\gamma)}$, starting from the couple (X_0, X_0) distributed according to $\zeta \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$, *i.e.*, $\zeta(\mathbb{R}^d, \cdot) = \mu_0^{(\gamma)} \in \mathcal{P}_2(\mathbb{R}^d)$ and $\zeta(\cdot, \mathbb{R}^d) = \pi$. Since $\log \pi$ is supposed m-strongly concave by **A**1, note that π belongs in $\mathcal{P}_2(\mathbb{R}^d)$. Based on independent d-dimensional standard Brownian motions $(\{\tilde{\mathsf{W}}_t, \{\tilde{\mathsf{W}}_t^i\}_{i=1}^b\})_{t\geq 0}$, we define $\mathsf{W}_t = \sqrt{\tau}\tilde{\mathsf{W}}_t + \sqrt{(1-\tau)/b}\sum_{i=1}^b \tilde{\mathsf{W}}_t^i$. For $k \in \mathbb{N}^\star$, we introduce $\tilde{Z}_k = \gamma^{-1/2}(\tilde{\mathsf{W}}_{k\gamma} - \tilde{\mathsf{W}}_{(k-1)\gamma})$, and for $i \in [b]$, we consider $\tilde{Z}_k^i = \gamma^{-1/2}(\tilde{\mathsf{W}}_{k\gamma}^i - \tilde{\mathsf{W}}_{(k-1)\gamma}^i)$. Therefore, for all $k \in \mathbb{N}^\star$ we can verify that $\mathsf{W}_{k\gamma} - \mathsf{W}_{(k-1)\gamma} = \sqrt{\gamma\tau}\tilde{Z}_k + \sqrt{\gamma(1-\tau)/b}\sum_{i=1}^b \tilde{Z}_k^i$. Moreover, consider $(\mathsf{X}_t)_{t\geq 0}$ the strong solution of the Langevin stochastic differential equation (SDE) given by

$$\mathrm{d}\mathsf{X}_t = -\frac{1}{b}\nabla U(\mathsf{X}_t)\,\mathrm{d}t + \sqrt{\frac{2}{b}}\,\mathrm{d}\mathsf{W}_t\,. \tag{26}$$

The Langevin diffusion defines a Markov semigroup $(\tilde{P}_t)_{t\geq 0}$ satisfying $\pi\tilde{P}_t = \pi$ for any $t \geq 0$, see for example Roberts and Tweedie (1996, Theorem 2.1). Note that X_t and X_k are distributed according to π and $\mu_k^{(\gamma)}$, respectively. From the definition of the Wasserstein distance of order 2 it follows that

$$\mathbf{W}_2(\pi, \mu_k^{(\gamma)}) \leq \mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right]^{1/2}\,.$$

So the proof consists mainly of upper bounding the squared norm $\|\mathsf{X}_{k\gamma} - X_k\|$, from which we derive an explicit bound on the Wasserstein distance by the previous inequality.

First upper bound on $\mathbb{E}^{\mathcal{F}_k}[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2]$. Under mild assumptions, we derive a first bound in Proposition 4 to control $\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2$ based on $\|\mathsf{X}_{k\gamma} - X_k\|^2$, $(1/b)\sum_{i=1}^b G_k^i$ and V_k. This decomposition highlights the different approximations brought by the discretization of the Langevin diffusion (26) between the averaged parameter $(X_k)_{k\in\mathbb{N}}$ defined in (23) and $\{\mathsf{X}_{k\gamma}\}_{k\in\mathbb{N}}$. Recall that $x_\star = \arg\min U$ and for all $k \in \mathbb{N}$, consider I_k the approximation error defined by

$$I_k = \int_{k\gamma}^{(k+1)\gamma} \left(\nabla\bar{U}(\mathsf{X}_s) - \nabla\bar{U}(\mathsf{X}_{k\gamma})\right)\mathrm{d}s\,. \tag{27}$$

For $\bar{\gamma} > 0$ small enough and $k \in \mathbb{N}$, for all $\gamma \in (0, \bar{\gamma}]$ and under the following assumption **H**1 we control the distance between the target distribution π and $\mu_k^{(\gamma)}$.

H1. *For any* $\{(x^i, y^i, c^i)\}_{i=1}^b \in \mathbb{R}^{3d}$, *we have*

$$\sum_{i=1}^b \int_{\mathsf{E}} \mathscr{G}^i\left(\left\{(x^j, y^j, c^j)\right\}_{j=1}^b, \xi^i\right)\mathrm{d}\nu_\xi(\xi^i) = \sum_{i=1}^b \nabla U^i(x^i)\,.$$

Proposition 4. *Assume **A**1, **H**1 hold and let* $\gamma \leq 2(3L)^{-1}$. *Then, for any* $k \in \mathbb{N}$, *we have*

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] \leq [1 - \gamma m(1 - 3\gamma L)]\,\|\mathsf{X}_{k\gamma} - X_k\|^2 + \gamma\left(\frac{2L^2}{m} + 3\gamma L^2\right)V_k \\ + \left(\frac{2}{\gamma m}\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2 + 3\mathbb{E}^{\mathcal{F}_k}\left[\|I_k\|^2\right]\right) + \gamma^2\operatorname{Var}^{\mathcal{F}_k}\left(\frac{1}{b}\sum_{i=1}^b G_k^i\right),\end{aligned}$$

where $V_k, \mathcal{F}_k, d_k$ *are defined in* (24), (22) *and* (25).

Proof. Let k be in $\mathbb{N}$ and γ in $\left(0, 2(3L)^{-1}\right]$. Recall the stochastic processes $X_{k+1}, \mathsf{X}_{(k+1)\gamma}$ are defined in (23) and (26) by

$$\begin{cases}\mathsf{X}_{(k+1)\gamma} = \mathsf{X}_{k\gamma} - \gamma\nabla\bar{U}(\mathsf{X}_{k\gamma}) - I_k + \sqrt{2/b}\left(\mathsf{W}_{(k+1)\gamma} - \mathsf{W}_{k\gamma}\right)\,, \\ X_{k+1} = \frac{1}{b}\sum_{i=1}^b\left[X_k^i - \gamma G_k^i + \sqrt{2\gamma}\left(\sqrt{\tau/b}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i\right)\right]\,,\end{cases}$$

with I_k defined in (27). Substracting the two above equations gives

$$\mathsf{X}_{(k+1)\gamma} - X_{k+1} = (\mathsf{X}_{k\gamma} - X_k) - \left(\int_{k\gamma}^{(k+1)\gamma}\nabla\bar{U}(\mathsf{X}_s)\mathrm{d}s - \frac{\gamma}{b}\sum_{i=1}^b G_k^i\right)\,.$$

Taking the conditional expectation of the above equation and developing the squared norm, we obtain

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] &= \mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] - 2\gamma\left\langle \mathsf{X}_{k\gamma} - X_k, \nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\rangle \\ &- 2\left\langle \mathsf{X}_{k\gamma} - X_k, \mathbb{E}^{\mathcal{F}_k}\left[I_k\right] + \gamma\nabla\bar{U}(X_k) - \frac{\gamma}{b}\sum_{i=1}^{b}\mathbb{E}^{\mathcal{F}_k}\left[G_k^i\right]\right\rangle + \mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k + \gamma\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \frac{\gamma}{b}\sum_{i=1}^{b}G_k^i\right\|^2\right] . \end{aligned} \tag{28}$$

Using that for all $\alpha > 0, (a,b) \in (\mathbb{R}^d)^2$, $2\left\langle a, b\right\rangle \leq \alpha \left\|a\right\|^2 + (1/\alpha)\left\|b\right\|^2$ combined with **H**1, for any $\epsilon > 0$ we have

$$\begin{aligned} - 2\left\langle \mathsf{X}_{k\gamma} - X_k, \mathbb{E}^{\mathcal{F}_k}\left[I_k\right] + \gamma\nabla\bar{U}(X_k) - \frac{\gamma}{b}\sum_{i=1}^{b}\mathbb{E}^{\mathcal{F}_k}\left[G_k^i\right]\right\rangle &\leq \epsilon\|\mathsf{X}_{k\gamma} - X_k\|^2 + \frac{2}{\epsilon}\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2 \\ &+ \frac{2\gamma^2}{\epsilon}\left\|\nabla\bar{U}(X_k) - \frac{1}{b}\sum_{i=1}^{b}\nabla U^i(X_k^i)\right\|^2 . \end{aligned} \tag{29}$$

In addition, the unbiased property **H**1 implies that

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k + \gamma\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \frac{\gamma}{b}\sum_{i=1}^{b}G_k^i\right\|^2\right] &= \gamma^2\operatorname{Var}^{\mathcal{F}_k}\left(\frac{1}{b}\sum_{i=1}^{b}G_k^i\right) \\ &+ \mathbb{E}^{\mathcal{F}_k}\left[\left\|\gamma\left(\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right) + I_k + \gamma\nabla\bar{U}(X_k) - \frac{\gamma}{b}\sum_{i=1}^{b}\nabla U^i(X_k^i)\right\|^2\right] . \end{aligned} \tag{30}$$

The Young inequality shows that

$$\begin{aligned}&\mathbb{E}^{\mathcal{F}_k}\left[\left\|\gamma\left(\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right) + I_k + \gamma\nabla\bar{U}(X_k) - \frac{\gamma}{b}\sum_{i=1}^{b}\nabla U^i(X_k^i)\right\|^2\right] \\ &\leq 3\gamma^2\left\|\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\|^2 + 3\mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k\right\|^2\right] + 3\gamma^2\left\|\nabla\bar{U}(X_k) - \frac{1}{b}\sum_{i=1}^{b}\nabla U^i(X_k^i)\right\|^2 . \end{aligned}$$

By **A**1 we know that $\bar{U}$ is L-smooth and convex which imply the co-coercivity of $\bar{U}$ (Nesterov, 2003, Theorem 2.1.5), that is for all $x, y \in \mathbb{R}^d$, $\left\|\nabla\bar{U}(y) - \nabla\bar{U}(x)\right\|^2 \leq L\left\langle\nabla\bar{U}(y) - \nabla\bar{U}(x), y - x\right\rangle$. Hence, we deduce that

$$\left\|\nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\|^2 \leq L\left\langle \mathsf{X}_{k\gamma} - X_k, \nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\rangle . \tag{31}$$

Setting $\epsilon = \gamma m$, we have $0 < \epsilon \leq 1$ and $1 + 1/\epsilon \leq 2(\gamma m)^{-1}$. Therefore, (29), (30) and (31) associated with (28) show that

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] &\leq (1 + \gamma m)\left\|\mathsf{X}_{k\gamma} - X_k\right\|^2 + \left(\frac{2}{\gamma m}\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2 + 3\mathbb{E}^{\mathcal{F}_k}\left[\left\|I_k\right\|^2\right]\right) \\ &- \gamma\left(2 - 3\gamma L\right)\left\langle \mathsf{X}_{k\gamma} - X_k, \nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\rangle \\ &+ \gamma^2\left(3 + \frac{2}{\gamma m}\right)\left\|\nabla\bar{U}(X_k) - \frac{1}{b}\sum_{i=1}^{b}\nabla U^i(X_k^i)\right\|^2 + \gamma^2\operatorname{Var}^{\mathcal{F}_k}\left(\frac{1}{b}\sum_{i=1}^{b}G_k^i\right) . \end{aligned} \tag{32}$$

For any $i \in [b]$, by **A**1, the m-convexity of $\bar{U}$ gives that

$$\left\langle \mathsf{X}_{k\gamma} - X_k, \nabla\bar{U}(\mathsf{X}_{k\gamma}) - \nabla\bar{U}(X_k)\right\rangle \geq m\|\mathsf{X}_{k\gamma} - X_k\|^2 \tag{33}$$

In addition, under **A**1 the Jensen inequality implies

$$\left\|\nabla\bar{U}(X_k) - \frac{1}{b}\sum_{i=1}^{b}\nabla U^i(X_k^i)\right\|^2 \leq L^2 V_k , \tag{34}$$

where V_k is defined in (24). Therefore, using the assumption on γ and plugging (33) and (34) in (32) yields the expected inequality. □

6.1 General supporting lemmas

In this subsection, we consider the stochastic processes $(X_k)_{k\in\mathbb{N}}$, $(\mathsf{X}_{k\gamma})_{k\in\mathbb{N}}$ defined in (23) and (26). We derive several lemmas which allow us to derive a recursion on $\mathbb{E}[\|\mathsf{X}_{k\gamma} - X_k\|^2]$.

Lemma 5. *Assume **A**1 holds. Then, for any $k \in \mathbb{N}$ and $\gamma > 0$ we have*

$$\mathbb{E}\left[\|I_k\|^2\right] \le \frac{d\gamma^3 L^2}{b}\left(1 + \frac{\gamma L^2}{2m} + \frac{\gamma^2 L^2}{12}\right).$$

Proof. Let k be in $\mathbb{N}$. Using the Jensen inequality, we have

$$\begin{aligned}
\mathbb{E}\left[\|I_k\|^2\right] &= \mathbb{E}\left[\left\|\int_{k\gamma}^{(k+1)\gamma} \left(\nabla\bar{U}(\mathsf{X}_s) - \nabla\bar{U}(\mathsf{X}_{k\gamma})\right)\mathrm{d}s\right\|^2\right] \\
&\le \gamma\int_{k\gamma}^{(k+1)\gamma} \mathbb{E}\left[\left\|\nabla\bar{U}(\mathsf{X}_s) - \nabla\bar{U}(\mathsf{X}_{k\gamma})\right\|^2\right]\mathrm{d}s \\
&\le L^2\gamma\int_{k\gamma}^{(k+1)\gamma} \mathbb{E}\left[\|\mathsf{X}_s - \mathsf{X}_{k\gamma}\|^2\right]\mathrm{d}s\,. \qquad (35)
\end{aligned}$$

Further, for any $s \in \mathbb{R}_+$, using Durmus and Moulines (2019, Lemma 21) applied to $(\mathsf{X}_{bt})_{t\in\mathbb{R}_+}$ we obtain

$$\mathbb{E}^{\mathcal{F}_{k\gamma}}\left[\|\mathsf{X}_s - \mathsf{X}_{k\gamma}\|^2\right] \le \frac{d(s-k\gamma)}{b}\left(2 + (s-k\gamma)^2\frac{L^2}{3}\right) + \frac{3}{2}(s-k\gamma)^2 L^2\|\mathsf{X}_{k\gamma} - x_\star\|^2\,.$$

Integrating the previous inequality on $[k\gamma, (k+1)\gamma]$, it implies

$$\int_{k\gamma}^{(k+1)\gamma} \mathbb{E}\left[\|\mathsf{X}_s - \mathsf{X}_{k\gamma}\|^2\right]\mathrm{d}s \le \frac{\gamma^2}{b}\left(d + \frac{bL^2\gamma}{2}\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - x_\star\|^2\right] + \frac{dL^2\gamma^2}{12}\right). \qquad (36)$$

Plugging (36) in (35) gives

$$\mathbb{E}\left[\|I_k\|^2\right] \le \frac{L^2\gamma^3}{b}\left(d + \frac{bL^2\gamma}{2}\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - x_\star\|^2\right] + \frac{dL^2\gamma^2}{12}\right). \qquad (37)$$

Applying Durmus and Moulines (2019, Proposition 1) to $(\mathsf{X}_{bt})_{t\in\mathbb{R}_+}$, we get

$$\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - x_\star\|^2\right] \le \frac{d}{bm}\,. \qquad (38)$$

Thus, combining (37) with (38) completes the proof. □

Lemma 6. *Assume **A**1 and **HX**1 hold. Then, for any $k \in \mathbb{N}$ and $\gamma > 0$ we have*

$$\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] \le \frac{2\gamma^4 d}{3b}\left(L^3 + \frac{d\tilde{L}^2}{b}\right),$$

where I_k is defined in (27).

Proof. Denote Δ the Laplacian defined, for all $x \in \mathbb{R}^d$, by $\Delta U(x) = \{\sum_{l=1}^d (\partial^2 U_j)(x)/\partial x_l^2\}_{j=1}^d$, moreover let $k \in \mathbb{N}$ be a fixed integer and $\gamma > 0$. Using the Itô formula, we have for $s \in [k\gamma, (k+1)\gamma]$

$$\nabla\bar{U}(\mathsf{X}_s) - \nabla\bar{U}(\mathsf{X}_{k\gamma}) = \int_{k\gamma}^{s} \frac{1}{b}\Delta(\nabla\bar{U})(\mathsf{X}_u) - \nabla^2\bar{U}(\mathsf{X}_u)\nabla\bar{U}(\mathsf{X}_u)\mathrm{d}u + \sqrt{\frac{2}{b}}\int_{k\gamma}^{s} \nabla^2\bar{U}(\mathsf{X}_u)\mathrm{d}B_u\,. \qquad (39)$$

We will upper bound separately the three terms of the previous equality. First, the L-Lipschitz property of $\nabla\bar{U}$ given by **A**1 implies for any $u \in \mathbb{R}_+$ that

$$\left\|\nabla^2\bar{U}(\mathsf{X}_u)\nabla\bar{U}(\mathsf{X}_u)\right\| \le L\left\|\nabla\bar{U}(\mathsf{X}_u) - \nabla\bar{U}(x_\star)\right\|. \qquad (40)$$

In addition, since for $u \in \mathbb{R}_+$, the random variable X_u is distributed according to the stationary distribution $\pi \propto \exp(-U)$, we know from Dalalyan (2017, Lemma 2) that

$$\mathbb{E}\left[\left\|\nabla \bar{U}(\mathsf{X}_u) - \nabla \bar{U}(x_\star)\right\|^2\right] \leq \frac{dL}{b}\,. \tag{41}$$

Therefore, we deduce from (40) and (41) the following bound

$$\mathbb{E}\left[\left\|\nabla^2 \bar{U}(\mathsf{X}_u)\nabla \bar{U}(\mathsf{X}_u)\right\|^2\right] \leq \frac{dL^3}{b}\,. \tag{42}$$

Denote $(e_i)_{i=1}^d$ the canonical basis of $\mathbb{R}^d$; using that U is three times continuously differentiable we can apply the Schwarz's theorem which combined with **HX**1, immediately yield that

$$\begin{aligned}\left\|\Delta(\nabla\bar{U})(x)\right\|^2 &= \sum_{i=1}^d \left|\sum_{j=1}^d \partial_j^2 \partial_i \bar{U}(x)\right|^2 \leq d\sum_{i=1}^d\sum_{j=1}^d \left|\partial_i\partial_j^2\bar{U}(x)\right|^2 \\ &= d\sum_{i=1}^d \lim_{\epsilon\to 0}\left\{\epsilon^{-2}\sum_{j=1}^d\left|\partial_j^2\bar{U}(x+\epsilon\cdot e_i) - \partial_j^2\bar{U}(x)\right|^2\right\} \\ &\leq d\sum_{i=1}^d \lim_{\epsilon\to 0}\left\{\epsilon^{-2}\left(\tilde{L}\|(x+\epsilon\cdot e_i) - x\|^{-1}\right)^2\right\} \leq \left(d\tilde{L}\right)^2\,. \end{aligned}\tag{43}$$

Lastly, we upper bound the third term derived in (39). Since the potentials $\{U^i\}_{i\in[b]}$ are supposed L-smooth and $\bar{U}$ twice continuously differentiable, for $s \in [k\gamma, (k+1)\gamma]$ we know that $\int_{k\gamma}^s \nabla^2\bar{U}(\mathsf{X}_u)\mathrm{d}B_u$ is a $\mathcal{F}_s$-martingale. Thus, for $k \geq 0$ we deduce that

$$\mathbb{E}^{\mathcal{F}_k}\left[\int_{k\gamma}^{(k+1)\gamma}\nabla^2\bar{U}(\mathsf{X}_u)\,\mathrm{d}u\right] = 0\,. \tag{44}$$

Eventually, combining (39), (42), (43) and (44) with the Jensen and Young inequalities give

$$\begin{aligned}&\frac{1}{\gamma}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] = \frac{1}{\gamma}\mathbb{E}\left[\left\|\int_{k\gamma}^{(k+1)\gamma}\mathbb{E}^{\mathcal{F}_k}\left[\nabla\bar{U}(\mathsf{X}_s) - \nabla\bar{U}(\mathsf{X}_{k\gamma})\right]\mathrm{d}s\right\|^2\right] \\ &\leq \int_{k\gamma}^{(k+1)\gamma}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[\nabla\bar{U}(\mathsf{X}_s) - \nabla\bar{U}(\mathsf{X}_{k\gamma})\right]\right\|^2\right]\mathrm{d}s \\ &= \int_{k\gamma}^{(k+1)\gamma}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[\int_{k\gamma}^s\frac{1}{b}\Delta(\nabla\bar{U})(\mathsf{X}_u) - \nabla^2\bar{U}(\mathsf{X}_u)\nabla\bar{U}(\mathsf{X}_u)\mathrm{d}u\right]\right\|^2\right]\mathrm{d}s \\ &\leq 2\int_{k\gamma}^{(k+1)\gamma}(s-k\gamma)\int_{k\gamma}^s\mathbb{E}\left[\frac{1}{b^2}\left\|\int_{k\gamma}^s\Delta(\nabla\bar{U})(\mathsf{X}_u)\mathrm{d}u\right\|^2 + \left\|\nabla^2\bar{U}(\mathsf{X}_u)\nabla\bar{U}(\mathsf{X}_u)\mathrm{d}u\right\|^2\right]\mathrm{d}s \\ &\leq 2\int_{k\gamma}^{(k+1)\gamma}(s-k\gamma)^2\left(\frac{dL^3}{b} + \frac{(d\tilde{L})^2}{b^2}\right)\mathrm{d}s = \frac{2\gamma^3 d}{3b}\left(L^3 + \frac{d\tilde{L}^2}{b}\right)\,.\end{aligned}$$

Multiplying this last inequality by $\gamma > 0$ proves the expected result. □

Lemma 7. *Assume* **A**1 *hold. Then, for any $k \in \mathbb{N}$ and $\gamma \in \left(0, (3m)^{-1}\right]$ we have*

$$\frac{2}{\gamma m}\mathbb{E}\left[\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\|^2\right] + 3\mathbb{E}\left[\|I_k\|^2\right] \leq \begin{cases}\frac{3\gamma^2 dL^2}{bm}\left(1+\frac{19\gamma L^2}{36m}\right) & \\ \frac{\gamma^3 d}{bm}\left(5L^3 + \frac{4d\tilde{L}^2}{3b}\right) & \textit{if } \textbf{\textit{HX}}1 \textit{ holds and } \gamma \leq L^{-1}.\end{cases}$$

Proof. Let k be in $\mathbb{N}$ and $\gamma \in \left(0, (3m)^{-1}\right]$, using Lemma 5 we have

$$\mathbb{E}\left[\|I_k\|^2\right] \leq \frac{\gamma^3 dL^2}{b}\left(1 + \frac{\gamma L^2}{2m} + \frac{\gamma^2 L^2}{12}\right)\,.$$

Therefore, we deduce

$$\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right] \leq \frac{3\gamma^2 dL^2}{bm}\left(1 + \frac{\gamma L^2}{2m} + \frac{\gamma^2 L^2}{12}\right) .$$

Moreover, if we additionally suppose the regularity of the Hessian of the potentials $(U^i)_{i=1}^b$ as stated in **HX**1, we sharpen the upper bound on $\mathbb{E}[\|\mathbb{E}^{\mathcal{F}_k}[I_k]\|^2]$. Indeed, we show in Lemma 6 that

$$\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] \leq \frac{4\gamma^3 d}{3bm}\left(L^3 + \frac{d\tilde{L}^2}{b}\right) .$$

Hence, we deduce that

$$\begin{aligned}\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right] &\leq \frac{3\gamma^3 dL^2}{b}\left(1 + \frac{\gamma L^2}{2m} + \frac{\gamma^2 L^2}{12}\right) + \frac{4\gamma^3 d}{3bm}\left(L^3 + \frac{d\tilde{L}^2}{b}\right) \\ &\leq \frac{\gamma^3 dL^3}{bm}\left(3 + \frac{4}{3} + \frac{19\gamma L}{36}\right) + \frac{4\gamma^3 d^2\tilde{L}^2}{3b^2 m} .\end{aligned}$$

□

6.2 Derivation of the central theorem

H2. *There exist $\alpha_v \in (0,1)$ and $(v_1, v_2) \in (\mathbb{R}_+)^2$ such that for any $k \in \mathbb{N}$, V_k satisfies*

$$\mathbb{E}\left[V_k\right] \leq v_1\alpha_v^k + v_2 ,$$

where V_k is defined in (24).

HX2. *There exist $q_c \in (0,1)$ and $\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4 \in \mathbb{R}_+$ satisfying $(1-q_c)(1+\alpha_0+\sqrt{(\alpha_0-1)^2+4\alpha_1}) < 2$ such that for $k \geq 0$ the following inequality holds*

$$\begin{aligned}(1-q_c)^{-1}\mathbb{E}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] \leq \alpha_0\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] + \alpha_1\sum_{l=0}^{k-1}(1-q_c)^{k-l}\mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right] \\ + \alpha_2\mathbb{E}\left[V_k\right] + \alpha_3\sum_{l=0}^{k-1}(1-q_c)^{k-l}\mathbb{E}\left[V_l\right] + \alpha_4 .\end{aligned}$$

With the notation introduced in **HX**2, consider

$$\delta = \frac{-1-\alpha_0+\sqrt{(\alpha_0-1)^2+4\alpha_1}}{2} . \tag{45}$$

At iteration $k \geq 0$, recall that $\mu_k^{(\gamma)}$ denotes the distribution of the average parameter X_k (23). The next result controls the Wasserstein distance between $\mu_k^{(\gamma)}$ and the posterior distribution π.

Theorem 8. *Assume **HX**2 and **H**2 hold. Then, for any probability measure $\mu_0^{(\gamma)} \in \mathcal{P}_2(\mathbb{R}^d)$, $k \in \mathbb{N}$, we have*

$$\begin{aligned}\mathbf{W}_2^2\left(\mu_k^{(\gamma)}, \pi\right) \leq (1+\alpha_0+\delta)^k(1-q_c)^k\,\mathbf{W}_2^2\left(\mu_0^{(\gamma)}, \pi\right) + (1-q_c)v_1\left(\alpha_2 + \frac{\alpha_3}{\alpha_0+\delta}\right)\frac{\alpha_v^k - (1+\alpha_0+\delta)^k(1-q_c)^k}{\alpha_v - (1+\alpha_0+\delta)(1-q_c)} \\ + \frac{1-q_c}{q_c - (1-q_c)(\alpha_0+\delta)}\left[\left(\alpha_2 + \frac{\alpha_3}{\alpha_0+\delta}\right)v_2 + \alpha_4\right] .\end{aligned}$$

Proof. For any $n \in \mathbb{N}$, define

$$\begin{aligned} u_n &= (1-q_c)^{-n}\,\mathbb{E}\left[\|\mathsf{X}_{n\gamma} - X_n\|^2\right] , \qquad S_n = \sum_{l=0}^{n} u_l , \\ v_n &= (1-q_c)^{-n}\left(\alpha_2\mathbb{E}\left[V_n\right] + \alpha_4\right) + \alpha_3\sum_{l=0}^{n-1}(1-q_c)^{-l}\mathbb{E}\left[V_l\right] .\end{aligned} \tag{46}$$

With the above notations, **HX**2 becomes

$$u_{k+1} \le \alpha_0 u_k + \alpha_1 \sum_{l=0}^{k-1} u_l + v_k \,,$$

which can be rewritten as

$$S_{k+1} - S_k \le \alpha_0 \left(S_k - S_{k-1}\right) + \alpha_1 S_{k-1} + v_k \,. \tag{47}$$

Since δ is solution of $\delta(1+\alpha_0+\delta)+\alpha_0-\alpha_1=0$, adding $(1+\delta)S_k$ in (47) gives that

$$\begin{aligned} S_{k+1} + \delta S_k &\le (1+\alpha_0+\delta)\left(S_k - \frac{\alpha_0-\alpha_1}{1+\alpha_0+\delta} S_{k-1}\right) + v_k \\ &= (1+\alpha_0+\delta)\left(S_k + \delta S_{k-1}\right) + v_k \,. \end{aligned}$$

Using the fact that $\alpha_0 \le 1+\sqrt{(\alpha_0-1)^2+4\alpha_1}$, we obtain $2(1+\delta) = 1-\alpha_0+\sqrt{(\alpha_0-1)^2+4\alpha_1} \ge 0$. Hence $1+\delta>0$, which leads to the following upper bound

$$u_{k+1} \le u_{k+1} + (1+\delta)\sum_{l=0}^{k} u_l = S_{k+1} + \delta S_k \,.$$

Thus, we obtain that

$$u_k \le S_k + \delta S_{k-1} \le (1+\alpha_0+\delta)^{k-1}\left(u_1 + (1+\delta)u_k\right) + \sum_{l=1}^{k-1} (1+\alpha_0+\delta)^{k-l-1} v_l \,.$$

Plugging the definition (46) of u_k and v_l inside the previous inequality, we get

$$\begin{aligned} (1-q_{\mathrm{c}})^{-k}\, \mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] &\le (1+\alpha_0+\delta)^{k-1}\left((1-q_{\mathrm{c}})^{-1}\, \mathbb{E}\left[\|\mathsf{X}_{\gamma} - X_1\|^2\right] + (1+\delta)\mathbb{E}\left[\|\mathsf{X}_0 - X_0\|^2\right]\right) \\ &+ \sum_{l=1}^{k-1} (1+\alpha_0+\delta)^{k-l-1}\left[(1-q_{\mathrm{c}})^{-l}\left(\alpha_2 \mathbb{E}\left[V_l\right] + \alpha_4\right) + \alpha_3 \sum_{j=0}^{l-1} (1-q_{\mathrm{c}})^{-j}\mathbb{E}\left[V_j\right]\right] \,. \end{aligned} \tag{48}$$

Moreover, using **HX**2 we obtain that

$$\mathbb{E}\left[\|\mathsf{X}_{\gamma} - X_1\|^2\right] \le (1-q_{\mathrm{c}})\alpha_0 \mathbb{E}\left[\|\mathsf{X}_0 - X_0\|^2\right] + (1-q_{\mathrm{c}})\alpha_2 \mathbb{E}\left[V_0\right] + \alpha_4 \,, \tag{49}$$

combining (48) with (49) yield

$$\begin{aligned} \mathbb{E}\left[\|\mathsf{X}_{k\gamma} - X_k\|^2\right] &\le (1+\alpha_0+\delta)^{k}(1-q_{\mathrm{c}})^{k}\, \mathbb{E}\left[\|\mathsf{X}_0 - X_0\|^2\right] + \alpha_2 \sum_{l=0}^{k-1} (1+\alpha_0+\delta)^{k-l-1}(1-q_{\mathrm{c}})^{k-l}\, \mathbb{E}\left[V_l\right] \\ &+ \alpha_3 \sum_{j=0}^{k-2} (1-q_{\mathrm{c}})^{k-j}\mathbb{E}\left[V_j\right] \sum_{l=j+1}^{k-1} (1+\alpha_0+\delta)^{k-l-1} + (1-q_{\mathrm{c}})\alpha_4 \sum_{l=0}^{k-1} (1+\alpha_0+\delta)^{l}(1-q_{\mathrm{c}})^{l} \,. \end{aligned} \tag{50}$$

Consider the function $f : a \in \mathbb{R} \to \mathbb{R}$ defined by $f(a) = a(1+\alpha_0+a)+\alpha_0-\alpha_1$. Using the definition (45) of δ combined with the increasing property of f, we deduce from $f(\delta) = 0 > f(-\alpha_0) = -\alpha_1$ that $\delta > -\alpha_0$, and thus we get $1+\alpha_0+\delta > 1$ which implies that

$$\begin{aligned} \sum_{l=j+1}^{k-1} (1+\alpha_0+\delta)^{k-l-1} &\le \sum_{l=0}^{k-j-2} (1+\alpha_0+\delta)^{k-j-l-2} \\ &\le \frac{(1+\alpha_0+\delta)^{k-j-1}}{\alpha_0+\delta} \,. \end{aligned} \tag{51}$$

Therefore, plugging (51) in (50) gives

$$\sum_{j=0}^{k-2} (1-q_{\mathrm{c}})^{k-j}\mathbb{E}\left[V_j\right] \sum_{l=j+1}^{k-1} (1+\alpha_0+\delta)^{k-l-1} \le \sum_{l=0}^{k-2} \frac{(1-q_{\mathrm{c}})^{k-l}(1+\alpha_0+\delta)^{k-l-1}}{\alpha_0+\delta}\mathbb{E}\left[V_l\right] \,. \tag{52}$$

In addition, since **HX**2 ensures that $(1-q_{\mathrm{c}})(1+\alpha_0+\delta)<1$, we have

$$\sum_{l=0}^{k-1}(1+\alpha_0+\delta)^l\,(1-q_{\mathrm{c}})^l \leq \frac{1}{q_{\mathrm{c}}-(1-q_{\mathrm{c}})(\alpha_0+\delta)}\,. \tag{53}$$

The last inequality combined with (50) and (52) show that

$$\begin{aligned}\mathbb{E}\left[\|\mathsf{X}_{k\gamma}-X_k\|^2\right] &\leq (1+\alpha_0+\delta)^k\,(1-q_{\mathrm{c}})^k\,\mathbb{E}\left[\|\mathsf{X}_0-X_0\|^2\right]\\ &\quad+\left(\alpha_2+\frac{\alpha_3}{\alpha_0+\delta}\right)\sum_{l=0}^{k-1}(1+\alpha_0+\delta)^{k-l-1}\,(1-q_{\mathrm{c}})^{k-l}\,\mathbb{E}\left[V_l\right]+\frac{(1-q_{\mathrm{c}})\alpha_4}{q_{\mathrm{c}}-(1-q_{\mathrm{c}})(\alpha_0+\delta)}\,.\end{aligned} \tag{54}$$

Further, since we assume **H**2, we have

$$\begin{aligned}\sum_{l=0}^{k-1}(1+\alpha_0+\delta)^{k-l-1}\,(1-q_{\mathrm{c}})^{k-l}\,\mathbb{E}\left[V_l\right] &\leq v_1\sum_{l=0}^{k-1}(1+\alpha_0+\delta)^{k-l-1}\,(1-q_{\mathrm{c}})^{k-l}\,\alpha_v^l\\ &\quad+v_2\sum_{l=0}^{k-1}(1+\alpha_0+\delta)^{k-l-1}\,(1-q_{\mathrm{c}})^{k-l}\,.\end{aligned} \tag{55}$$

A calculation gives that

$$\sum_{l=0}^{k-1}(1+\alpha_0+\delta)^{k-l-1}\,(1-q_{\mathrm{c}})^{k-l}\,\alpha_v^l \leq (1-q_{\mathrm{c}})\frac{\alpha_v^k-(1+\alpha_0+\delta)^k\,(1-q_{\mathrm{c}})^k}{\alpha_v-(1+\alpha_0+\delta)\,(1-q_{\mathrm{c}})} \tag{56}$$

and combining (53), (55) with (56), we find that

$$\sum_{l=0}^{k-1}(1+\alpha_0+\delta)^{k-l-1}\,(1-q_{\mathrm{c}})^{k-l}\,\mathbb{E}\left[V_l\right] \leq (1-q_{\mathrm{c}})v_1\frac{\alpha_v^k-(1+\alpha_0+\delta)^k\,(1-q_{\mathrm{c}})^k}{\alpha_v-(1+\alpha_0+\delta)\,(1-q_{\mathrm{c}})}+\frac{(1-q_{\mathrm{c}})v_2}{q_{\mathrm{c}}-(1-q_{\mathrm{c}})(\alpha_0+\delta)}\,. \tag{57}$$

Therefore, plugging (57) inside (54) shows that

$$\begin{aligned}\mathbb{E}\left[\|\mathsf{X}_{k\gamma}-X_k\|^2\right] &\leq (1+\alpha_0+\delta)^k\,(1-q_{\mathrm{c}})^k\,\mathbb{E}\left[\|\mathsf{X}_0-X_0\|^2\right]\\ &\quad+(1-q_{\mathrm{c}})v_1\left(\alpha_2+\frac{\alpha_3}{\alpha_0+\delta}\right)\frac{\alpha_v^k-(1+\alpha_0+\delta)^k\,(1-q_{\mathrm{c}})^k}{\alpha_v-(1+\alpha_0+\delta)\,(1-q_{\mathrm{c}})}\\ &\quad+\frac{1-q_{\mathrm{c}}}{q_{\mathrm{c}}-(1-q_{\mathrm{c}})(\alpha_0+\delta)}\left[\left(\alpha_2+\frac{\alpha_3}{\alpha_0+\delta}\right)v_2+\alpha_4\right]\,.\end{aligned} \tag{58}$$

Eventually, since the Wasserstein distance $\mathbf{W}_2(\pi,\mu_k^{(\gamma)})$ is the infimum over all couplings, we obtain that $\mathbf{W}_2^2(\pi,\mu_k^{(\gamma)}) \leq \mathbb{E}[\|\mathsf{X}_{k\gamma}-X_k\|^2]$. Moreover, it follows from the strongly convex assumption **A**1 that $\pi\in\mathcal{P}_2(\mathbb{R}^d)$. Thus, we can apply Villani (2009, Theorem 4.1) to prove the existence of an optimal coupling ζ such that taking (X_0,X_0) distributed according to ζ implies that $\mathbb{E}[\|\mathsf{X}_0-X_0\|^2]^{1/2}=\mathbf{W}_2(\pi,\mu_0^{(\gamma)})$. Substituting these results into (58) completes the proof. □

6.3 Upper bound on V_k

The goal of this subsection is to prove the upper bound derived in Lemma 11 for $(\mathbb{E}\left[V_k\right])_{k\in\mathbb{N}}$ to ensure that **H**2 holds. Recall that for $k\geq 0$, V_k is defined in (24), d_k in (25), G_k^i in (17) and we introduce $\bar{G}_k^i=\mathbb{E}^{\mathcal{F}_k}[G_k^i]$. To prove the central lemma of this subsection, we also consider the assumptions **HX**3 and **HX**4 given below.

HX3. *There exist $A_d, A_\sigma\in(0,1)\,, B_d, B_\sigma, C_d, C_\sigma, D_d, D_\sigma\in\mathbb{R}_+$, such that for any $k\in\mathbb{N}$, we have*

$$\begin{aligned}\mathbb{E}\left[d_{k+1}^2\right] &\leq (1-A_d)\,\mathbb{E}\left[d_k^2\right]+B_d\mathbb{E}\left[\sigma_k^2\right]+C_d\mathbb{E}\left[V_k\right]+D_d\,,\\ \mathbb{E}\left[\sigma_{k+1}^2\right] &\leq (1-A_\sigma)\,\mathbb{E}\left[\sigma_k^2\right]+B_\sigma\mathbb{E}\left[d_k^2\right]+C_\sigma\mathbb{E}\left[V_k\right]+D_\sigma\,.\end{aligned}$$

HX4. *There exist $A, \bar{A}, B, \bar{B}, C, \bar{C}, D, \bar{D} \geq 0$ such that for any $i \in [b], k \in \mathbb{N}$, we have*

$$\frac{1}{b}\sum_{i=1}^{b} \mathbb{E}\left[\left\|\bar{G}_k^i\right\|^2\right] \leq \bar{A}\mathbb{E}\left[V_k\right] + \bar{B}\mathbb{E}\left[d_k^2\right] + \bar{C}\mathbb{E}\left[\sigma_k^2\right] + \bar{D}\,,$$
$$\frac{1}{b}\sum_{i=1}^{b} \mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right] \leq A\mathbb{E}\left[V_k\right] + B\mathbb{E}\left[d_k^2\right] + C\mathbb{E}\left[\sigma_k^2\right] + D\,.$$

With the notation considered in **HX**3 and **HX**4, for any $\gamma > 0$ we also introduce the following quantities:

$$\begin{aligned}
\mathrm{C}^\gamma &= \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}} - 4A_d}\left[B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B} + \frac{B_\sigma}{A_\sigma - A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right]\,,\\
\mathrm{C}_r^\gamma &= \frac{9\gamma^2\left(1-p_{\mathrm{c}}\right)C_\sigma}{p_{\mathrm{c}} - 4A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) + 3\mathrm{C}^\gamma\left(C_d + \frac{B_d C_\sigma}{A_\sigma - A_d}\right)\,,\\
\mathrm{C}_\sigma^\gamma &= \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}} - 4A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) + \mathrm{C}^\gamma B_d\left(2 + \frac{3}{A_\sigma - A_d}\right)\,, \quad \mathrm{C}_d^\gamma = 7\mathrm{C}^\gamma\,, \quad \mathrm{C}_V^\gamma = 1 + 2\mathrm{C}^\gamma C_d\,,\\
\mathrm{C}_\delta^\gamma &= \frac{4(1-p_{\mathrm{c}})\gamma^2 D_\sigma}{A_\sigma(p_{\mathrm{c}} - 4A_d)}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) + \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right)\\
&\quad + \frac{\mathrm{C}^\gamma}{A_d}\left(1 + \frac{2B_d B_\sigma}{A_d(A_\sigma - A_d)}\right)\left(D_d + \frac{B_d D_\sigma}{A_\sigma}\right) + \frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{b p_{\mathrm{c}}}\,.
\end{aligned} \tag{59}$$

If $A_d \leq A_\sigma/2$ and $A_d A_\sigma \geq 8B_d B_\sigma$, we also introduce a convergence rate (proved later in Lemma 10) defined by

$$\alpha = A_d - \frac{2(A_\sigma - A_d)^{-1}B_d B_\sigma}{1 + \sqrt{1 + 4(1-A_d)^{-1}(A_\sigma - A_d)^{-1}B_d B_\sigma}}\,. \tag{60}$$

Lemma 9. *Assume **HX**3 and also that $A_d \leq A_\sigma/2$, $A_d A_\sigma \geq 8B_d B_\sigma$ hold. Then, we have*

$$A_d/2 < \alpha \leq A_d\,.$$

Proof. First, introduce $\delta_\alpha \in \mathbb{R}_+$ the unique non-negative solution of

$$\delta_\alpha^2 + \delta_\alpha = \frac{B_d B_\sigma}{(1-A_d)(A_\sigma - A_d)}\,.$$

Since we suppose $A_d \leq A_\sigma/2$, thus we have $A_d \leq 1/2$ which implies that $(1-A_d)(A_d^2/4 + A_d/2) \geq A_d/4$. In addition, using $A_d A_\sigma \geq 8B_d B_\sigma$, we get that

$$(1-A_d)\left(\frac{A_d^2}{4} + \frac{A_d}{2}\right) \geq \frac{A_d}{4} \geq \frac{2B_d B_\sigma}{A_\sigma} \geq (1-A_d)\left(\delta_\alpha^2 + \delta_\alpha\right)\,.$$

Hence, the increasing property of the function $x \in \mathbb{R}_+ \mapsto x^2 + x$ combined with the fact that $\delta_\alpha \geq 0$ prove that $A_d \geq 2\delta_\alpha$. Moreover, a calculation shows that α satisfies $\alpha = 1 - (1-A_d)(1+\delta_\alpha)$. Thus, using $0 \leq 2\delta_\alpha \leq A_d$ implies that $\alpha \in (A_d/2, A_d]$. □

The random variable V_k given in (24) measures the averaged distance between the global parameter X_k and the local ones $(X_k^i)_{i\in[b]}$. The first lines of the proof of the next lemma are based on Gorbunov et al. (2021, Lemma E.3), however their purpose was to upper bound $\sum_l w_l \mathbb{E}V_l$ for some weights $w_l > 0$, while we prefer to control $\mathbb{E}V_k$ to combine this bound with that of Proposition 4. Moreover, the assumptions considered in this work are different, so the proof requires the development of other techniques

Lemma 10. *Assume **HX**3, **HX**4 hold with $A_d < \min(A_\sigma/2, p_{\mathrm{c}}/4), A_d A_\sigma \geq 8B_d B_\sigma$ and consider $\gamma \leq p_{\mathrm{c}}^{1/2}(2-2p_{\mathrm{c}})^{-1/2}[A + (1+2/p_{\mathrm{c}})\bar{A}]^{-1/2}$. Then, for any $k \in \mathbb{N}$, we have*

$$\mathbb{E}\left[V_k\right] \leq (1-\alpha)^k\left(\mathrm{C}_V^\gamma\mathbb{E}\left[V_0\right] + \mathrm{C}_d^\gamma\mathbb{E}\left[d_0^2\right] + \mathrm{C}_\sigma^\gamma\mathbb{E}\left[\sigma_0^2\right] + 2D_d\right) + \mathrm{C}_r^\gamma\sum_{i=0}^{k-2}(1-\alpha)^{k-i-1}\mathbb{E}\left[V_i\right] + \mathrm{C}_\delta^\gamma\,,$$

where V_k is defined in (24).

Proof. Let $k \in \mathbb{N}^\star$, using for $i \in [b]$ the definitions (20), (23) of X_k^i and X_k

$$X_{k+1}^i = X_k^i - \gamma G_k^i + \sqrt{2\gamma}\left(\sqrt{\tau/b}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i\right),$$
$$X_{k+1} = X_k - \frac{\gamma}{b}\sum_{j=1}^{b} G_k^i + \sqrt{\frac{2\gamma\tau}{b}}\tilde{Z}_{k+1} + \frac{\sqrt{2(1-\tau)\gamma}}{b}\sum_{i=1}^{b} Z_{k+1}^i\,.$$

First upper bound on $\mathbb{E}\left[V_k\right]$. Substracting the two above equations combined with the Jensen inequality give

$$\begin{aligned}
&\mathbb{E}\left[V_{k+1}\right] = \frac{1}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|X_{k+1}^i - X_{k+1}\right\|^2\right]\\
&= \frac{1-p_{\mathrm{c}}}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(G_k^i - G^k) + \sqrt{2(1-\tau)\gamma}Z_{k+1}^i - \frac{\sqrt{2(1-\tau)\gamma}}{b}\sum_{j=1}^{b} Z_{k+1}^j\right\|^2\right]\\
&= \frac{1-p_{\mathrm{c}}}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(\bar{G}_k^i - \bar{G}^k)\right\|^2\right] + \frac{(1-p_{\mathrm{c}})\gamma^2}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|(G_k^i - \bar{G}_k^i) - (G^k - \bar{G}^k)\right\|^2\right]\\
&\quad + 2(1-\tau)\gamma\mathbb{E}\left[\left\|Z_{k+1}^i - \frac{1}{b}\sum_{j=1}^{b} Z_{k+1}^j\right\|^2\right]
\end{aligned}$$

Hence, we get

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] &\leq \frac{1-p_{\mathrm{c}}}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(\bar{G}_k^i - \bar{G}^k)\right\|^2\right] + \frac{(1-p_{\mathrm{c}})\gamma^2}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right]\\
&\quad + 2\left(1-\tau\right)\left(1-1/b\right)\gamma d\\
&\leq \frac{(1-p_{\mathrm{c}})(1+p_{\mathrm{c}}/2)}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|X_k^i - X_k\right\|^2\right] + \frac{(1-p_{\mathrm{c}})\gamma^2}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right]\\
&\quad + \frac{(1-p_{\mathrm{c}})(1+2/p_{\mathrm{c}})\gamma^2}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|\bar{G}_k^i - \bar{G}^k\right\|^2\right] + 2\left(1-\tau\right)\left(1-1/b\right)\gamma d\,.
\end{aligned}$$

Using $(1-p_{\mathrm{c}})(1+p_{\mathrm{c}}/2) \leq 1 - p_{\mathrm{c}}/2$, we finally obtain

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] &\leq \left(1-p_{\mathrm{c}}/2\right)\mathbb{E}\left[V_k\right] + \frac{(1-p_{\mathrm{c}})(2+p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|\bar{G}_k^i\right\|^2\right]\\
&\quad + \frac{(1-p_{\mathrm{c}})\gamma^2}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right] + 2\left(1-\tau\right)\left(1-1/b\right)\gamma d\,.
\end{aligned}$$

Combining the last inequality with **HX4**, it shows

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] &\leq \left(1 - \frac{p_{\mathrm{c}}}{2} + (1-p_{\mathrm{c}})\gamma^2\left[A + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{A}\right]\right)\mathbb{E}\left[V_k\right] + (1-p_{\mathrm{c}})\gamma^2\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right)\\
&\quad + (1-p_{\mathrm{c}})\gamma^2\left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\mathbb{E}\left[d_k^2\right] + (1-p_{\mathrm{c}})\gamma^2\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_k^2\right] + 2\left(1-\tau\right)\left(1-1/b\right)\gamma d\,.
\end{aligned}$$

Since $\gamma \leq \frac{p_{\mathrm{c}}^{1/2}}{2(1-p_{\mathrm{c}})^{1/2}\left[A+(1+2/p_{\mathrm{c}})\bar{A}\right]^{1/2}}$, the above inequality implies that

$$\mathbb{E}\left[V_{k+1}\right] \leq \left(1 - \frac{p_{\mathrm{c}}}{4}\right)\mathbb{E}\left[V_k\right] + (1-p_{\mathrm{c}})\gamma^2\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + 2\left(1-\tau\right)\left(1-1/b\right)\gamma d$$

$$+ (1-p_{\mathrm{c}})\gamma^2 \left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right) \mathbb{E}\left[d_k^2\right] + (1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \mathbb{E}\left[\sigma_k^2\right] .$$

Using by convention that $\sum_{l=0}^{-1} = 0$, an induction shows that

$$\begin{aligned}\mathbb{E}\left[V_k\right] \leq \left(1-\frac{p_{\mathrm{c}}}{4}\right)^k \mathbb{E}\left[V_0\right] + \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + \frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{bp_{\mathrm{c}}} \\ + (1-p_{\mathrm{c}})\gamma^2 \left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right) \sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} \mathbb{E}\left[d_l^2\right] \\ + (1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} \mathbb{E}\left[\sigma_l^2\right] . \qquad (61)\end{aligned}$$

Moreover, for any $l \in \mathbb{N}^\star$ the assumption **HX**3 implies that

$$\mathbb{E}\left[d_l^2\right] \leq (1-A_d)\,\mathbb{E}\left[d_{l-1}^2\right] + B_d\mathbb{E}\left[\sigma_{l-1}^2\right] + C_d\mathbb{E}\left[V_{l-1}\right] + D_d\,,$$

and unrolling the recursion gives that

$$\mathbb{E}\left[d_l^2\right] \leq (1-A_d)^l\,\mathbb{E}\left[d_0^2\right] + \sum_{j=1}^{l}(1-A_d)^{l-j}\left(B_d\mathbb{E}\left[\sigma_{j-1}^2\right] + C_d\mathbb{E}\left[V_{j-1}\right]\right) + \frac{D_d}{A_d} . \qquad (62)$$

Similarly, we also have

$$\mathbb{E}\left[\sigma_l^2\right] \leq (1-A_\sigma)^l\,\mathbb{E}\left[\sigma_0^2\right] + \sum_{j=1}^{l}(1-A_\sigma)^{l-j}\left(B_\sigma\mathbb{E}\left[d_{j-1}^2\right] + C_\sigma\mathbb{E}\left[V_{j-1}\right]\right) + \frac{D_\sigma}{A_\sigma} . \qquad (63)$$

Hence, by plugging (63) in (61) we obtain that

$$\begin{aligned}\mathbb{E}\left[V_k\right] \leq \left(1-\frac{p_{\mathrm{c}}}{4}\right)^k \mathbb{E}\left[V_0\right] + \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + \frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{bp_{\mathrm{c}}} \\ + (1-p_{\mathrm{c}})\gamma^2 \left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right) \sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} \mathbb{E}\left[d_l^2\right] \\ + (1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} (1-A_\sigma)^l\,\mathbb{E}\left[\sigma_0^2\right] \\ + B_\sigma(1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \sum_{l=0}^{k-1}\sum_{j=1}^{l}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} (1-A_\sigma)^{l-j}\,\mathbb{E}\left[d_{j-1}^2\right] \\ + C_\sigma(1-p_{\mathrm{c}})\gamma^2 \left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \sum_{l=0}^{k-1}\sum_{j=1}^{l}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1} (1-A_\sigma)^{l-j}\,\mathbb{E}\left[V_{j-1}\right] \\ + \frac{4(1-p_{\mathrm{c}})\gamma^2 D_\sigma}{A_\sigma(p_{\mathrm{c}}-4A_d)}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) . \qquad (64)\end{aligned}$$

In addition, interchanging the summations gives

$$\sum_{l=0}^{k-1}\sum_{j=1}^{l}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}(1-A_\sigma)^{l-j}\,\mathbb{E}\left[V_{j-1}^2\right] = \sum_{i=0}^{k-2}\left[\sum_{l=0}^{k-i-2}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-i-2-l}(1-A_\sigma)^l\right]\mathbb{E}\left[V_i\right] .$$

Thus, using that $\sum_{l=0}^{k-i-2}\left(1-p_{\mathrm{c}}/4\right)^{k-i-2-l}\left(1-A_\sigma\right)^l \leq 4\left(1-A_d\right)^{k-i-1}\left(p_{\mathrm{c}}-4A_d\right)^{-1}$, we can simplify the upper bound of $\mathbb{E}\left[V_k\right]$ derived in (64). Indeed, we can write

$$\begin{aligned}\mathbb{E}\left[V_k\right] \leq \left(1-\frac{p_{\mathrm{c}}}{4}\right)^k \mathbb{E}\left[V_0\right] &+ \frac{4(1-p_{\mathrm{c}})\gamma^2\left(1-A_d\right)^k}{p_{\mathrm{c}}-4A_d}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_0^2\right] \\ &+ \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + \frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{bp_{\mathrm{c}}} + \frac{4(1-p_{\mathrm{c}})\gamma^2 D_\sigma}{A_\sigma(p_{\mathrm{c}}-4A_d)}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) \\ &+ (1-p_{\mathrm{c}})\gamma^2\left(B+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}\mathbb{E}\left[d_l^2\right] \\ &+ B_\sigma(1-p_{\mathrm{c}})\gamma^2\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}\sum_{j=0}^{l-1}\left(1-A_\sigma\right)^{l-j-1}\mathbb{E}\left[d_j^2\right] \\ &+ \frac{4(1-p_{\mathrm{c}})\gamma^2 C_\sigma}{p_{\mathrm{c}}-4A_d}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\sum_{l=0}^{k-2}\left(1-A_d\right)^{k-l-1}\mathbb{E}\left[V_l\right]. \qquad (65)\end{aligned}$$

Upper bound on $\mathbb{E}\left[d_k^2\right]$. For $l \geq 1$, plugging (63) into (62) yields the following upper bound

$$\begin{aligned}\mathbb{E}\left[d_l^2\right] \leq \left(1-A_d\right)^l \mathbb{E}\left[d_0^2\right] &+ C_d\sum_{j=1}^{l}\left(1-A_d\right)^{l-j}\mathbb{E}\left[V_{j-1}\right] + \frac{D_d}{A_d} \\ &+ B_d\sum_{j=1}^{l}\left(1-A_d\right)^{l-j}\left[\left(1-A_\sigma\right)^{j-1}\mathbb{E}\left[\sigma_0^2\right] + \sum_{i=1}^{j-1}\left(1-A_\sigma\right)^{j-i-1}\left(B_\sigma\mathbb{E}\left[d_{i-1}^2\right] + C_\sigma\mathbb{E}\left[V_{i-1}\right]\right) + \frac{D_\sigma}{A_\sigma}\right].\end{aligned}$$

The above inequality leads to the next inequality

$$\begin{aligned}\mathbb{E}\left[d_l^2\right] \leq \left(1-A_d\right)^l \mathbb{E}\left[d_0^2\right] &+ B_d\sum_{j=1}^{l}\left(1-A_d\right)^{l-j}\left(1-A_\sigma\right)^{j-1}\mathbb{E}\left[\sigma_0^2\right] \\ &+ C_d\sum_{j=1}^{l}\left(1-A_d\right)^{l-j}\mathbb{E}\left[V_{j-1}\right] + B_dC_\sigma\sum_{j=1}^{l}\sum_{i=1}^{j-1}\left(1-A_\sigma\right)^{j-i-1}\left(1-A_d\right)^{l-j}\mathbb{E}\left[V_{i-1}\right] \\ &+ B_dB_\sigma\sum_{j=1}^{l}\sum_{i=1}^{j-1}\left(1-A_d\right)^{l-j}\left(1-A_\sigma\right)^{j-i-1}\mathbb{E}\left[d_{i-1}^2\right] + \frac{D_d}{A_d} + \frac{B_dD_\sigma}{A_dA_\sigma}. \qquad (66)\end{aligned}$$

By interchanging the double summations in (66), we obtain

$$\begin{aligned}&\sum_{j=1}^{l}\sum_{i=1}^{j-1}\left(1-A_d\right)^{l-j}\left(1-A_\sigma\right)^{j-i-1}\mathbb{E}\left[d_{i-1}^2\right] = \sum_{i=1}^{l-1}\left[\sum_{j=i+1}^{l}\left(1-A_d\right)^{l-j}\left(1-A_\sigma\right)^{j-i-1}\right]\mathbb{E}\left[d_{i-1}^2\right] \\ &= \sum_{i=0}^{l-2}\left[\sum_{j=0}^{l-i-2}\left(1-A_d\right)^{l-i-2-j}\left(1-A_\sigma\right)^{j}\right]\mathbb{E}\left[d_i^2\right] \leq \frac{1}{A_\sigma-A_d}\sum_{i=0}^{l-2}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[d_i^2\right]. \qquad (67)\end{aligned}$$

Similarly, we can also get that

$$\sum_{j=1}^{l}\sum_{i=1}^{j-1}\left(1-A_d\right)^{l-j}\left(1-A_\sigma\right)^{j-i-1}\mathbb{E}\left[V_{i-1}\right] \leq \frac{1}{A_\sigma-A_d}\sum_{i=0}^{l-2}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[V_i\right]. \qquad (68)$$

Plugging back (67) and (68) in (66) shows

$$\begin{aligned}\mathbb{E}\left[d_l^2\right] \leq \left(1-A_d\right)^l \mathbb{E}\left[d_0^2\right] &+ \frac{B_d\left(1-A_d\right)^l}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right] + \frac{B_dB_\sigma}{A_\sigma-A_d}\sum_{i=0}^{l-2}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[d_i^2\right] \\ &+ C_d\sum_{i=0}^{l-1}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[V_i\right] + \frac{B_dC_\sigma}{A_\sigma-A_d}\sum_{i=0}^{l-2}\left(1-A_d\right)^{l-i-1}\mathbb{E}\left[V_i\right] + \frac{D_d}{A_d} + \frac{B_dD_\sigma}{A_dA_\sigma}. \qquad (69)\end{aligned}$$

Now, we want to control $\sum_{i=0}^{l-2}(1-A_d)^{l-i-1}\,\mathbb{E}\left[d_i^2\right]$. For this, for any $l \in \mathbb{N}$ define

$$U_l = \mathbb{E}\left[d_0^2\right] + \frac{B_d}{A_\sigma - A_d}\mathbb{E}\left[\sigma_0^2\right] + \frac{D_d\,(1-A_d)^{-l}}{A_d} + \frac{B_d D_\sigma\,(1-A_d)^{-l}}{A_d A_\sigma} \\ + C_d\sum_{i=0}^{l-1}(1-A_d)^{-i-1}\,\mathbb{E}\left[V_i\right] + \frac{B_d C_\sigma}{A_\sigma - A_d}\sum_{i=0}^{l-2}(1-A_d)^{-i-1}\,\mathbb{E}\left[V_i\right] \quad (70)$$

and consider

$$S_l = \sum_{i=0}^{l}(1-A_d)^{-i}\,\mathbb{E}\left[d_i^2\right]\ .$$

With the above notation, (69) can be rewritten as

$$S_l - S_{l-1} \le \frac{B_d B_\sigma}{(1-A_d)\,(A_\sigma - A_d)}S_{l-2} + U_l\,. \quad (71)$$

For $l \ge 2$, using the upper bound derived in (71) gives

$$\mathbb{E}\left[d_l^2\right] = (1-A_d)^l\,(S_l - S_{l-1}) \le \frac{B_d B_\sigma\,(1-A_d)^{l-1}\,S_{l-2}}{(A_\sigma - A_d)} + (1-A_d)^l\,U_l\,. \quad (72)$$

Finally, we define

$$\delta_\alpha = \frac{-1+\sqrt{1+4(1-A_d)^{-1}\,(A_\sigma - A_d)^{-1}B_d B_\sigma}}{2}$$

such that δ_α is solution of the equation

$$\delta_\alpha^2 + \delta_\alpha = \frac{B_d B_\sigma}{(1-A_d)\,(A_\sigma - A_d)} \quad (73)$$

Thus for $l \ge 2$, the definition of δ_α combined with (71) show

$$S_l + \delta_\alpha S_{l-1} \le (1+\delta_\alpha)\,(S_{l-1} + \delta_\alpha S_{l-2}) + U_l\,.$$

Unrolling this recursion gives

$$S_k + \delta_\alpha S_{k-1} \le (1+\delta_\alpha)^{k-1}\,(S_1 + \delta_\alpha S_0) + \sum_{l=2}^{k}(1+\delta_\alpha)^{k-l}\,U_l\,. \quad (74)$$

Upper bound on $\sum_{l=0}^{k-1}(1-\tilde{\alpha})^{l-j-1}\,\mathbb{E}[d_j^2]$. Let consider a fixed $\tilde{\alpha} \in \{p_c/4, A_\sigma\}$, by assumption we have $A_d < \tilde{\alpha} < 1$. Since we want to control $\sum_{l=0}^{k-1}(1-p_c/4)^{k-l-1}\mathbb{E}[d_l^2]$ and $\sum_{l=0}^{k-1}(1-p_c/4)^{k-l-1}\sum_{j=0}^{l-1}(1-A_\sigma)^{l-j-1}\,\mathbb{E}[d_j^2]$ involved in the inequality (65), we first study $\sum_{l=0}^{k-1}(1-\tilde{\alpha})^{k-l-1}\mathbb{E}[d_l^2]$. From (72), we deduce that

$$\sum_{l=0}^{k-1}(1-\tilde{\alpha})^{k-l-1}\,\mathbb{E}\left[d_l^2\right] \le \frac{B_d B_\sigma}{(1-A_d)(A_\sigma - A_d)}\sum_{l=0}^{k-1}(1-A_d)^l\,(1-\tilde{\alpha})^{k-l-1}\,S_{l-2} \\ + \sum_{l=0}^{k-1}(1-A_d)^l\,(1-\tilde{\alpha})^{k-l-1}\,U_l\,. \quad (75)$$

Since we suppose **HX**3 and $A_d \le A_\sigma/2$, $A_d A_\sigma \ge 8B_d B_\sigma$ we can apply Lemma 9 which shows that $1-\alpha = (1-A_d)(1+\delta_\alpha) \in (0, 1-\tilde{\alpha})$ and leads to

$$\sum_{l=0}^{k-1}(1-A_d)^l\,(1-\tilde{\alpha})^{k-l-1}\,(1+\delta_\alpha)^{l-3} \le (1+\delta_\alpha)^{-3}\sum_{l=0}^{k-1}(1-\alpha)^l\,(1-\tilde{\alpha})^{k-l-1} \\ \le \frac{(1-\alpha)^k}{(\tilde{\alpha}-\alpha)(1+\delta_\alpha)^3}\,. \quad (76)$$

Moreover, for $l \geq 2$ applying the result given by (74), we have

$$S_{l-2} \leq (1+\delta_\alpha)^{l-3}(S_1 + \delta_\alpha S_0) + \sum_{j=2}^{l-2} (1+\delta_\alpha)^{l-j-2} U_j \,. \tag{77}$$

Using the definition of U_l given by (70), we can write the following equality

$$\begin{aligned}
&\sum_{l=0}^{k-1} (1-A_d)^l (1-\tilde{\alpha})^{k-l-1} \sum_{j=2}^{l-2} (1+\delta_\alpha)^{l-j-2} U_j \\
&\qquad = \left(\mathbb{E}\left[d_0^2\right] + \frac{B_d}{A_\sigma - A_d}\mathbb{E}\left[\sigma_0^2\right]\right) \sum_{l=0}^{k-1}\sum_{j=2}^{l-2} (1-A_d)^l (1-\tilde{\alpha})^{k-l-1} (1+\delta_\alpha)^{l-j-2} \\
&\qquad\quad + \left(\frac{D_d}{A_d} + \frac{B_d D_\sigma}{A_d A_\sigma}\right) \sum_{l=0}^{k-1}\sum_{j=2}^{l-2} (1-A_d)^{l-j} (1-\tilde{\alpha})^{k-l-1} (1+\delta_\alpha)^{l-j-2} \\
&\qquad\quad + \left(C_d + \frac{B_d C_\sigma}{A_\sigma - A_d}\right) \sum_{l=0}^{k-1}\sum_{j=2}^{l-2} (1-A_d)^l (1-\tilde{\alpha})^{k-l-1} (1+\delta_\alpha)^{l-j-2} \sum_{i=0}^{j-1} (1-A_d)^{-i-1} \mathbb{E}\left[V_i\right]
\end{aligned} \tag{78}$$

We now upper bound each quantity separately. Regarding the first double sum, since $(1-A_d)(1+\delta_\alpha) = 1-\alpha$ we get

$$\begin{aligned}
&\sum_{l=0}^{k-1}\sum_{j=2}^{l-2} (1-A_d)^l (1-\tilde{\alpha})^{k-l-1} (1+\delta_\alpha)^{l-j-2} \\
&= \sum_{j=2}^{k-3} (1-A_d)^{j+2} \sum_{l=j+2}^{k-1} (1-\tilde{\alpha})^{k-l-1} (1-\alpha)^{l-j-2} \\
&\leq \frac{1}{\tilde{\alpha}-\alpha} \sum_{j=4}^{k-1} (1-A_d)^j (1-\alpha)^{k-j} \leq \frac{(1-A_d)^4 (1-\alpha)^{k-3}}{(A_d-\alpha)(\tilde{\alpha}-\alpha)} \,.
\end{aligned} \tag{79}$$

Using $(1-A_d)(1+\delta_\alpha) = 1-\alpha$ combined with $\sum_{l=j+2}^{k-1}(1-\alpha)^{l-j-2}(1-\tilde{\alpha})^{k-l-1} \leq (\tilde{\alpha}-\alpha)^{-1}(1-\alpha)^{k-j-2}$ give

$$\begin{aligned}
&\sum_{l=0}^{k-1}\sum_{j=2}^{l-2} (1-A_d)^{l-j} (1-\tilde{\alpha})^{k-l-1} (1+\delta_\alpha)^{l-j-2} \\
&= (1-A_d)^2 \sum_{l=0}^{k-1}\sum_{j=2}^{l-2} (1-\alpha)^{l-j-2} (1-\tilde{\alpha})^{k-l-1} \\
&= (1-A_d)^2 \sum_{j=2}^{k-3}\sum_{l=j+2}^{k-1} (1-\alpha)^{l-j-2} (1-\tilde{\alpha})^{k-l-1} \\
&\leq \frac{(1-A_d)^2}{\tilde{\alpha}-\alpha} \sum_{j=2}^{k-3} (1-\alpha)^{k-j-2} \leq \frac{(1-\alpha)(1-A_d)^2}{\alpha(\tilde{\alpha}-\alpha)} \,.
\end{aligned} \tag{80}$$

The same arguments show that

$$\begin{aligned}
&\sum_{l=0}^{k-1}\sum_{j=2}^{l-2} (1-A_d)^l (1-\tilde{\alpha})^{k-l-1} (1+\delta_\alpha)^{l-j-2} \sum_{i=0}^{j-1} (1-A_d)^{-i-1} \mathbb{E}\left[V_i\right] \\
&\leq \sum_{i=0}^{k-4}\sum_{j=i+1}^{k-3}\sum_{l=j+2}^{k-1} (1-A_d)^l (1-\tilde{\alpha})^{k-l-1} (1+\delta_\alpha)^{l-j-2} (1-A_d)^{-i-1} \mathbb{E}\left[V_i\right] \\
&\leq \sum_{i=0}^{k-4} \mathbb{E}\left[V_i\right] \sum_{j=i+1}^{k-3} (1-A_d)^{j-i+1} \sum_{l=j+2}^{k-1} (1-\tilde{\alpha})^{k-l-1} (1-\alpha)^{l-j-2}
\end{aligned}$$

$$\begin{aligned}
&\le \frac{1}{\tilde{\alpha}-\alpha}\sum_{i=0}^{k-4}\mathbb{E}\left[V_i\right]\sum_{j=i+1}^{k-3}\left(1-A_d\right)^{j-i+1}\left(1-\alpha\right)^{k-j-2} \\
&= \frac{\left(1-\alpha\right)\left(1-A_d\right)^2}{\tilde{\alpha}-\alpha}\sum_{i=0}^{k-4}\mathbb{E}\left[V_i\right]\sum_{j=i+1}^{k-3}\left(1-A_d\right)^{j-i-1}\left(1-\alpha\right)^{k-j-3} \\
&\le \frac{\left(1-\alpha\right)^{-1}\left(1-A_d\right)^2}{(A_d-\alpha)(\tilde{\alpha}-\alpha)}\sum_{i=0}^{k-4}\left(1-\alpha\right)^{k-i-1}\mathbb{E}\left[V_i\right] . \qquad (81)
\end{aligned}$$

Therefore, plugging (79), (80), (81) inside (78) implies

$$\begin{aligned}
\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2}U_j &\le \frac{\left(1-\alpha\right)\left(1-A_d\right)^2}{\alpha(\tilde{\alpha}-\alpha)}\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right) \\
&+ \frac{\left(1-A_d\right)^4\left(1-\alpha\right)^{k-3}}{(A_d-\alpha)(\tilde{\alpha}-\alpha)}\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right) \\
&+ \frac{\left(1-\alpha\right)^{-1}\left(1-A_d\right)^2}{(A_d-\alpha)(\tilde{\alpha}-\alpha)}\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-4}\left(1-\alpha\right)^{k-i-1}\mathbb{E}\left[V_i\right] . \qquad (82)
\end{aligned}$$

In addition, by definition of U_l provides in (70) we have

$$\begin{aligned}
\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}U_l &= \left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\sum_{l=0}^{k-1}\left(1-\tilde{\alpha}\right)^{k-l-1} \\
&+ \left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1} \\
&+ \left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}\sum_{i=0}^{l-1}\left(1-A_d\right)^{-i-1}\mathbb{E}\left[V_i\right] .
\end{aligned}$$

Thus, a calculation yields that

$$\begin{aligned}
\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}U_l &\le \frac{\left(1-A_d\right)^k}{\tilde{\alpha}-A_d}\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right) \\
&+ \frac{1}{\tilde{\alpha}}\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)+\frac{1}{\tilde{\alpha}-A_d}\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-2}\left(1-A_d\right)^{k-i-1}\mathbb{E}\left[V_i\right] . \qquad (83)
\end{aligned}$$

Plugging (77) in (75) shows

$$\begin{aligned}
\sum_{l=0}^{k-1}\left(1-\tilde{\alpha}\right)^{k-l-1}\mathbb{E}\left[d_l^2\right] &\le \frac{B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)}{(1-A_d)(A_\sigma-A_d)}\sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-3} \\
&+ \frac{B_dB_\sigma}{(1-A_d)(A_\sigma-A_d)}\sum_{l=0}^{k-1}\sum_{j=2}^{l-2}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}\left(1+\delta_\alpha\right)^{l-j-2}U_j \\
&+ \sum_{l=0}^{k-1}\left(1-A_d\right)^l\left(1-\tilde{\alpha}\right)^{k-l-1}U_l . \qquad (84)
\end{aligned}$$

Hence, by combining (76), (82), (83) and (84) we obtain for $A_d > \alpha$, that

$$\sum_{l=0}^{k-1}\left(1-\tilde{\alpha}\right)^{k-l-1}\mathbb{E}\left[d_l^2\right] \le \frac{B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)\left(1-\alpha\right)^k}{(1-A_d)(A_\sigma-A_d)(\tilde{\alpha}-\alpha)(1+\delta_\alpha)^3}$$

$$
\begin{aligned}
&+\left(\frac{(1-A_d)^k}{\tilde{\alpha}-A_d}+\frac{B_dB_\sigma\,(1-\alpha)^k}{(A_\sigma-A_d)(A_d-\alpha)(\tilde{\alpha}-\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\\
&+\left(\frac{1}{\tilde{\alpha}}+\frac{B_dB_\sigma}{\alpha(\tilde{\alpha}-\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\\
&+\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-2}\left(\frac{(1-A_d)^{k-i-1}}{\tilde{\alpha}-A_d}+\frac{B_dB_\sigma\,(1-\alpha)^{k-i-1}}{(A_\sigma-A_d)(A_d-\alpha)(\tilde{\alpha}-\alpha)}\right)\mathbb{E}\left[V_i\right]\,. \qquad (85)
\end{aligned}
$$

In addition, the above bound holds even if $A_d=\alpha$ by considering that $(A_d-\alpha)^{-1}B_dB_\sigma=0$.

Upper bound on $\sum_{l=0}^{k-1}(1-p_{\mathrm{c}}/4)^{k-l-1}\mathbb{E}\left[d_l^2\right]$. Applying (85) with $\tilde{\alpha}=p_{\mathrm{c}}/4$ gives

$$
\begin{aligned}
\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}\mathbb{E}\left[d_l^2\right] \le{}& \frac{4B_dB_\sigma\,(S_1+\delta_\alpha S_0)\,(1-\alpha)^k}{(1-A_d)(A_\sigma-A_d)(p_{\mathrm{c}}-4\alpha)(1+\delta_\alpha)^3}\\
&+\left(\frac{4\,(1-A_d)^k}{p_{\mathrm{c}}-4A_d}+\frac{4B_dB_\sigma\,(1-\alpha)^k}{(A_\sigma-A_d)(A_d-\alpha)(p_{\mathrm{c}}-4\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\\
&+\left(\frac{4}{p_{\mathrm{c}}}+\frac{4B_dB_\sigma}{\alpha(p_{\mathrm{c}}-4\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\\
&+4\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-2}\left(\frac{(1-A_d)^{k-i-1}}{p_{\mathrm{c}}-4A_d}+\frac{B_dB_\sigma\,(1-\alpha)^{k-i-1}}{(A_\sigma-A_d)(A_d-\alpha)(p_{\mathrm{c}}-4\alpha)}\right)\mathbb{E}\left[V_i\right]\,. \qquad (86)
\end{aligned}
$$

Upper bound on $\sum_{l=0}^{k-1}(1-p_{\mathrm{c}}/4)^{k-l-1}\sum_{j=0}^{l-1}(1-A_\sigma)^{l-j-1}\mathbb{E}[d_j^2]$. Recall that we consider that $(A_d-\alpha)^{-1}B_dB_\sigma=0$ in the specific case where $A_d=\alpha$. This time, setting $\tilde{\alpha}=A_\sigma$ in (85) shows that

$$
\begin{aligned}
\sum_{j=0}^{l-1}(1-A_\sigma)^{l-j-1}\mathbb{E}\left[d_l^2\right] \le{}& \frac{B_dB_\sigma\,(S_1+\delta_\alpha S_0)\,(1-\alpha)^l}{(1-A_d)(A_\sigma-A_d)(A_\sigma-\alpha)(1+\delta_\alpha)^3}\\
&+\left(\frac{(1-A_d)^l}{A_\sigma-A_d}+\frac{B_dB_\sigma\,(1-\alpha)^l}{(A_\sigma-A_d)\,(A_d-\alpha)\,(A_\sigma-\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\\
&+\left(\frac{1}{A_\sigma}+\frac{B_dB_\sigma}{\alpha(A_\sigma-\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\\
&+\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{l-2}\left(\frac{(1-A_d)^{l-i-1}}{A_\sigma-A_d}+\frac{B_dB_\sigma\,(1-\alpha)^{l-i-1}}{(A_\sigma-A_d)(A_d-\alpha)(A_\sigma-\alpha)}\right)\mathbb{E}\left[V_i\right]\,. \qquad (87)
\end{aligned}
$$

Moreover, we have the two following bounds

$$
\begin{aligned}
\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}(1-A_d)^l &\le \frac{4\,(1-A_d)^k}{p_{\mathrm{c}}-4A_d}\,,\\
\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}(1-\alpha)^l &\le \frac{4\,(1-\alpha)^k}{p_{\mathrm{c}}-4\alpha}\,. \qquad (88)
\end{aligned}
$$

Therefore, permuting the summations implies

$$
\begin{aligned}
\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}\sum_{i=0}^{l-2}(1-A_d)^{l-i-1}\mathbb{E}\left[V_i\right] &\le \sum_{i=0}^{k-3}\mathbb{E}\left[V_i\right]\sum_{l=i+2}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}(1-A_d)^{l-i-1}\\
&\le \frac{4}{p_{\mathrm{c}}-4A_d}\sum_{i=0}^{k-3}(1-A_d)^{k-i-1}\mathbb{E}\left[V_i\right]\,. \qquad (89)
\end{aligned}
$$

In a similar way, we obtain

$$\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}\sum_{i=0}^{l-2}(1-\alpha)^{l-i-1}\,\mathbb{E}\left[V_i\right] \le \frac{4}{p_{\mathrm{c}}-4\alpha}\sum_{i=0}^{k-3}(1-\alpha)^{k-i-1}\,\mathbb{E}\left[V_i\right]\,. \tag{90}$$

Hence, the combination of (87) with (88), (89), (90) yields

$$\begin{aligned}
\sum_{l=0}^{k-1}\left(1-\frac{p_{\mathrm{c}}}{4}\right)^{k-l-1}\sum_{j=0}^{l-1}(1-A_\sigma)^{l-j-1}\,\mathbb{E}\left[d_l^2\right] &\le \frac{4B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)(1-\alpha)^k}{(p_{\mathrm{c}}-4\alpha)(1-A_d)(A_\sigma-A_d)(A_\sigma-\alpha)(1+\delta_\alpha)^3}\\
&+\frac{4}{A_\sigma-A_d}\left(\frac{(1-A_d)^k}{p_{\mathrm{c}}-4A_d}+\frac{B_dB_\sigma(1-\alpha)^k}{(p_{\mathrm{c}}-4\alpha)(A_d-\alpha)(A_\sigma-\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\\
&+\frac{4}{p_{\mathrm{c}}}\left(\frac{1}{A_\sigma}+\frac{B_dB_\sigma}{\alpha(A_\sigma-\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\\
&+\frac{4}{A_\sigma-A_d}\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-3}\left(\frac{(1-A_d)^{k-i-1}}{p_{\mathrm{c}}-4A_d}+\frac{B_dB_\sigma(1-\alpha)^{k-i-1}}{(p_{\mathrm{c}}-4\alpha)(A_d-\alpha)(A_\sigma-\alpha)}\right)\mathbb{E}\left[V_i\right]\,.
\end{aligned} \tag{91}$$

Upper bound on $\mathbb{E}\left[V_k\right]$. Plugging (86) and (91) in (65), we obtain

$$\begin{aligned}
\mathbb{E}\left[V_k\right] &\le \left(1-\frac{p_{\mathrm{c}}}{4}\right)^k\mathbb{E}\left[V_0\right]+\frac{4(1-p_{\mathrm{c}})\gamma^2(1-A_d)^k}{p_{\mathrm{c}}-4A_d}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_0^2\right]\\
&+\frac{4(1-p_{\mathrm{c}})\gamma^2D_\sigma}{A_\sigma(p_{\mathrm{c}}-4A_d)}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)+\frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right)+\frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{bp_{\mathrm{c}}}\\
&+(1-p_{\mathrm{c}})\gamma^2\left(B+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\Bigg[\frac{4B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)(1-\alpha)^k}{(1-A_d)(A_\sigma-A_d)(p_{\mathrm{c}}-4\alpha)(1+\delta_\alpha)^3}\\
&+\left(\frac{4\left(1-A_d\right)^k}{p_{\mathrm{c}}-4A_d}+\frac{4B_dB_\sigma(1-\alpha)^k}{(A_\sigma-A_d)(A_d-\alpha)(p_{\mathrm{c}}-4\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\\
&+\left(\frac{4}{p_{\mathrm{c}}}+\frac{4B_dB_\sigma}{\alpha(p_{\mathrm{c}}-4\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\\
&+4\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-2}\left(\frac{(1-A_d)^{k-i-1}}{p_{\mathrm{c}}-4A_d}+\frac{B_dB_\sigma(1-\alpha)^{k-i-1}}{(A_\sigma-A_d)(A_d-\alpha)(p_{\mathrm{c}}-4\alpha)}\right)\mathbb{E}\left[V_i\right]\Bigg]\\
&+4(1-p_{\mathrm{c}})\gamma^2B_\sigma\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\Bigg[\frac{B_dB_\sigma\left(S_1+\delta_\alpha S_0\right)(1-\alpha)^k}{(p_{\mathrm{c}}-4\alpha)(1-A_d)(A_\sigma-A_d)(A_\sigma-\alpha)(1+\delta_\alpha)^3}\\
&+\frac{1}{A_\sigma-A_d}\left(\frac{(1-A_d)^k}{p_{\mathrm{c}}-4A_d}+\frac{B_dB_\sigma(1-\alpha)^k}{(p_{\mathrm{c}}-4\alpha)(A_d-\alpha)(A_\sigma-\alpha)}\right)\left(\mathbb{E}\left[d_0^2\right]+\frac{B_d}{A_\sigma-A_d}\mathbb{E}\left[\sigma_0^2\right]\right)\\
&+\frac{1}{p_{\mathrm{c}}}\left(\frac{1}{A_\sigma}+\frac{B_dB_\sigma}{\alpha(A_\sigma-\alpha)(A_\sigma-A_d)}\right)\left(\frac{D_d}{A_d}+\frac{B_dD_\sigma}{A_dA_\sigma}\right)\\
&+\frac{1}{A_\sigma-A_d}\left(C_d+\frac{B_dC_\sigma}{A_\sigma-A_d}\right)\sum_{i=0}^{k-3}\left(\frac{(1-A_d)^{k-i-1}}{p_{\mathrm{c}}-4A_d}+\frac{B_dB_\sigma(1-\alpha)^{k-i-1}}{(p_{\mathrm{c}}-4\alpha)(A_d-\alpha)(A_\sigma-\alpha)}\right)\mathbb{E}\left[V_i\right]\Bigg]\\
&+\frac{4(1-p_{\mathrm{c}})\gamma^2C_\sigma}{p_{\mathrm{c}}-4A_d}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\sum_{l=0}^{k-2}(1-A_d)^{k-l-1}\,\mathbb{E}\left[V_l\right]\,.
\end{aligned} \tag{92}$$

For any negative number $j<0$, using the convention that $\sum_{l=0}^{j}=0$ and simplifying the calculations provided by (92), we find that

$$\mathbb{E}\left[V_k\right] \le \left(1-\frac{p_{\mathrm{c}}}{4}\right)^k\mathbb{E}\left[V_0\right]+\frac{4(1-p_{\mathrm{c}})\gamma^2(1-A_d)^k}{p_{\mathrm{c}}-4A_d}\left(C+\frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_0^2\right]$$

$$
\begin{aligned}
&+ \frac{4(1-p_{\mathrm{c}})\gamma^2 B_d B_\sigma \left(S_1 + \delta_\alpha S_0\right) (1-\alpha)^k}{(p_{\mathrm{c}}-4\alpha)(1-A_d)(A_\sigma - A_d)(1+\delta_\alpha)^3} \left[B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B} + \frac{B_\sigma}{A_\sigma - \alpha}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right] \\
&+ \frac{4(1-p_{\mathrm{c}})\gamma^2 D_\sigma}{A_\sigma(p_{\mathrm{c}}-4A_d)}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) + \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + \frac{8\,(1-\tau)\,(b-1)\,\gamma d}{b p_{\mathrm{c}}} \\
&+ 4(1-p_{\mathrm{c}})\gamma^2 \Bigg[\left(\frac{1}{p_{\mathrm{c}}} + \frac{B_d B_\sigma}{\alpha(p_{\mathrm{c}}-4\alpha)(A_\sigma - A_d)}\right)\left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right) \\
&\quad + \frac{B_\sigma}{p_{\mathrm{c}}}\left(\frac{1}{A_\sigma} + \frac{B_d B_\sigma}{\alpha(A_\sigma-\alpha)(A_\sigma - A_d)}\right)\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\Bigg]\left(\frac{D_d}{A_d} + \frac{B_d D_\sigma}{A_d A_\sigma}\right) \\
&+ \frac{4\gamma^2\,(1-p_{\mathrm{c}})\,(1-A_d)^k}{p_{\mathrm{c}}-4A_d}\left[B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B} + \frac{B_\sigma}{A_\sigma - A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right]\left(\mathbb{E}\left[d_0^2\right] + \frac{B_d}{A_\sigma - A_d}\mathbb{E}\left[\sigma_0^2\right]\right) \\
&+ \frac{4\gamma^2\,(1-p_{\mathrm{c}})\,B_d B_\sigma\,(1-\alpha)^k}{(p_{\mathrm{c}}-4\alpha)\,(A_d-\alpha)\,(A_\sigma - A_d)}\left[B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B} + \frac{B_\sigma}{A_\sigma - \alpha}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right]\left(\mathbb{E}\left[d_0^2\right] + \frac{B_d}{A_\sigma - A_d}\mathbb{E}\left[\sigma_0^2\right]\right) \\
&+ \frac{4\gamma^2\,(1-p_{\mathrm{c}})}{p_{\mathrm{c}}-4A_d}\left[C_\sigma\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) + \left(C_d + \frac{B_d C_\sigma}{A_\sigma - A_d}\right)\left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B} + \frac{B_\sigma}{A_\sigma - A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right)\right] \\
&\qquad \times \sum_{i=0}^{k-2}(1-A_d)^{k-i-1}\,\mathbb{E}\left[V_i\right] \\
&+ \frac{4\gamma^2\,(1-p_{\mathrm{c}})\,B_d B_\sigma}{(p_{\mathrm{c}}-4\alpha)\,(A_d-\alpha)\,(A_\sigma - A_d)}\left(C_d + \frac{B_d C_\sigma}{A_\sigma - A_d}\right)\left[B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B} + \frac{B_\sigma}{A_\sigma - \alpha}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right] \\
&\qquad \times \sum_{i=0}^{k-3}(1-\alpha)^{k-i-1}\,\mathbb{E}\left[V_i\right]. \qquad (93)
\end{aligned}
$$

As explained in (73), recall that

$$
\delta_\alpha^2 + \delta_\alpha = \frac{B_d B_\sigma}{(1-A_d)\,(A_\sigma - A_d)}\,, \qquad \alpha = A_d - \delta_\alpha(1-A_d)\,.
$$

Thus, when $B_d B_\sigma \neq 0$ then $\delta_\alpha \neq 0$, which implies that $A_d \neq \alpha$ and gives

$$
\frac{B_d B_\sigma}{(A_d - \alpha)\,(A_\sigma - A_d)} = 1 + \delta_\alpha\,.
$$

In addition, in the proof of Lemma 9 we saw that $2\delta_\alpha \leq A_d \leq 1/2$ and also that $A_d/2 \leq \alpha \leq A_d$. Therefore, we can regroup several terms in (93) and write

$$
\begin{aligned}
\mathbb{E}\left[V_k\right] \leq{}& \left(1 - \frac{p_{\mathrm{c}}}{4}\right)^k \mathbb{E}\left[V_0\right] + \frac{4(1-p_{\mathrm{c}})\gamma^2\,(1-A_d)^k}{p_{\mathrm{c}}-4A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_0^2\right] \\
&+ \frac{4(1-p_{\mathrm{c}})\gamma^2 \delta_\alpha\left(S_1 + \delta_\alpha S_0\right)(1-\alpha)^k}{(p_{\mathrm{c}}-4A_d)(1+\delta_\alpha)^2}\left[B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B} + \frac{B_\sigma}{A_\sigma - A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right] \\
&+ \frac{9\gamma^2\,(1-p_{\mathrm{c}})\,(1-\alpha)^k}{p_{\mathrm{c}}-4A_d}\left[B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B} + \frac{B_\sigma}{A_\sigma - A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right]\left(\mathbb{E}\left[d_0^2\right] + \frac{B_d}{A_\sigma - A_d}\mathbb{E}\left[\sigma_0^2\right]\right) \\
&+ \frac{4(1-p_{\mathrm{c}})\gamma^2 D_\sigma}{A_\sigma(p_{\mathrm{c}}-4A_d)}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) + \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + \frac{8\,(1-\tau)\,(b-1)\,\gamma d}{b p_{\mathrm{c}}} \\
&+ 4(1-p_{\mathrm{c}})\gamma^2\Bigg[\left(\frac{1}{p_{\mathrm{c}}} + \frac{2B_d B_\sigma}{A_d(p_{\mathrm{c}}-4A_d)(A_\sigma - A_d)}\right)\left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right) \\
&\quad + \frac{B_\sigma}{p_{\mathrm{c}}}\left(\frac{1}{A_\sigma} + \frac{2B_d B_\sigma}{A_d(A_\sigma - A_d)^2}\right)\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\Bigg]\left(\frac{D_d}{A_d} + \frac{B_d D_\sigma}{A_d A_\sigma}\right) \\
&+ \frac{9\gamma^2\,(1-p_{\mathrm{c}})}{p_{\mathrm{c}}-4A_d}\left[C_\sigma\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) + \left(C_d + \frac{B_d C_\sigma}{A_\sigma - A_d}\right)\left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B} + \frac{B_\sigma}{A_\sigma - A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right)\right]
\end{aligned}
$$

$$\times \sum_{i=0}^{k-2} (1-\alpha)^{k-i-1} \mathbb{E}\left[V_i\right] . \qquad (94)$$

Recall that we defined C^γ in (59) by

$$\mathrm{C}^\gamma = \frac{4(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c} - 4A_d}\left[B + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{B} + \frac{B_\sigma}{A_\sigma - A_d}\left(C + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{C}\right)\right] .$$

Hence, using (94) we get that

$$\begin{aligned}
\mathbb{E}\left[V_k\right] \leq \left(1-\frac{p_\mathrm{c}}{4}\right)^k \mathbb{E}\left[V_0\right] &+ \frac{4(1-p_\mathrm{c})\gamma^2 D_\sigma}{A_\sigma(p_\mathrm{c}-4A_d)}\left(C + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{C}\right) + \frac{4(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}}\left(D + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{D}\right) \\
&+ \frac{\mathrm{C}^\gamma}{A_d}\left(1 + \frac{2B_dB_\sigma}{A_d(A_\sigma - A_d)}\right)\left(D_d + \frac{B_dD_\sigma}{A_\sigma}\right) + \frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{bp_\mathrm{c}} \\
&+ \left(\frac{4(1-p_\mathrm{c})\gamma^2\left(1-A_d\right)^k}{p_\mathrm{c}-4A_d}\left(C + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{C}\right) + \frac{9\mathrm{C}^\gamma B_d\left(1-\alpha\right)^k}{4\left(A_\sigma - A_d\right)}\right)\mathbb{E}\left[\sigma_0^2\right] \\
&+ \frac{9}{4}\mathrm{C}^\gamma\left(1-\alpha\right)^k \mathbb{E}\left[d_0^2\right] + \mathrm{C}^\gamma\left(1-\alpha\right)^{k-2}\left(A_d-\alpha\right)\left(1-A_d\right)\left(S_1 + \frac{A_d-\alpha}{1-A_d}S_0\right) \\
&+ \left[\frac{9\gamma^2\left(1-p_\mathrm{c}\right)C_\sigma}{p_\mathrm{c}-4A_d}\left(C + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{C}\right) + 3\mathrm{C}^\gamma\left(C_d + \frac{B_dC_\sigma}{A_\sigma - A_d}\right)\right]\sum_{i=0}^{k-2}(1-\alpha)^{k-i-1}\mathbb{E}\left[V_i\right] .
\end{aligned}$$

Finally, we conclude the proof remarking that

$$\begin{aligned}
&\mathrm{C}^\gamma\left(1-\alpha\right)^{k-2}\left(A_d-\alpha\right)\left[(1-A_d)S_1 + (A_d-\alpha)S_0\right] \\
&\qquad \leq \mathrm{C}^\gamma\left(1-\alpha\right)^{k-2}\left(A_d-\alpha\right)\left[(2-A_d-\alpha)\mathbb{E}\left[d_0^2\right] + B_d\mathbb{E}\left[\sigma_0^2\right] + C_d\mathbb{E}\left[V_0\right] + D_d\right] \\
&\qquad\qquad \leq \mathrm{C}^\gamma\left(1-\alpha\right)^k\left(4\mathbb{E}\left[d_0^2\right] + 2B_d\mathbb{E}\left[\sigma_0^2\right] + 2C_d\mathbb{E}\left[V_0\right] + 2D_d\right) .
\end{aligned}$$

□

In order to ease notation, with the definitions used in **HX**4 and (59), consider for any $\gamma \in \mathbb{R}_+$ the variable $\mathrm{C}_\epsilon^\gamma \in \mathbb{R}_+$ defined by

$$\mathrm{C}_\epsilon^\gamma = \mathrm{C}_V^\gamma\mathbb{E}\left[V_0\right] + \mathrm{C}_d^\gamma\mathbb{E}\left[d_0^2\right] + \mathrm{C}_\sigma^\gamma\mathbb{E}\left[\sigma_0^2\right] + 2D_d \qquad (95)$$

In addition, with the previous notations consider

$$\delta = \frac{2\left(1-A_d/2\right)^{-1}\mathrm{C}_r^\gamma}{1+\sqrt{1+4\left(1-A_d/2\right)^{-1}\mathrm{C}_r^\gamma}}$$

and define

$$\gamma_V = \frac{p_\mathrm{c}^{1/2}}{(2-2p_\mathrm{c})^{1/2}\left[A + (1+2/p_\mathrm{c})\bar{A}\right]^{1/2}} .$$

Lemma 11. *Assume **HX**3, **HX**4 hold with $4\mathrm{C}_r^\gamma \leq A_d < \min(A_\sigma/2, p_\mathrm{c}/4)$, $A_dA_\sigma \geq 8B_dB_\sigma$ and let $\gamma \in (0, \gamma_V]$. Then, for any $k \geq 1$, we have*

$$\mathbb{E}\left[V_k\right] \leq \left(1-\frac{A_d}{4}\right)^k\left(2\mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_r^\gamma\mathrm{C}_\delta^\gamma}{A_d}\right) + \mathrm{C}_\delta^\gamma ,$$

where V_k is defined in (24), $\mathrm{C}_\epsilon^\gamma, \mathrm{C}_r^\gamma, \mathrm{C}_\delta^\gamma$ in (59) and (95).

Proof. Let k in $\mathbb{N}$ be fixed. Since the assumptions of Lemma 10 are satisfied, we know that

$$\mathbb{E}\left[V_k\right] \leq \left(1-\alpha\right)^k\mathrm{C}_\epsilon^\gamma + \mathrm{C}_r^\gamma\sum_{l=0}^{k-2}\left(1-\alpha\right)^{k-l-1}\mathbb{E}\left[V_l\right] + \mathrm{C}_\delta^\gamma ,$$

where α is defined in (60). In addition, Lemma 9 shows that $A_d/2 \le \alpha$. Hence, multiplying the last inequality by the weight ω_k defined for any $l \in \mathbb{N}$, by

$$\omega_l = (1 - A_d/2)^{-l}\ ,$$

we obtain the following inequality

$$\omega_k \mathbb{E}\left[V_k\right] \le \mathrm{C}_\epsilon^\gamma + \frac{\mathrm{C}_r^\gamma}{1 - A_d/2} \sum_{l=0}^{k-2} \omega_l \mathbb{E}\left[V_l\right] + \mathrm{C}_\delta^\gamma \omega_k\ .$$

Applying the sharp Grönwall inequality (Holte, 2009), we get

$$\omega_k \mathbb{E}\left[V_k\right] \le \mathrm{C}_\epsilon^\gamma + \omega_k \mathrm{C}_\delta^\gamma + \frac{\mathrm{C}_r^\gamma}{1 - A_d/2} \sum_{l=0}^{k-1} \left(\mathrm{C}_\epsilon^\gamma + \omega_l \mathrm{C}_\delta^\gamma\right) \left(1 + \frac{\mathrm{C}_r^\gamma}{1 - A_d/2}\right)^{k-l-1}\ .$$

Therefore, a calculation shows that

$$\omega_k \mathbb{E}\left[V_k\right] \le \mathrm{C}_\epsilon^\gamma + \omega_k \mathrm{C}_\delta^\gamma + \mathrm{C}_\epsilon^\gamma \left(1 + \frac{\mathrm{C}_r^\gamma}{1 - A_d/2}\right)^{k} + \frac{\mathrm{C}_r^\gamma \mathrm{C}_\delta^\gamma}{1 - A_d/2} \sum_{l=0}^{k-1} \omega_l \left(1 + \frac{\mathrm{C}_r^\gamma}{1 - A_d/2}\right)^{k-l-1}\ ,$$

and simplifying the previous inequality gives the following upper bound:

$$\mathbb{E}\left[V_k\right] \le \mathrm{C}_\delta^\gamma + \omega_k^{-1} \mathrm{C}_\epsilon^\gamma + \mathrm{C}_\epsilon^\gamma \left(1 - \frac{A_d}{2} + \mathrm{C}_r^\gamma\right)^{k} + \mathrm{C}_r^\gamma \mathrm{C}_\delta^\gamma \sum_{l=0}^{k-1} \left(1 - \frac{A_d}{2} + \mathrm{C}_r^\gamma\right)^{k-l-1}\ . \tag{96}$$

In addition, using $4\mathrm{C}_r^\gamma < A_d < p_{\mathrm{c}}/4$ implies $0 < 1 - A_d/2 + \mathrm{C}_r^\gamma < 1$ which combined with (96) gives

$$\mathbb{E}\left[V_k\right] \le \mathrm{C}_\delta^\gamma + \omega_k^{-1} \mathrm{C}_\epsilon^\gamma + \mathrm{C}_\epsilon^\gamma \left(1 - \frac{A_d}{2} + \mathrm{C}_r^\gamma\right)^{k} + \frac{\mathrm{C}_r^\gamma \mathrm{C}_\delta^\gamma}{A_d/2 - \mathrm{C}_r^\gamma} \left(1 - \frac{A_d}{2} + \mathrm{C}_r^\gamma\right)^{k}\ .$$

Eventually, combining the last inequality with the assumption $4\mathrm{C}_r^\gamma < A_d$ completes the proof. □

With the notation of the assumptions **HX**3 and **HX**4, we define

$$\alpha_d = \frac{4\gamma^2}{p_{\mathrm{c}} A_d} \max\left\{p_{\mathrm{c}} B + 3\bar{B}, \frac{4B_\sigma}{A_\sigma}\left(p_{\mathrm{c}} C + 3\bar{C}\right)\right\}\ , \qquad \alpha_\sigma = \frac{4\gamma^2\left(p_{\mathrm{c}} C + 3\bar{C}\right)}{p_{\mathrm{c}} A_\sigma}\ . \tag{97}$$

The following lemma is used in the convergence proof of VR-FALD⋆ (see Lemma 26).

Lemma 12. *Assume **HX**3, **HX**4 hold with*

$$A_d \le \min\left(A_\sigma, \frac{p_{\mathrm{c}}}{4}\right)\ , \qquad \alpha_d C_d + \alpha_\sigma C_\sigma \le \frac{p_{\mathrm{c}}}{8}\ , \qquad \alpha_d B_d + \gamma^2\left(C + \frac{3}{p_{\mathrm{c}}}\bar{C}\right) \le \frac{\alpha_\sigma A_\sigma}{2}\ ,$$

and consider $\gamma \le p_{\mathrm{c}}^{1/2}(2-2p_{\mathrm{c}})^{-1/2}[A + (1+2/p_{\mathrm{c}})\bar{A}]^{-1/2}$. Then, for any $k \in \mathbb{N}$, we have

$$\begin{aligned}\mathbb{E}\left[V_k\right] + \alpha_d \mathbb{E}\left[d_k^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_k^2\right] \le \left(1 - \frac{A_d}{2}\right)^{k} \left(\mathbb{E}\left[V_0\right] + \alpha_d \mathbb{E}\left[d_0^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_0^2\right]\right) \\ + \frac{2(1-p_{\mathrm{c}})\gamma^2}{A_d}\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + \frac{2\alpha_d D_d + 2\alpha_\sigma D_\sigma}{A_d} + \frac{4\left(1-\tau\right)\left(b-1\right)\gamma d}{b A_d}\ ,\end{aligned}$$

where V_k is defined in (24).

Proof. Let $k \in \mathbb{N}^\star$, using for $i \in [b]$ the definitions (20), (23) of X_k^i and X_k

$$X_{k+1}^i = X_k^i - \gamma G_k^i + \sqrt{2\gamma}\left(\sqrt{\tau/b}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i\right)\ ,$$

$$X_{k+1} = X_k - \frac{\gamma}{b}\sum_{j=1}^{b} G_k^i + \sqrt{\frac{2\gamma\tau}{b}}\tilde{Z}_{k+1} + \frac{\sqrt{2(1-\tau)\gamma}}{b}\sum_{i=1}^{b} Z_{k+1}^i \,.$$

Substracting the two above equations combined with the Jensen inequality give

$$\begin{aligned}
&\mathbb{E}\left[V_{k+1}\right] = \frac{1}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|X_{k+1}^i - X_{k+1}\right\|^2\right] \\
&= \frac{1-p_{\mathrm{c}}}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(G_k^i - G^k) + \sqrt{2(1-\tau)\gamma}Z_{k+1}^i - \frac{\sqrt{2(1-\tau)\gamma}}{b}\sum_{j=1}^{b} Z_{k+1}^j\right\|^2\right] \\
&= \frac{1-p_{\mathrm{c}}}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(\bar{G}_k^i - \bar{G}^k)\right\|^2\right] + \frac{(1-p_{\mathrm{c}})\gamma^2}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|(G_k^i - \bar{G}_k^i) - (G^k - \bar{G}^k)\right\|^2\right] \\
&\quad + 2(1-\tau)\gamma\mathbb{E}\left[\left\|Z_{k+1}^i - \frac{1}{b}\sum_{j=1}^{b} Z_{k+1}^j\right\|^2\right]
\end{aligned}$$

Hence, we get

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] &\leq \frac{1-p_{\mathrm{c}}}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|(X_k^i - X_k) - \gamma(\bar{G}_k^i - \bar{G}^k)\right\|^2\right] + \frac{(1-p_{\mathrm{c}})\gamma^2}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right] \\
&\quad + 2\left(1-\tau\right)\left(1-1/b\right)\gamma d \\
&\leq \frac{(1-p_{\mathrm{c}})(1+p_{\mathrm{c}}/2)}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|X_k^i - X_k\right\|^2\right] + \frac{(1-p_{\mathrm{c}})\gamma^2}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right] \\
&\quad + \frac{(1-p_{\mathrm{c}})(1+2/p_{\mathrm{c}})\gamma^2}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|\bar{G}_k^i - \bar{G}^k\right\|^2\right] + 2\left(1-\tau\right)\left(1-1/b\right)\gamma d \,.
\end{aligned}$$

We finally obtain

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] \leq \left(1-p_{\mathrm{c}}/2\right)\mathbb{E}\left[V_k\right] + \frac{(1-p_{\mathrm{c}})(2+p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|\bar{G}_k^i\right\|^2\right] \\
+ \frac{(1-p_{\mathrm{c}})\gamma^2}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right] + 2\left(1-\tau\right)\left(1-\frac{1}{b}\right)\gamma d \,.
\end{aligned}$$

Combining the last inequality with **HX4** shows

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] &\leq \left(1 - \frac{p_{\mathrm{c}}}{2} + (1-p_{\mathrm{c}})\gamma^2\left[A + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{A}\right]\right)\mathbb{E}\left[V_k\right] + (1-p_{\mathrm{c}})\gamma^2\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) \\
&\quad + (1-p_{\mathrm{c}})\gamma^2\left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\mathbb{E}\left[d_k^2\right] + (1-p_{\mathrm{c}})\gamma^2\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_k^2\right] + 2\left(1-\tau\right)\left(1-\frac{1}{b}\right)\gamma d \,.
\end{aligned}$$

Since $\gamma \leq \frac{p_{\mathrm{c}}^{1/2}}{2(1-p_{\mathrm{c}})^{1/2}\left[A+(1+2/p_{\mathrm{c}})\bar{A}\right]^{1/2}}$, the above inequality implies that

$$\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] &\leq \left(1 - \frac{p_{\mathrm{c}}}{4}\right)\mathbb{E}\left[V_k\right] + (1-p_{\mathrm{c}})\gamma^2\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + 2\left(1-\tau\right)\left(1-1/b\right)\gamma d \\
&\quad + (1-p_{\mathrm{c}})\gamma^2\left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\mathbb{E}\left[d_k^2\right] + (1-p_{\mathrm{c}})\gamma^2\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\mathbb{E}\left[\sigma_k^2\right] \,.
\end{aligned}$$

The previous bound combined with **HX3** gives that

$$
\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] + \alpha_d \mathbb{E}\left[d_{k+1}^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_{k+1}^2\right] \leq & \left[\left(1 - \frac{p_{\mathrm{c}}}{4}\right) + \alpha_d C_d + \alpha_\sigma C_\sigma\right] \mathbb{E}\left[V_k\right] \\
& + \left[\alpha_d(1 - A_d) + \alpha_\sigma B_\sigma + (1 - p_{\mathrm{c}})\gamma^2 \left(B + \frac{2 + p_{\mathrm{c}}}{p_{\mathrm{c}}} \bar{B}\right)\right] \mathbb{E}\left[d_k^2\right] \\
& + \left[\alpha_\sigma(1 - A_\sigma) + \alpha_d B_d + (1 - p_{\mathrm{c}})\gamma^2 \left(C + \frac{2 + p_{\mathrm{c}}}{p_{\mathrm{c}}} \bar{C}\right)\right] \mathbb{E}\left[\sigma_k^2\right] \\
& + (1 - p_{\mathrm{c}})\gamma^2 \left(D + \frac{2 + p_{\mathrm{c}}}{p_{\mathrm{c}}} \bar{D}\right) + 2\,(1 - \tau)\, \frac{(b-1)}{b} \gamma d + \alpha_d D_d + \alpha_\sigma D_\sigma \,. \qquad (98)
\end{aligned}
$$

By assumption, we have

$$
\begin{aligned}
& \alpha_d C_d + \alpha_\sigma C_\sigma \leq \frac{p_{\mathrm{c}}}{8} \,, \\
& \alpha_d B_d + \gamma^2 \left(C + \frac{3}{p_{\mathrm{c}}} \bar{C}\right) \leq \frac{\alpha_\sigma A_\sigma}{2} \,,
\end{aligned} \qquad (99)
$$

and by definition of α_d, α_σ given in (97), we know that $\alpha_\sigma B_\sigma + \gamma^2(B + 3\bar{B}/p_{\mathrm{c}}) \leq \alpha_d A_d/2$. In addition, since we suppose that $A_d \leq \min(p_{\mathrm{c}}/4, A_\sigma)$, the last inequalities combined with (99) imply

$$
\begin{aligned}
& 1 - \frac{p_{\mathrm{c}}}{4} + \alpha_d C_d + \alpha_\sigma C_\sigma \leq 1 - \frac{A_d}{2} \\
& 1 - A_d + \frac{\alpha_\sigma}{\alpha_d} B_\sigma + \frac{(1 - p_{\mathrm{c}})\gamma^2}{\alpha_d} \left(B + \frac{2 + p_{\mathrm{c}}}{p_{\mathrm{c}}} \bar{B}\right) \leq 1 - \frac{A_d}{2} \\
& 1 - A_\sigma + \frac{\alpha_d}{\alpha_\sigma} B_d + \frac{(1 - p_{\mathrm{c}})\gamma^2}{\alpha_\sigma} \left(C + \frac{2 + p_{\mathrm{c}}}{p_{\mathrm{c}}} \bar{C}\right) \leq 1 - \frac{A_d}{2} \,.
\end{aligned} \qquad (100)
$$

Thus, by taking up (98) and using (100), we get

$$
\begin{aligned}
\mathbb{E}\left[V_{k+1}\right] + \alpha_d \mathbb{E}\left[d_{k+1}^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_{k+1}^2\right] \leq & \left(1 - \frac{A_d}{2}\right) \left(\mathbb{E}\left[V_k\right] + \alpha_d \mathbb{E}\left[d_k^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_k^2\right]\right) \\
& + (1 - p_{\mathrm{c}})\gamma^2 \left(D + \frac{2 + p_{\mathrm{c}}}{p_{\mathrm{c}}} \bar{D}\right) + 2\,(1 - \tau) \left(1 - \frac{1}{b}\right) \gamma d + \alpha_d D_d + \alpha_\sigma D_\sigma \,.
\end{aligned}
$$

Finally, the stated result follows by induction. □

7 Main results

Section 7 is divided into four subsections in which we prove theoretical results for the FALD and VR-FALD* algorithms. These analyses are presented in Theorem 20 and Theorem 28. The proofs are based on Lemma 11 proved in Section 6.3 to ensure that the local parameters $\{X_k^i\}_{i\in[b]}$ do not deviate too much from X_k, then we apply the general result given in Section 6 to obtain explicit upper bounds for $\mathbf{W}_2(\pi, \mu_k^{(\gamma)})$.

Until the end of the paper, we consider a family of independent random variables $(\xi^i)_{i=1}^b$ distributed according to $\nu_\xi^{\otimes b}$, and we denote $(H^i)_{i=1}^b$ a family of functions defined on $\mathbb{R}^d \times \mathsf{E} \to \mathbb{R}^d$ such that for each $i \in [b], x \in \mathbb{R}^d$, $H^i(x, \xi^i(\cdot))$ is measurable on $(\mathsf{E}, \mathcal{E})$ and satisfies the following condition:

A4. *Assume there exists $\hat{L} \geq 0$, such that for any $i \in [b], x, y \in \mathbb{R}^d$, we have*

$$
\begin{aligned}
& \mathbb{E}\left[H^i(x, \xi^i)\right] = \nabla U^i(x) \,, \\
& \mathbb{E}\left[\left\|H^i(y, \xi^i) - H^i(x, \xi^i)\right\|^2\right] \leq \hat{L}^2 \left\|y - x\right\|^2 \,.
\end{aligned}
$$

The assumption **A**4 is equivalent to **A**2 written in the main paper, though for clarity we prefer to replace the stochastic gradient $\widehat{\nabla} U_k^i$ by $H^i(\cdot, \xi^i)$. To simplify the notation, in what follows we consider the random variable $\xi = (\xi^1, \ldots, \xi^b)$, and we denote

$$
H : \begin{cases} \mathbb{R}^d \times \mathsf{E}^b \to \mathbb{R}^d \\ (x, z) \mapsto \sum_{i=1}^b H^i(x, z^i) \end{cases} \,.
$$

Thus, for each $x \in \mathbb{R}^d$, with this notation we have $H(x,\xi) = \sum_{i=1}^b H^i(x,\xi^i)$. We also introduce the averaged versions $\bar{U}, \bar{H}$ of the local potentials $\{U^i\}_{i\in[b]}$ and the stochastic gradients $\{H^i\}_{i\in[b]}$ defined by

$$\bar{U}(x) = \frac{1}{b}\sum_{i=1}^b U^i(x)\,, \qquad \bar{H}(x,z) = \frac{1}{b}\sum_{i=1}^b H^i(x,z^i)\,.$$

Remark 13. *In the mini-batch scenario without replacement, the ith client draws a mini-batch J_i subset$[N_i]$ of size $n_i = |J_i| \in [N_i]$ among N_i data and computes its stochastic gradient, which for $x \in \mathbb{R}^d$ is given by $H^i(x,\xi^i) = \sum_{j\in J_i} \nabla U^{i,j}(x)$. Using the result provided in Vono et al. (2022b, Lemma S4), we know that*

$$\begin{aligned}\mathbb{E}\left[\left\|H^i(y,\xi^i) - H^i(x,\xi^i)\right\|^2\right] &= \left\|\nabla U^i(y) - \nabla U^i(x)\right\|^2 + \mathrm{Var}\left(H^i(y,\xi^i) - H^i(x,\xi^i)\right)\\ &\le \left(1 + \frac{n_i(N_i - n_i)\max_{j=1}^{N_i} L_j^i}{N_i(N_i-1)L}\right) L^2 \left\|y - x\right\|^2\,.\end{aligned}$$

*Therefore, **A**4 is satisfied for a choice of $\hat{L} > 0$ such that*

$$\hat{L} \le L\sqrt{1 + \max_{i=1}^b \left\{n_i(N_i - n_i)[N_i(N_i-1)]^{-1}(\max_{j=1}^{N_i} L_j^i)L^{-1}\right\}}\,.$$

A5. *For $i \in [b]$, $j \in [N_i]$, assume that $U^{i,j}$ is continuously differentiable, convex and there exists $L_j^i > 0$ such that for any $x, y \in \mathbb{R}^d$,*

$$U^{i,j}(y) \le U^{i,j}(x) + \left\langle \nabla U^{i,j}(x), y - x\right\rangle + \frac{L_j^i}{2}\left\|y - x\right\|^2\,.$$

A6. *Assume there exists $\tilde{\omega} > 0$ such that for any $x \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\left\|H(x,\xi) - H(x_\star,\xi) - \nabla U(x)\right\|^2\right] \le \tilde{\omega} b^2 \left\|x - x_\star\right\|^2\,.$$

A1 combined with **A**4 implies **A**6 with $\tilde{\omega} = 2L^2 + 2\hat{L}^2$. However, this new assumption **A**6 is interesting because without stochastic gradient we obtain $\tilde{\omega} = 0$, which allows us to recover the classical Langevin bounds.

Remark 14. *Consider the same scenario as detailed in Remark 13 and define*

$$\tilde{\omega} = \left(\sum_{i=1}^b \frac{n_i(N_i - n_i)\max_{j=1}^{N_i} L_j^i}{b^2 N_i(N_i - 1)}\right) L\,.$$

Applying Vono et al. (2022b, Lemma S4) we have the following lines

$$\begin{aligned}\mathbb{E}\left[\left\|\bar{H}(x,\xi) - \bar{H}(x_\star,\xi) - \nabla\bar{U}(x)\right\|^2\right] &= \mathrm{Var}\left(\bar{H}(x,\xi) - \bar{H}(x_\star,\xi)\right)\\ &= \frac{1}{b^2}\sum_{i=1}^b \mathrm{Var}\left(H^i(x,\xi^i) - H^i(x_\star,\xi^i)\right) \le \tilde{\omega}\left\|x - x_\star\right\|^2\,.\end{aligned}$$

*Therefore, **A**6 is satisfied and in the deterministic case where all data are used to calculate the gradient, we have $\tilde{\omega} = 0$.*

To deal with variance reduction based algorithms, we consider the following assumption **A**7, which is also implied by **A**1-**A**4, however the constant ω vanishes with exact gradient computation.

A7. *Assume there exists $\omega \ge 0$ such that for any $i \in [b]$ and $x, y \in \mathbb{R}^d$,*

$$\mathbb{E}\left[\left\|H^i(x,\xi^i) - H^i(y,\xi^i) - \nabla U^i(x) + \nabla U^i(y)\right\|^2\right] \le \omega\left\|x - y\right\|^2\,.$$

Remark 15. *In the mini-batch scenario without replacement detailed in Remark 13, the use of Vono et al. (2022b, Lemma S4) implies that*

$$\mathbb{E}\left[\left\|H^i(x,\xi^i) - H^i(y,\xi^i) - \nabla U^i(x) + \nabla U^i(y)\right\|^2\right] = \mathrm{Var}\left(H^i(x,\xi^i) - H^i(y,\xi^i)\right)$$

$$\le \frac{n_i(N_i-n_i)}{N_i(N_i-1)} L \max_{j=1}^{N_i} L_j^i \|x-y\|^2 .$$

*Thus, **A**7 is satisfied by setting*

$$\omega = \max_{i=1}^{b} \left\{ \frac{n_i(N_i-n_i)}{N_i(N_i-1)} \max_{j=1}^{N_i} L_j^i \right\} L .$$

In the deterministic case, we obtain $\omega = 0$. Similarly, in the mini-batch scenario with replacement it is sufficient to set

$$\omega = \frac{N_i - n_i}{n_i} \sum_{j=1}^{N_i} \left(L_j^i\right)^2$$

*to ensure that **A**7 holds.*

7.1 Study of FALD

7.1.1 Remark on the theoretical analysis of Deng et al. (2021)

FALD has been proposed in Deng et al. (2021), the authors develop an MCMC algorithm targeting the distribution proportional to $\exp(-b^{-1}\sum_{i=1}^b U^i)$ and also establish non-asymptotic bounds. They introduce (Deng et al., 2021, Lemma B.2) the stochastic processes $\{(\bar\theta_t^i)_{t\ge0}\}_{i\in[b]}$ satisfying the Langevin stochastic differential equations for $t \ge 0$, $\mathrm{d}\bar\theta_t^i = -\nabla U^i(\bar\theta_t^i) + \sqrt{2b}\,\mathrm{dW}_t^i$ where $\{(\mathrm{W}_t^i)_{t\ge0}\}_{i\in[b]}$ are independent d-dimensional standard Brownian motion and define $\bar\theta_t = b^{-1}\sum_{i=1}^b \bar\theta_t^i$. Then, it is asserted (Deng et al., 2021, Lemma B.5) that $(\bar\theta_t)$ is solution of the Langevin stochastic differential equation $\mathrm{d}\bar\theta_t = -b^{-1}\sum_{i=1}^b \nabla U^i(\bar\theta_t) + \sqrt{2}\,\mathrm{dW}_t$, where $\mathrm{W}_t = b^{-1/2}\sum_{i=1}^b \mathrm{W}_t^i$. However, this statement cannot hold in all generalities, and we give a counter-example. For instance, consider the Gaussian potentials $\{U^i : x \in \mathbb{R}^d \mapsto \Sigma_i^{-1}(x - \mathrm{m}^i)\}_{i\in[b]}$ where $\{(\mathrm{m}^i, \Sigma_i)\}_{i\in[b]}$ are the mean and the covariance parameters; if for $i \in [b]$, $\bar\theta_0^i$ is distributed according to $\exp(-U^i)$, then $b^{-1}\sum_{i=1}^b \bar\theta_t^i$ follows $\mathbf{N}(b^{-1}\sum_{i=1}^b \mathrm{m}^i, b^{-2}\sum_{i=1}^b \Sigma_i)$ whereas $\exp(-b^{-1}\sum_{i=1}^b U^i)$ corresponds to the density of the Gaussian $\mathbf{N}(\sum_{i=1}^b(\bar\Sigma\Sigma_i^{-1})\mathrm{m}^i, b\bar\Sigma)$ where $\bar\Sigma = (\sum_{i=1}^b \Sigma_i^{-1})^{-1}$. Therefore, for any $t \ge 0$, in this case $\bar\theta_t$ is distributed according to $\mathbf{N}(b^{-1}\sum_{i=1}^b \mathrm{m}^i, b^{-2}\sum_{i=1}^b \Sigma_i)$ and thus cannot be distributed according to $\exp(-b^{-1}\sum_{i=1}^b U^i)$ as crucially used in the proof of Deng et al. (2021, Lemma B.5).

7.1.2 Theoretical analysis

In this section, we prove the first theoretical guarantee on FALD stated in Theorem 1. Similar to McMahan et al. (2017), the clients update their local parameters $\{X_k^i\}_{i\in[b]}$ several times before transmitting them to the server with probability $p_{\mathrm{c}} \in (0,1]$. Then, the server aggregates the local parameters to update its own parameter X_k as in (23). For all $i \in [b]$, $k \in \mathbb{N}$, consider the stochastic gradients defined by

$$G_k^i = H^i(X_k^i, \xi_{k+1}^i) , \tag{101}$$
$$\bar G_k^i = \nabla U^i(X_k^i) . \tag{102}$$

Lemma 16. *Assume **A**1, **A**4 and **A**6 hold. Then for any $k \in \mathbb{N}$, we have*

$$\frac{1}{b}\sum_{i=1}^b \mathbb{E}\left[\|\bar G_k^i\|^2\right] \le 3L^2\mathbb{E}\left[V_k\right] + 3L^2\mathbb{E}\left[d_k^2\right] + \frac{3}{b}\sum_{i=1}^b \|\nabla U^i(x_\star)\|^2 ,$$
$$\frac{1}{b}\sum_{i=1}^b \mathbb{E}\left[\|G_k^i - \bar G_k^i\|^2\right] \le 3\hat L^2\mathbb{E}\left[V_k\right] + 3\tilde\omega\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\left\|\bar H(x_\star, \xi)\right\|^2\right] .$$

For any $i \in [b]$, $k \in \mathbb{N}$, recall the stochastic gradients $G_k^i, \bar G_k^i$ are defined in (101) and (102), respectively

Algorithm 2 Stochastic Averaging Langevin Dynamics - FALD

```
Input: initial vectors $(X_0^i)_{i\in[b]}$, noise parameter $\tau \in [0,1]$, number of communication rounds $K$, probability $p_c$ of communication, step-size $\gamma$.
for $k = 0$ to $K-1$ do
    // On each client
    Draw $B_{k+1} \sim \mathcal{B}(p_c)$, $\tilde{Z}_{k+1} \sim \mathbf{N}(0_d, \mathrm{I}_d)$
    // In parallel on the $b$ clients
    for $i = 1$ to $b$ do
        Draw $\xi_{k+1}^i \sim \nu_\xi$ and $\tilde{Z}_{k+1}^i \sim \mathbf{N}(0_d, \mathrm{I}_d)$
        Compute $G_k^i = H^i(X_k^i, \xi_{k+1}^i)$
        Set $\tilde{X}_{k+1}^i = X_k^i - \gamma G_k^i + \sqrt{2\gamma}\,(\sqrt{\tau/b}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i)$
        if $B_{k+1} = 1$ then
            Broadcast $\tilde{X}_{k+1}^i$ to the server
        else
            Update $X_{k+1}^i \leftarrow \tilde{X}_{k+1}^i$
    if $B_{k+1} = 1$ then
        // On the central server
        Update then broadcast the global parameter $X_{k+1} = \frac{1}{b}\sum_{i=1}^b \tilde{X}_{k+1}^i$
        // On each client
        Update the local parameter $X_{k+1}^i \leftarrow X_{k+1}$
Output: samples $\{X_\ell\}_{\{\ell\in[K]\,:\,B_\ell=1\}}$.
```

Proof. Using the Young inequality combined with the Lipschitz property **A**1 of the gradients $(U^i)_i^b$, for $k \geq 0$ we get

$$\frac{1}{b}\sum_{i=1}^b \mathbb{E}\left[\|\bar{G}_k^i\|^2\right] = \frac{1}{b}\sum_{i=1}^b \mathbb{E}\left[\|\nabla U^i(X_k^i) - \nabla U^i(X_k) + \nabla U^i(X_k) - \nabla U^i(x_\star) + \nabla U^i(x_\star)\|^2\right]$$

$$\leq 3L^2\mathbb{E}\left[V_k\right] + 3L^2\mathbb{E}\left[d_k^2\right] + \frac{3}{b}\sum_{i=1}^b \|\nabla U^i(x_\star)\|^2 \,.$$

In addition, since the random variables $(G_k^i - \bar{G}_k^i)_{i=1}^b$ are centered and independent, the Young and the Jensen inequality imply that

$$\begin{aligned}
\frac{1}{b}\sum_{i=1}^b \mathbb{E}\left[\left\|G_k^i - \bar{G}_k^i\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^b \left(G_k^i - \bar{G}_k^i\right)\right\|^2\right] \\
&= \mathbb{E}\Bigg[\Bigg\|\frac{1}{b}\sum_{i=1}^b H^i(X_k^i, \xi_{k+1}^i) - \bar{H}(X_k, \xi_{k+1}) + \bar{H}(X_k, \xi_{k+1}) - \bar{H}(x_\star, \xi_{k+1}) \\
&\qquad + \bar{H}(x_\star, \xi_{k+1}) - \nabla\bar{U}(X_k) + \nabla\bar{U}(X_k) - \frac{1}{b}\sum_{i=1}^b \nabla U^i(X_k^i)\Bigg\|^2\Bigg] \\
&\leq 3\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^b H^i(X_k^i, \xi_{k+1}^i) - \bar{H}(X_k, \xi_{k+1})\right\|^2\right] \\
&\qquad + 3\mathbb{E}\left[\left\|\bar{H}(X_k, \xi_{k+1}) - \nabla\bar{U}(X_k) - \bar{H}(x_\star, \xi_{k+1})\right\|^2\right] + 3\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] \\
&\leq 3\hat{L}^2\mathbb{E}\left[V_k\right] + 3\tilde{\omega}\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] \,.
\end{aligned}$$

□

Lemma 17. *Assume* **A**1 *and* **A**4 *hold. Then, for any* $\gamma \in (0, m(6\hat{L}^2)^{-1}]$, *we have*

$$\mathbb{E}\left[d_{k+1}^2\right] \leq \left(1 - \frac{\gamma m}{2}\right)\mathbb{E}\left[d_k^2\right] + \frac{2\gamma L^2}{m}\mathbb{E}\left[V_k\right] + 3\gamma^2\mathbb{E}\left[\|\bar{H}(x_\star, \xi)\|^2\right] + \frac{2\gamma d}{b} \,,$$

where V_k, d_k are defined in (24) and (25).

Proof. Let k be in $\mathbb{N}$. Rewriting the expression of X_{k+1} defined in (23), we obtain

$$\begin{aligned}
\mathbb{E}\left[d_{k+1}^2\right] &= \mathbb{E}\left[\|X_{k+1}-x_\star\|^2\right] \\
&= \mathbb{E}\left[\left\|X_k - x_\star - \frac{\gamma}{b}\sum_{i=1}^{b} H^i(X_k^i,\xi_{k+1}^i) + \sqrt{2\gamma}\left(\sqrt{\frac{\tau}{b}}\,\tilde{Z}_{k+1} + \frac{\sqrt{1-\tau}}{b}\sum_{i=1}^{b} Z_{k+1}^i\right)\right\|^2\right] \\
&= \mathbb{E}\left[\|X_k - x_\star\|^2\right] - 2\gamma\mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{b}\sum_{i=1}^{b} H^i(X_k^i,\xi_{k+1}^i)\right\rangle\right] \\
&\quad + \gamma^2\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b} H^i(X_k^i,\xi_{k+1}^i)\right\|^2\right] + \frac{2\gamma d}{b}\,. \qquad (103)
\end{aligned}$$

Further, the Young inequality combined with **A**4 give

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b} H^i(X_k^i,\xi_{k+1}^i)\right\|^2\right] &\le \frac{3}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|H^i(X_k^i,\xi_{k+1}^i) - H^i(X_k,\xi_{k+1}^i)\right\|^2\right] + 3\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right] \\
&\quad + 3\mathbb{E}\left[\left\|\bar{H}(X_k,\xi_{k+1}) - \bar{H}(x_\star,\xi)\right\|^2\right] \\
&\le 3\hat{L}^2\mathbb{E}\left[V_k\right] + 3\hat{L}^2\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right]\,. \qquad (104)
\end{aligned}$$

In addition, using the fact that for any vectors $a, b \in \mathbb{R}^d$, $2\left|\langle a, b\rangle\right| \le m\left\|a\right\|^2 + \left\|b\right\|^2/m$ we can upper bound the inner product derived in (103) as follows

$$\begin{aligned}
-\mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{b}\sum_{i=1}^{b} H^i(X_k^i,\xi_{k+1}^i)\right\rangle\right] &= -\mathbb{E}\left[\left\langle X_k - x_\star, \nabla\bar{U}(X_k)\right\rangle\right] \\
&\quad + \mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{b}\sum_{i=1}^{b}\left[H^i(X_k,\xi_{k+1}^i) - H^i(X_k^i,\xi_{k+1}^i)\right]\right\rangle\right] \\
&\le -\mathbb{E}\left[\left\langle X_k - x_\star, \nabla\bar{U}(X_k)\right\rangle\right] + m\mathbb{E}\left[d_k^2\right]/2 + L^2\mathbb{E}\left[V_k\right]/(2m) \\
&\le -m\mathbb{E}\left[d_k^2\right]/2 + L^2\mathbb{E}\left[V_k\right]/(2m)\,. \qquad (105)
\end{aligned}$$

Therefore, plugging (104) and (105) in (103) shows

$$\mathbb{E}\left[d_{k+1}^2\right] \le \left(1 - \gamma\left[m - 3\gamma\hat{L}^2\right]\right)\mathbb{E}\left[d_k^2\right] + \gamma\left(3\gamma\hat{L}^2 + \frac{L^2}{m}\right)\mathbb{E}\left[V_k\right] + 3\gamma^2\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right] + \frac{2\gamma d}{b}\,.$$

Eventually, the assumption $\gamma \le m(6\hat{L}^2)^{-1}$ completes the proof. □

For any $\gamma \in (0, m(6\hat{L}^2)^{-1}]$, under **A**1, **A**4 and **A**6 using Lemma 16 and Lemma 17 we have shown that **HX**3 and **HX**4 hold with the following quantities

$$\begin{array}{llll}
A = 3\hat{L}^2\,, & B = 3\tilde{\omega}\,, & C = 0\,, & D = 3\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right]\,, \\
\bar{A} = 3L^2\,, & \bar{B} = 3L^2\,, & \bar{C} = 0\,, & \bar{D} = (3/b)\sum_{i=1}^{b}\left\|\nabla U^i(x_\star)\right\|^2\,, \\
A_d = \gamma m/2\,, & B_d = 0\,, & C_d = 2\gamma L^2/m\,, & D_d = 3\gamma^2\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right] + 2\gamma d/b\,, \\
A_\sigma = 1\,, & B_\sigma = 0\,, & C_\sigma = 0\,, & D_\sigma = 0\,.
\end{array} \qquad (106)$$

For any $\gamma > 0$, consider the following variables

$$\begin{aligned}
&\mathrm{C}^\gamma = \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}} - 4A_d}\left(B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B}\right)\,, \quad \mathrm{C}_r^\gamma = 3\mathrm{C}^\gamma C_d\,, \quad \mathrm{C}_V^\gamma = 1 + 2C_d\mathrm{C}^\gamma\,, \\
&\mathrm{C}_\epsilon^\gamma = \mathrm{C}_V^\gamma\mathbb{E}\left[V_0\right] + 7\mathrm{C}^\gamma\mathbb{E}\left[d_0^2\right] + 2D_d\,, \quad \mathrm{C}_\delta^\gamma = \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}}}\left(D + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{D}\right) + \frac{\mathrm{C}^\gamma D_d}{A_d} + \frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{b p_{\mathrm{c}}}\,.
\end{aligned} \qquad (107)$$

We also introduce γ_1 and I_γ, which are defined for any $\gamma > 0$ by

$$\gamma_1 = \frac{p_{\mathrm{c}}^{1/2}}{(2-2p_{\mathrm{c}})^{1/2}\left[A+(1+2/p_{\mathrm{c}})\bar{A}\right]^{1/2}} \wedge \frac{m}{6\hat{L}^2} \wedge \frac{p_{\mathrm{c}}}{2m} \wedge \frac{q_{\mathrm{c}}}{m}\,,$$
$$I_\gamma = \{\gamma \in (0,\gamma_1)\,:\, \gamma m \geq 8\mathrm{C}_r^\gamma\}\,.$$

Based on Lemma 11, we derive the following result.

Lemma 18. *Assume* **A**1*,* **A**4 *and* **A**6 *hold. Then, for any $\gamma \in I_\gamma$ and $k \geq 1$, we have*

$$\mathbb{E}\left[V_k\right] \leq \left(1-\frac{A_d}{4}\right)^k \left(2\mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_\delta^\gamma \mathrm{C}_r^\gamma}{A_d}\right) + \mathrm{C}_\delta^\gamma\,.$$

where V_k is defined in (24) *and $\mathrm{C}_\epsilon^\gamma, \mathrm{C}_r^\gamma, \mathrm{C}_\delta^\gamma$ in* (107)*.*

Proof. For any $\gamma \in I_\gamma$, we have $4\mathrm{C}_r^\gamma \leq A_d$ and moreover it is easy to check that $A_d < \min(A_\sigma/2, p_{\mathrm{c}}/4)$, $A_d A_\sigma \geq 8B_d B_\sigma = 0$. In addition, since **A**1, **A**4 and **A**6 are satisfied we can apply Lemma 16 and Lemma 17 which show that **HX**3, **HX**4 hold with the variables introduced in (106). Therefore, we can use Lemma 11 to complete the proof. □

Based on the results presented in this section, we can rewrite the upper bound on $(\mathbb{E}\left[V_k\right])_{k\in\mathbb{N}}$ given in Lemma 18 into the format of **H**2. We consider for $\gamma > 0$,

$$v_1 = 2\mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_\delta^\gamma \mathrm{C}_r^\gamma}{A_d}\,, \qquad\qquad v_2 = \mathrm{C}_\delta^\gamma\,. \tag{108}$$

Lemma 19. *Assume* **A**1*,* **H**1*,* **A**4 *hold and let $\gamma \leq 2(3L)^{-1}$. Then for any $k \in \mathbb{N}$, we have*

$$\begin{aligned}\mathbb{E}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] \leq \left[1-\gamma m\left(1-3\gamma L\right)+3\gamma^2\hat{L}^2\right]\mathbb{E}\left[\|\mathsf{X}_{k\gamma}-X_k\|^2\right] + \gamma\left(\frac{2L^2}{m}+3\gamma(L^2+\hat{L}^2)\right)\mathbb{E}\left[V_k\right] \\ + \left(\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right]+3\mathbb{E}\left[\|I_k\|^2\right]\right) + \frac{3\gamma^2}{b^2}\int_{\mathbb{R}^d}\mathrm{Var}^{\mathcal{F}_0}\left(H(x,\xi)\right)\pi(\mathrm{d}x)\,.\end{aligned}$$

Proof. For any $k \in \mathbb{N}$, recall that $\mathcal{F}_k$ is defined in (22) and using Proposition 4 we obtain

$$\begin{aligned}\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma}-X_{k+1}\|^2\right] \leq \left[1-\gamma m\left(1-3\gamma L\right)\right]\|\mathsf{X}_{k\gamma}-X_k\|^2 + \gamma\left(\frac{2L^2}{m}+3\gamma L^2\right)V_k \\ + \left(\frac{2}{\gamma m}\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2+3\mathbb{E}^{\mathcal{F}_k}\left[\|I_k\|^2\right]\right) + \gamma^2\,\mathrm{Var}^{\mathcal{F}_k}\left(\frac{1}{b}\sum_{i=1}^b G_k^i\right)\,.\end{aligned} \tag{109}$$

Since the stochastic gradients $(H^i(\cdot,\xi_{k+1}^i))_{i=1}^b$ are unbiased, **A**4 with the Young inequality imply that

$$\begin{aligned}&\mathrm{Var}^{\mathcal{F}_k}\left(\frac{1}{b}\sum_{i=1}^b G_k^i\right) = \mathbb{E}^{\mathcal{F}_k}\left[\left\|\frac{1}{b}\sum_{i=1}^b\left[H^i(X_k^i,\xi_{k+1}^i)-\nabla U^i(X_k^i)\right]\right\|^2\right] \\ &= \mathbb{E}^{\mathcal{F}_k}\left[\left\|\frac{1}{b}\sum_{i=1}^b H^i(X_k^i,\xi_{k+1}^i) - \bar{H}(X_k,\xi_{k+1}) - \frac{1}{b}\sum_{i=1}^b\nabla U^i(X_k^i) + \nabla\bar{U}(X_k)\right.\right. \\ &\left.\left. + \bar{H}(X_k,\xi_{k+1}) - \bar{H}(\mathsf{X}_{k\gamma},\xi_{k+1}) - \nabla\bar{U}(X_k) + \nabla\bar{U}(\mathsf{X}_{k\gamma}) + \bar{H}(\mathsf{X}_{k\gamma}^i,\xi_{k+1}) - \nabla\bar{U}(\mathsf{X}_{k\gamma})\right\|^2\right] \\ &\leq 3\hat{L}^2 V_k + 3\hat{L}^2\|X_k-\mathsf{X}_{k\gamma}\|^2 + 3\,\mathrm{Var}^{\mathcal{F}_k}\left(\bar{H}(\mathsf{X}_{k\gamma},\xi_{k+1})\right)\,.\end{aligned}$$

Taking the expectation and using that $\mathsf{X}_{k\gamma}$ has distribution π combined with (109) complete the proof. □

For notational convenience, we also introduce the time step-size γ_2 defined by

$$\gamma_2 = \frac{p_{\mathrm{c}}}{4m} \wedge \frac{1}{6(L+\hat{L}^2/m)} \wedge \frac{p_{\mathrm{c}} m}{38(1-p_{\mathrm{c}})^{1/2}\,(p_{\mathrm{c}}\tilde{\omega}+3L^2)^{1/2}\,L}\,.$$

Theorem 20. *Assume* **A**1*,* **A**4 *and* **A**6 *hold and let* $\gamma \in (0, \gamma_1 \wedge \gamma_2)$*. Then, for any initial probability measure* $\mu_0^{(\mathrm{F})} \in \mathcal{P}_2(\mathbb{R}^d)$*,* $k \in \mathbb{N}$*, we have*

$$\mathbf{W}_2^2\left(\mu_k^{(\mathrm{F})}, \pi\right) \leq \left(1 - \frac{\gamma m}{2}\right)^k \mathbf{W}_2^2\left(\mu_0^{(\mathrm{F})}, \pi\right) + \frac{8L^2}{m^2} v_1 \left(1 - \frac{\gamma m}{8}\right)^k + \frac{6L^2}{m^2} v_2 + \frac{6\gamma d}{bm^2}\kappa_I + \frac{6\gamma}{b^2 m}\int_{\mathbb{R}^d} \mathrm{Var}^{\mathcal{F}_0}\left(H\left(x, \xi_1\right)\right)\pi(\mathrm{d}x)\,.$$

where v_1*,* v_2 *are defined in* (108) *and* $\kappa_I = L^2(1+\gamma L^2/m)$*. If in addition we suppose* **HX**1*, set* $\kappa_I = 2\gamma(L^3 + d\tilde{L}^2/b)$*.*

Proof. We know that **H**1 is satisfied since for any $i \in [b], x \in \mathbb{R}^d$ the stochastic gradient $H^i(x, \xi_1^i)$ is unbiased. The constraint $\gamma \leq \gamma_1$ combined with Lemma 17 implies **HX**3 and plugging the expression of $A_d, A_\sigma, B_d, C, \bar{C}, C_d, C_\sigma$ provided in (106) into C_r^γ defined in (107) gives that

$$\mathrm{C}_r^\gamma = \frac{72\gamma^3(1-p_{\mathrm{c}})L^2\left(\tilde{\omega} + (1+2/p_{\mathrm{c}})\,L^2\right)}{(p_{\mathrm{c}} - 2\gamma m)m}\,.$$

For any $\gamma \in (0, \gamma_2]$, we have $(p_{\mathrm{c}} - 2\gamma m)m^2 \geq 576(1-p_{\mathrm{c}})\gamma^2 L^2\left(\tilde{\omega} + (1+2/p_{\mathrm{c}})\,L^2\right)$ which shows that $\gamma \in I_\gamma$. Thus, we can apply Lemma 18 which proves that **H**2 holds with $q_{\mathrm{c}} = \gamma m$ and $\alpha_v = 1 - A_d/4$ and v_1, v_2 defined in (108). Since the assumptions of Lemma 19 are satisfied, **HX**2 holds, and therefore we can apply Theorem 8 with

$$(1-q_{\mathrm{c}})\alpha_0 = 1 - \gamma m\,(1 - 3\gamma L) + 3\gamma^2\hat{L}^2\,, \quad \alpha_1 = 0\,, \quad (1-q_{\mathrm{c}})\alpha_2 = \gamma\left(\frac{2L^2}{m} + 3\gamma(L^2 + \hat{L}^2)\right)\,, \quad \alpha_3 = 0\,,$$

$$(1-q_{\mathrm{c}})\alpha_4 = \left(\frac{2}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right]\right) + \frac{3\gamma^2}{b^2}\int_{\mathbb{R}^d}\mathrm{Var}^{\mathcal{F}_0}\left(H(x,\xi_1)\right)\pi(\mathrm{d}x)\,.$$

Furthermore, using Lemma 7 we have

$$\frac{2}{\gamma m}\mathbb{E}\left[\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\|^2\right] + 3\mathbb{E}\left[\|I_k\|^2\right] \leq \frac{3\gamma^2 dL^2}{bm}\left(1 + \frac{19\gamma L^2}{36m}\right)\,. \tag{110}$$

Moreover, if we suppose **HX**1, we obtain

$$\frac{2}{\gamma m}\mathbb{E}\left[\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\|^2\right] + 3\mathbb{E}\left[\|I_k\|^2\right] \leq \frac{\gamma^3 d}{bm}\left(5L^3 + \frac{4d\tilde{L}^2}{3b}\right)\,. \tag{111}$$

Finally, with the notation of Theorem 8 we obtain $1 + \delta = 0$, and using $\gamma \leq (6(L + m^{-1}\hat{L}^2))^{-1}$ combined with (110) or (111) if we suppose **HX**1 give the expected result. □

Now, consider the time stepsizes γ_3 and $\gamma_\star$ defined by

$$\gamma_3 = \frac{p_{\mathrm{c}} m}{3L^2 + p_{\mathrm{c}}\tilde{\omega}}\,, \qquad\qquad \gamma_\star = \gamma_1 \wedge \gamma_2 \wedge \gamma_3\,.$$

From the previous result, the next corollary controls the asymptotic bias obtained by Algorithm 2.

Corollary 21. *Assume* **A**1*,* **A**4 *and* **A**6 *hold and let* $\gamma \in (0, \gamma_\star)$*,* $\tau = 1$*. Then, for any initial probability measure* $\mu_0^{(\mathrm{F})} \in \mathcal{P}_2(\mathbb{R}^d)$*,* $k \in \mathbb{N}$*, we have*

$$\frac{6^{-4}b}{\gamma d}\limsup_{k\to\infty}\mathbf{W}_2^2\left(\mu_k^{(\mathrm{F})}, \pi\right) \leq \frac{\int_{\mathbb{R}^d}\mathrm{Var}^{\mathcal{F}_0}\left(H\left(x,\xi_1\right)\right)\pi(\mathrm{d}x)}{bdm} + \frac{\tilde{\kappa}_I}{m^2} + \frac{(1-p_{\mathrm{c}})\gamma L^2}{p_{\mathrm{c}}^2 m^2}\left(\frac{1}{d}\sum_{i=1}^b\left\|\nabla U^i(x_\star)\right\|^2 + \frac{p_{\mathrm{c}}}{bd}\mathbb{E}\left[\left\|H(x_\star,\xi)\right\|^2\right] + \frac{L^2 + p_{\mathrm{c}}\tilde{\omega}}{m}\right)\,.$$

where $\tilde{\kappa}_I = L^2$ *and if we suppose* **HX**1*,* $\tilde{\kappa}_I = \gamma(L^3 + d\tilde{L}^2/b)$*.*

Proof. Using Theorem 20 combined with $\gamma \leq \gamma_1 \wedge \gamma_2$ gives that

$$\limsup_{k\to\infty} \mathbf{W}_2^2\left(\mu_k^{(\mathrm{F})}, \pi\right) \leq \frac{6\gamma}{b^2 m}\int_{\mathbb{R}^d} \mathrm{Var}^{\mathcal{F}_0}\left(H\left(x, \xi_1\right)\right)\pi(\mathrm{d}x) + \frac{6\gamma d}{bm^2}\kappa_I + \frac{6L^2}{m^2}\upsilon_2\,. \tag{112}$$

Further, recall that $A_d, B, \bar{B}, D, \bar{D}, D_d$ are provided in (106) and C_δ^γ is defined in (107) by

$$\begin{aligned}
\mathrm{C}_\delta^\gamma &= \frac{4(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}}\left(D + \frac{2+p_\mathrm{c}}{p_\mathrm{c}}\bar{D}\right) + \frac{\mathrm{C}^\gamma D_d}{A_d} + \frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{bp_\mathrm{c}}\\
&\leq \frac{12(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}}\left[1 + \frac{12\gamma}{m}\left(\tilde{\omega} + \frac{3}{p_\mathrm{c}}L^2\right)\right]\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] + \frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{bp_\mathrm{c}}\\
&\quad + \frac{36(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}^2 b}\sum_{i=1}^b \left\|\nabla U^i(x_\star)\right\|^2 + \frac{96(1-p_\mathrm{c})\gamma^2 d}{p_\mathrm{c} bm}\left(\tilde{\omega} + \frac{3}{p_\mathrm{c}}L^2\right)\\
&\leq \frac{156(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}}\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] + \frac{36(1-p_\mathrm{c})\gamma^2}{p_\mathrm{c}^2 b}\sum_{i=1}^b \left\|\nabla U^i(x_\star)\right\|^2\\
&\quad + \frac{96(1-p_\mathrm{c})\gamma^2 d}{p_\mathrm{c} bm}\left(\tilde{\omega} + \frac{3}{p_\mathrm{c}}L^2\right) + \frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{bp_\mathrm{c}}\,.
\end{aligned} \tag{113}$$

Finally, setting $\tau = 1$ combined with (112) and (113) show that

$$\begin{aligned}
\limsup_{k\to\infty} \mathbf{W}_2^2\left(\mu_k^{(\mathrm{F})}, \pi\right) &\leq \frac{6\gamma}{b^2 m}\int_{\mathbb{R}^d} \mathrm{Var}^{\mathcal{F}_0}\left(H\left(x, \xi_1\right)\right)\pi(\mathrm{d}x) + \frac{6\gamma d}{bm^2}\kappa_I\\
&\quad + \frac{8(1-p_\mathrm{c})\gamma^2 L^2}{bp_\mathrm{c} m^2}\left[\frac{156}{b}\mathbb{E}\left[\|H(x_\star, \xi)\|^2\right] + \frac{36}{p_\mathrm{c}}\sum_{i=1}^b \left\|\nabla U^i(x_\star)\right\|^2 + \frac{96d}{m}\left(\tilde{\omega} + \frac{3}{p_\mathrm{c}}L^2\right)\right].
\end{aligned}$$

□

7.2 Study of VR-FALD*

In this alternative of FALD derived in Section 7.1, we introduce control variates to cope with both heterogeneity and variance in local gradients. Instead of using $H^i(X_k^i)$ to update the local parameter X_k^i, this time the ith client uses the proxy $H^i(X_k^i, \xi_{k+1}^i) - H^i(Y_k, \xi_{k+1}^i) + \nabla U^i(Y_k)$ based on an analog of the SVRG algorithm (Johnson and Zhang, 2013; Karimireddy et al., 2020) and where Y_k is a global reference point updated with probability $q_\mathrm{c} \in (0, 1]$. We derive an explicit upper bound on the Wasserstein distance between the distribution of the server parameter $\mathsf{X}_{k\gamma}$ and the target distribution π. We also show how this new global control variate mitigates the effect of heterogeneity in the convergence rate. To do so, we consider the stochastic gradients defined for any $i \in [b], k \in \mathbb{N}$, by

$$G_k^i = H^i(X_k^i, \xi_{k+1}^i) - H^i(Y_k, \xi_{k+1}^i) + C_k\,, \tag{114}$$

$$\bar{G}_k^i = \nabla U^i(X_k^i) - \nabla U^i(Y_k) + C_k \tag{115}$$

and denote

$$\sigma_k = \left(\frac{1}{b}\sum_{i=1}^b \mathbb{E}^{\mathcal{F}_k}\left[\left\|H^i(Y_k, \xi_{k+1}^i) - H^i(x_\star, \xi_{k+1}^i)\right\|^2\right]\right)^{1/2}\,. \tag{116}$$

Lemma 22. *Assume* **A***1,* **A***4 and* **A***6 hold. Then for any $k \in \mathbb{N}$, we have*

$$\frac{1}{b}\sum_{i=1}^b \mathbb{E}\left[\|\bar{G}_k^i\|^2\right] \leq 3L^2\mathbb{E}\left[V_k\right] + 3L^2\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\sigma_k^2\right]\,,$$

$$\frac{1}{b}\sum_{i=1}^b \mathbb{E}\left[\|G_k^i - \bar{G}_k^i\|^2\right] \leq 3\hat{L}^2\mathbb{E}\left[V_k\right] + 3\tilde{\omega}\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\sigma_k^2\right]\,.$$

For any $i \in [b], k \in \mathbb{N}$, recall the stochastic gradients $G_k^i, \bar{G}_k^i$ are defined in (114) and (115), respectively

Algorithm 3 VR-FALD*

```
Input: initial vectors $(X_0^i)_{i\in[b]}$, noise parameter $\tau \in [0,1]$, number of communication rounds $K$, probability $p_{\mathrm{c}}$ of communication, probability $q_{\mathrm{c}}$ to update the control variates, step-size $\gamma$ and batch size $r$.
Initialize $Y_0 = (1/b)\sum_{i=1}^b X_0^i$ and $C_0 = (1/b)\nabla U(Y_0)$
for $k = 0$ to $K-1$ do
    // On each client
    Draw $B_{k+1} \sim \mathcal{B}(p_{\mathrm{c}})$, $\tilde{Z}_{k+1} \sim \mathbf{N}(0_d, \mathrm{I}_d)$
    // In parallel on the $b$ clients
    for $i = 1$ to $b$ do
        Draw $\xi_{k+1}^i \sim \nu_\xi$, $\tilde{Z}_{k+1}^i \sim \mathbf{N}(0_d, \mathrm{I}_d)$
        Compute $G_k^i = H^i(X_k^i, \xi_{k+1}^i) - H^i(Y_k, \xi_{k+1}^i) + C_k$
        Set $\tilde{X}_{k+1}^i = X_k^i - \gamma G_k^i + \sqrt{2\gamma}\,(\sqrt{\tau/b}\,\tilde{Z}_{k+1} + \sqrt{1-\tau}\,\tilde{Z}_{k+1}^i)$
        if $B_{k+1} = 1$ then
            Broadcast $\tilde{X}_{k+1}^i$ to the server
        else
            Update $X_{k+1}^i \leftarrow \tilde{X}_{k+1}^i$
        if $\tilde{B}_{k+1} = 1$ then
            Broadcast $X_k^i$ to the server
        else
            Update $Y_{k+1} \leftarrow Y_k$ and $C_{k+1} \leftarrow C_k$
    if $B_{k+1} = 1$ then
        // On the central server
        Update then broadcast the global parameter $X_{k+1} \leftarrow (1/b)\sum_{i=1}^b \tilde{X}_{k+1}^i$
        // On each client
        Update the local parameter $X_{k+1}^i \leftarrow X_{k+1}$
    if $\tilde{B}_{k+1} = 1$ then
        // On the central server
        Update then broadcast $Y_{k+1} \leftarrow (1/b)\sum_{i=1}^b X_k^i$
        // On each client
        Compute and broadcast $\nabla U^i(Y_{k+1})$
        // On the central server
        Update then broadcast $C_{k+1} \leftarrow (1/b)\nabla U(Y_{k+1})$
Output: samples $\{X_\ell\}_{\{\ell\in[K]\,:\,B_\ell=1\}}$.
```

Proof. For $k \geq 0$, Lipschitz property of $\{\nabla U^i\}_{i\in[b]}$ supposed in **A**1 gives that

$$
\begin{aligned}
\frac{1}{b}\sum_{i=1}^{b}\mathbb{E}\left[\|\bar{G}_k^i\|^2\right] &= \frac{1}{b}\sum_{i=1}^{b}\mathbb{E}\left[\|\nabla U^i(X_k^i) - \nabla U^i(Y_k) + \nabla \bar{U}(Y_k)\|^2\right] \\
&\leq \frac{3}{b}\sum_{i=1}^{b}\mathbb{E}\left[\|\nabla U^i(X_k^i) - \nabla U^i(X_k)\|^2\right] + \frac{3}{b}\sum_{i=1}^{b}\mathbb{E}\left[\|\nabla U^i(Y_k) - \nabla U^i(x_\star)\|^2\right] \\
&\quad + \frac{3}{b}\sum_{i=1}^{b}\mathbb{E}\left[\|\nabla U^i(X_k) - \nabla U^i(x_\star)\|^2\right] \\
&\leq 3L^2\mathbb{E}\left[V_k\right] + 3L^2\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\sigma_k^2\right]
\end{aligned}
$$

and the proof is concluded by noting that **A**4 gives

$$
\begin{aligned}
\frac{1}{b}\sum_{i=1}^{b}\mathbb{E}\|G_k^i - \bar{G}_k^i\|^2 &= \mathbb{E}\left[\operatorname{Var}^{\mathcal{F}_k}\left(\frac{1}{b}\sum_{i=1}^{b}G_k^i\right)\right] \\
&\leq \mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}H^i(X_k^i, \xi_{k+1}^i) - \bar{H}(X_k, \xi_{k+1})\right\|^2\right] \\
&\leq 3\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}H^i(X_k^i, \xi_{k+1}^i) - \bar{H}(X_k, \xi_{k+1})\right\|^2\right] + 3\mathbb{E}\left[\left\|\bar{H}(Y_k, \xi_{k+1}) - \bar{H}(x_\star, \xi_{k+1})\right\|^2\right] \\
&\quad + 3\mathbb{E}\left[\left\|\bar{H}(X_k, \xi_{k+1}) - \bar{H}(x_\star, \xi_{k+1}) - \nabla\bar{U}(X_k)\right\|^2\right] .
\end{aligned}
$$

□

Lemma 23. *Assume* **A**1 *and* **A**4 *hold. Then, for any* $\gamma \in (0, m(6\hat{L}^2)^{-1}]$, *we have*

$$
\mathbb{E}\left[d_{k+1}^2\right] \leq \left(1 - \frac{\gamma m}{2}\right)\mathbb{E}\left[d_k^2\right] + \frac{2\gamma L^2}{m}\mathbb{E}\left[V_k\right] + 4\gamma^2\mathbb{E}\left[\sigma_k^2\right] + 10\gamma^2\mathbb{E}\left[\|\bar{H}(x_\star, \xi)\|^2\right] + \frac{2\gamma d}{b} ,
$$

where V_k, d_k, σ_k *are defined in* (24), (25) *and* (116).

Proof. Let k be in $\mathbb{N}$. Writing the expression of X_{k+1} defined in (23) and developing the expectation of the squared norm give

$$
\begin{aligned}
&\mathbb{E}\left[d_{k+1}^2\right] = \mathbb{E}\left[\|X_{k+1} - x_\star\|^2\right] \\
&= \mathbb{E}\left[\left\|X_k - x_\star - \frac{\gamma}{b}\sum_{i=1}^{b}H^i(X_k^i, \xi_{k+1}^i) + \gamma\bar{H}(Y_k, \xi_{k+1}) - \gamma\nabla\bar{U}(Y_k) + \sqrt{2\gamma}\left(\sqrt{\frac{\tau}{b}}\,\tilde{Z}_{k+1} + \frac{\sqrt{1-\tau}}{b}\sum_{i=1}^{b}Z_{k+1}^i\right)\right\|^2\right] \\
&= \mathbb{E}\left[\|X_k - x_\star\|^2\right] - 2\gamma\mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{b}\sum_{i=1}^{b}H^i(X_k^i, \xi_{k+1}^i)\right\rangle\right] + \gamma^2\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}H^i(X_k^i, \xi_{k+1}^i)\right\|^2\right] \\
&\quad - 2\gamma^2\mathbb{E}\left[\left\langle\frac{1}{b}\sum_{i=1}H^i(X_k^i, \xi_{k+1}^i), \bar{H}(Y_k, \xi_{k+1}) - \gamma\nabla\bar{U}(Y_k)\right\rangle\right] + \gamma^2\mathbb{E}\left[\left\|\bar{H}(Y_k, \xi_{k+1}) - \nabla\bar{U}(Y_k)\right\|^2\right] + \frac{2\gamma d}{b} \\
&= \mathbb{E}\left[d_k^2\right] - 2\gamma\mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{b}\sum_{i=1}^{b}\nabla U^i(X_k^i)\right\rangle\right] + 2\gamma^2\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}H^i(X_k^i, \xi_{k+1}^i)\right\|^2\right] \\
&\quad + 2\gamma^2\mathbb{E}\left[\left\|\bar{H}(Y_k, \xi_{k+1}) - \nabla\bar{U}(Y_k)\right\|^2\right] + \frac{2\gamma d}{b} .
\end{aligned} \tag{117}
$$

Using the Young inequality combined with **A**4 show

$$
\mathbb{E}\left[\left\|\frac{1}{b}\sum_{i=1}^{b}H^i(X_k^i, \xi_{k+1}^i)\right\|^2\right] \leq \frac{3}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|H^i(X_k^i, \xi_{k+1}^i) - H^i(X_k, \xi_{k+1}^i)\right\|^2\right]
$$

$$
\begin{aligned}
&+ 3\mathbb{E}\left[\left\|\bar{H}(X_k,\xi_{k+1}) - \bar{H}(x_\star,\xi)\right\|^2\right] + 3\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right] \\
&\leq 3\hat{L}^2\mathbb{E}\left[V_k\right] + 3\hat{L}^2\mathbb{E}\left[d_k^2\right] + 3\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right] \,. \qquad (118)
\end{aligned}
$$

We also have that

$$
\begin{aligned}
\mathbb{E}\left[\left\|\bar{H}(Y_k,\xi_{k+1}) - \nabla\bar{U}(Y_k)\right\|^2\right] &\leq 2\mathbb{E}\left[\left\|\bar{H}(Y_k,\xi_{k+1}) - \bar{H}(x_\star,\xi_{k+1})\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right] \\
&\leq 2\mathbb{E}\left[\sigma_k^2\right] + 2\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right] \,. \qquad (119)
\end{aligned}
$$

In addition, using the fact that for any vectors $a, b \in \mathbb{R}^d$, $2|\langle a, b\rangle| \leq m\|a\|^2 + \|b\|^2/m$, we can upper bound the inner product derived in (117) as follows

$$
\begin{aligned}
-\mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{b}\sum_{i=1}^{b}\nabla U^i(X_k^i)\right\rangle\right] &= -\mathbb{E}\left[\left\langle X_k - x_\star, \nabla\bar{U}(X_k)\right\rangle\right] \\
&\quad + \mathbb{E}\left[\left\langle X_k - x_\star, \frac{1}{b}\sum_{i=1}^{b}\left[H^i(X_k,\xi_{k+1}^i) - H^i(X_k^i,\xi_{k+1}^i)\right]\right\rangle\right] \\
&\leq -\mathbb{E}\left[\left\langle X_k - x_\star, \nabla\bar{U}(X_k)\right\rangle\right] + m\mathbb{E}\left[d_k^2\right]/2 + L^2\mathbb{E}\left[V_k\right]/(2m) \\
&\leq -m\mathbb{E}\left[d_k^2\right]/2 + L^2\mathbb{E}\left[V_k\right]/(2m) \,. \qquad (120)
\end{aligned}
$$

Hence, combining (117), (118), (119) and (120) implies that

$$
\mathbb{E}\left[d_{k+1}^2\right] \leq \left(1 - \gamma m + 6\gamma^2\hat{L}^2\right)\mathbb{E}\left[d_k^2\right] + \left(\frac{\gamma L^2}{m} + 6\gamma^2\hat{L}^2\right)\mathbb{E}\left[V_k\right] + 4\gamma^2\mathbb{E}\left[\sigma_k^2\right] + 10\gamma^2\mathbb{E}\left[\|\bar{H}(x_\star,\xi)\|^2\right] + \frac{2\gamma d}{b} \,.
$$

Using the assumption on γ completes the proof. □

Lemma 24. *Assume the L-smoothness of the potentials $\{U^i\}_{i\in[b]}$ and* **A**4 *hold. Then, for any $k \in \mathbb{N}$, we have*

$$
\mathbb{E}\left[\sigma_{k+1}^2\right] \leq (1 - q_c)\mathbb{E}\left[\sigma_k^2\right] + 2q\hat{L}^2\mathbb{E}\left[d_k^2\right] + 2q\hat{L}^2\mathbb{E}\left[V_k\right] \,,
$$

where V_k, d_k, σ_k are defined in (24), (25) *and* (116).

Proof. Let's consider $k \geq 0$, using **A**4 implies that

$$
\begin{aligned}
&\mathbb{E}\left[\sigma_{k+1}^2\right] = \frac{1}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|H^i(Y_{k+1}^i,\xi_{k+1}^i) - H^i(x_\star,\xi_{k+1}^i)\right\|^2\right] \\
&= \frac{1-q_c}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|H^i(Y_k^i,\xi_{k+1}^i) - H^i(x_\star,\xi_{k+1}^i)\right\|^2\right] + \frac{q_c}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|H^i(X_k^i,\xi_{k+1}^i) - H^i(x_\star,\xi_{k+1}^i)\right\|^2\right] \\
&= (1 - q_c)\mathbb{E}\left[\sigma_k^2\right] + \frac{2q}{b}\sum_{i=1}^{b}\mathbb{E}\left[\left\|H^i(X_k^i,\xi_{k+1}^i) - H^i(X_k,\xi_{k+1}^i)\right\|^2 + \left\|H^i(X_k,\xi_{k+1}^i) - H^i(x_\star,\xi_{k+1}^i)\right\|^2\right] \\
&\leq (1 - q_c)\mathbb{E}\left[\sigma_k^2\right] + 2q\hat{L}^2\mathbb{E}\left[d_k^2\right] + 2q\hat{L}^2\mathbb{E}\left[V_k\right] \,.
\end{aligned}
$$

Which shows the expected result. □

For any $\gamma \in (0, m(6\hat{L}^2)^{-1}]$, under **A**1, **A**4 and **A**6 we have shown that Lemma 22 and Lemma 23 imply **HX**3 and **HX**4 with

$$
\begin{array}{llll}
A = c_V = 3\hat{L}^2\,, & B = c_d = 3\tilde{\omega}\,, & C = c_\sigma = 3\,, & D = c = 0\,, \\
\bar{A} = 3L^2\,, & \bar{B} = 3L^2\,, & \bar{C} = 3\,, & \bar{D} = 0\,, \\
A_d = \gamma m/2\,, & B_d = 4\gamma^2\,, & C_d = 2\gamma L^2/m\,, & D_d = (10\gamma^2)\mathbb{E}\left[\left\|\bar{H}(x_\star,\xi)\right\|^2\right] + 2\gamma d/b\,, \\
A_\sigma = q\,, & B_\sigma = 2q\hat{L}^2\,, & C_\sigma = 2q\hat{L}^2\,, & D_\sigma = 0\,.
\end{array} \qquad (121)
$$

For any $\gamma > 0$, consider the following variables

$$\alpha_d = \frac{4\gamma^2}{p_{\mathrm{c}} A_d} \max\left\{p_{\mathrm{c}} B + 3\bar{B}, \frac{4B_\sigma}{A_\sigma}\left(p_{\mathrm{c}} C + 3\bar{C}\right)\right\} , \qquad \alpha_\sigma = \frac{4\gamma^2\left(p_{\mathrm{c}} C + 3\bar{C}\right)}{p_{\mathrm{c}} A_\sigma} . \tag{122}$$

Lemma 25. *Assume **A**1, **A**4 and **A**6 hold with*

$$A_d \leq \min\left(A_\sigma, \frac{p_{\mathrm{c}}}{4}\right) , \qquad \alpha_d C_d + \alpha_\sigma C_\sigma \leq \frac{p_{\mathrm{c}}}{8} , \qquad \alpha_d B_d + \gamma^2\left(C + \frac{3}{p_{\mathrm{c}}}\bar{C}\right) \leq \frac{\alpha_\sigma A_\sigma}{2} ,$$

and consider $\gamma \leq m(6\hat{L}^2)^{-1} \wedge p_{\mathrm{c}}^{1/2}(2-2p_{\mathrm{c}})^{-1/2}[A + (1+2/p_{\mathrm{c}})\bar{A}]^{-1/2}$. Then, for any $k \in \mathbb{N}$, we have

$$\mathbb{E}\left[V_k\right] \leq \left(1 - \frac{A_d}{2}\right)^k \left(\mathbb{E}\left[V_0\right] + \alpha_d \mathbb{E}\left[d_0^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_0^2\right]\right) + \frac{2\alpha_d D_d}{A_d} + \frac{4\left(1-\tau\right)\left(b-1\right)\gamma d}{b A_d} ,$$

where V_k is defined in (24).

Proof. Applying Lemma 12 with the variables provided in (121) gives the result. □

Let's introduce $\gamma_1 > 0$ such that

$$\gamma_1 \leq \frac{m}{128\hat{L}^2} \wedge \frac{m}{8\max\left(3L^2 + p_{\mathrm{c}}\tilde{\omega}, 24\hat{L}^2\right)} \wedge \frac{2q}{m} \wedge \frac{p_{\mathrm{c}}}{2m} \wedge \frac{p_{\mathrm{c}}}{\left[2(1-p_{\mathrm{c}})(p_{\mathrm{c}} A + 3\bar{A})\right]^{1/2}} \wedge \frac{p_{\mathrm{c}}}{8\left[6\left(\frac{L^2}{m^2}\max\left(3L^2 + p_{\mathrm{c}}\tilde{\omega}, 24\hat{L}^2\right)\right) + \frac{2}{q_{\mathrm{c}}}\right]^{1/2}} .$$

Under **A**1, **A**4 and **A**6, for all $\gamma \in (0, \gamma_1]$ the assumptions of Lemma 25 are satisfied. The upper bound on $(\mathbb{E}\left[V_k\right])_{k\in\mathbb{N}}$ derived in Lemma 25 can be rewritten into the format of **H**2 by considering

$$\tilde{v}_1 = \mathbb{E}\left[V_0\right] + \alpha_d \mathbb{E}\left[d_0^2\right] + \alpha_\sigma \mathbb{E}\left[\sigma_0^2\right] , \qquad \tilde{v}_2 = \frac{2\alpha_d D_d}{A_d} + \frac{4\left(1-\tau\right)\left(b-1\right)\gamma d}{b A_d} . \tag{123}$$

In addition, for any $\gamma > 0$, consider the following variables

$$\begin{aligned}
\mathrm{C}^\gamma &= \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}} - 4A_d}\left[B + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{B} + \frac{B_\sigma}{A_\sigma - A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right)\right] , \\
\mathrm{C}^\gamma_r &= \frac{9\gamma^2\left(1-p_{\mathrm{c}}\right)C_\sigma}{p_{\mathrm{c}} - 4A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) + 3\mathrm{C}^\gamma\left(C_d + \frac{B_d C_\sigma}{A_\sigma - A_d}\right) , \\
\mathrm{C}^\gamma_\sigma &= \frac{4(1-p_{\mathrm{c}})\gamma^2}{p_{\mathrm{c}} - 4A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) + \mathrm{C}^\gamma B_d\left(2 + \frac{3}{A_\sigma - A_d}\right) , \\
\mathrm{C}^\gamma_d &= 7\mathrm{C}^\gamma , \qquad \mathrm{C}^\gamma_V = 1 + 2\mathrm{C}^\gamma C_d , \\
\mathrm{C}^\gamma_\delta &= \frac{\mathrm{C}^\gamma D_d}{A_d}\left(1 + \frac{2B_d B_\sigma}{A_d(A_\sigma - A_d)}\right) + \frac{8\left(1-\tau\right)\left(b-1\right)\gamma d}{b p_{\mathrm{c}}} , \\
\mathrm{C}^\gamma_\epsilon &= \mathrm{C}^\gamma_V \mathbb{E}\left[V_0\right] + \mathrm{C}^\gamma_d \mathbb{E}\left[d_0^2\right] + \mathrm{C}^\gamma_\sigma \mathbb{E}\left[\sigma_0^2\right] + 2D_d .
\end{aligned} \tag{124}$$

Based on Lemma 10, we derive the following result.

Lemma 26. *Assume **A**1, **A**4 and **A**6 hold and consider $\gamma \in (0, \gamma_1]$. Then, for any $k \in \mathbb{N}$, we have*

$$\mathbb{E}\left[V_k\right] \leq \left(1 - \frac{A_d}{4}\right)^k\left(\mathrm{C}^\gamma_\epsilon + \frac{4\mathrm{C}^\gamma_r \tilde{v}_1}{A_d}\right) + \frac{2\mathrm{C}^\gamma_r \tilde{v}_2}{A_d} + \mathrm{C}^\gamma_\delta ,$$

where V_k is defined in (24) and $\mathrm{C}^\gamma_\epsilon, \mathrm{C}^\gamma_r, \mathrm{C}^\gamma_\delta$ in (124).

Proof. Since we suppose **A**1, **A**4 and **A**6 hold with $\gamma \leq \gamma_1$, the assumptions of Lemma 25 are satisfied. Therefore, for any $l \in \mathbb{N}$, we obtain

$$\mathbb{E}\left[V_l\right] \leq \left(1 - \frac{A_d}{2}\right)^l \tilde{v}_1 + \tilde{v}_2 \,. \tag{125}$$

Moreover, the condition $\gamma \leq m/128\hat{L}^2$ ensures that $A_d A_\sigma = q\gamma m/2 \geq 8 B_d B_\sigma = 64 q \gamma^2 \hat{L}^2$, hence we can apply Lemma 10. Then, plugging (125) in the bound derived in Lemma 10 gives

$$\mathbb{E}\left[V_k\right] \leq (1-\alpha)^k \mathrm{C}_\epsilon^\gamma + \mathrm{C}_r^\gamma \sum_{i=0}^{k-2} (1-\alpha)^{k-i-1} \mathbb{E}\left[V_i\right] + \mathrm{C}_\delta^\gamma \,, \tag{126}$$

where α is defined in (60) by

$$\alpha = A_d - \frac{2(A_\sigma - A_d)^{-1} B_d B_\sigma}{1 + \sqrt{1 + 4(1-A_d)^{-1}(A_\sigma - A_d)^{-1} B_d B_\sigma}} \,.$$

Using Lemma 9, we know that $A_d/2 < \alpha \leq A_d$ and combining this bound with (125) and (126) leads to

$$\mathbb{E}\left[V_k\right] \leq \left(1 - \frac{A_d}{4}\right)^k \left(\mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_r^\gamma \tilde{v}_1}{A_d}\right) + \frac{2\mathrm{C}_r^\gamma \tilde{v}_2}{A_d} + \mathrm{C}_\delta^\gamma \,.$$

□

In order to rewrite the upper bound on $(\mathbb{E}\left[V_k\right])_{k \in \mathbb{N}}$ given in Lemma 26 in the format of **H**2, we consider for $\gamma > 0$

$$v_1 = \mathrm{C}_\epsilon^\gamma + \frac{4\mathrm{C}_r^\gamma \tilde{v}_1}{A_d} \,, \qquad\qquad v_2 = \frac{2\mathrm{C}_r^\gamma \tilde{v}_2}{A_d} + \mathrm{C}_\delta^\gamma \,. \tag{127}$$

Lemma 27. *Assume **A**1, **A**7, **H**1 and hold and let $\gamma \leq (6L)^{-1}$. Using the convention that $\sum_0^{-1} = 0$, then for any $k \in \mathbb{N}$, we have*

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\right\|^2\right] &\leq \left[1 - \gamma m + \gamma^2\left(3mL + 4\omega\right)\right] \mathbb{E}\left[\left\|\mathsf{X}_{k\gamma} - X_k\right\|^2\right] \\
&\quad + 4\gamma^2 \omega q_{\mathrm{c}} \sum_{l=0}^{k-1} (1-q_{\mathrm{c}})^{k-l-1} \mathbb{E}\left[\left\|\mathsf{X}_{l\gamma} - X_l\right\|^2\right] + \gamma\left(\frac{2L^2}{m} + 3\gamma L^2 + 4\gamma\omega\right) \mathbb{E}\left[V_k\right] \\
&\quad + \left(\frac{2}{\gamma m} \mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2\right] + 3\mathbb{E}\left[\left\|I_k\right\|^2\right]\right) + \frac{16\gamma^3 \omega d}{b q_{\mathrm{c}}}\left(1 + \frac{\gamma L}{q_{\mathrm{c}}}\right) .
\end{aligned}$$

Proof. For $k \in \mathbb{N}$, using the independence of $(\xi_{k+1}^i)_{i \in [b]}$ combined with **H**1 and **A**7, we obtain

$$\begin{aligned}
\mathrm{Var}^{\mathcal{F}_k}\left(\frac{1}{b}\sum_{i=1}^b G_k^i\right) &= \mathbb{E}^{\mathcal{F}_k}\left[\left\|\frac{1}{b}\sum_{i=1}^b \left[\nabla U^i(X_k^i) - \nabla U^i(Y_k) - H^i(X_k^i, \xi_{k+1}^i) + H^i(Y_k, \xi_{k+1}^i)\right]\right\|^2\right] \\
&= \frac{1}{b}\sum_{i=1}^b \mathbb{E}^{\mathcal{F}_k}\left[\left\|\nabla U^i(X_k^i) - \nabla U^i(Y_k) - H^i(X_k^i, \xi_{k+1}^i) + H^i(Y_k, \xi_{k+1}^i)\right\|^2\right] \\
&\leq \frac{\omega}{b}\sum_{i=1}^b \left\|X_k^i - Y_k\right\|^2 .
\end{aligned} \tag{128}$$

Denote $t_k \in \mathbb{N}$ the time when the reference point of the control variate is updated, therefore we have

$$t_k = \begin{cases} 0\,, & \text{if } k = 0 \\ \max\left\{l \in \{0, \dots, k-1\} \,:\, Y_k = b^{-1}\sum_{i=1}^b X_k^i\right\}\,, & \text{if } k \geq 1 \end{cases} . \tag{129}$$

Hence, for any $i \in [b], k \geq 0$, we have

$$X_k^i - Y_k = (X_k^i - X_k) + (X_k - \mathsf{X}_{k\gamma}) + (\mathsf{X}_{k\gamma} - \mathsf{X}_{t_k\gamma}) + (\mathsf{X}_{t_k\gamma} - Y_k)\,.$$

Thus for $k \geq 0$, combining the previous line with Young's inequality, it yields that

$$\frac{1}{b}\sum_{i=1}^{b} \mathbb{E}\left[\left\|X_k^i - Y_k\right\|^2\right] \leq 4\mathbb{E}\left[V_k\right] + 4\mathbb{E}\left[\|X_k - \mathsf{X}_{k\gamma}\|^2\right] + 4\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - \mathsf{X}_{t_k\gamma}\|^2\right] + 4\mathbb{E}\left[\|\mathsf{X}_{t_k\gamma} - Y_k\|^2\right]\,. \tag{130}$$

For $k \geq 1$, by definition of t_k, we have

$$\mathbb{E}\left[V_{t_k}\right] = \sum_{l=0}^{k-1} \mathbb{P}\left(t_k = l\right)\mathbb{E}\left[V_l\right] = q\sum_{l=0}^{k-1}(1-q_{\mathrm{c}})^{k-l-1}\mathbb{E}\left[V_l\right]\,.$$

Moreover, for $k \geq 1$ we get

$$\begin{aligned}
\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - \mathsf{X}_{t_k\gamma}\|^2\right] &= \sum_{l=0}^{k-1}\mathbb{P}\left(t_k = l\right)\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - \mathsf{X}_{l\gamma}\|^2\right] \\
&= q\sum_{l=0}^{k-1}(1-q_{\mathrm{c}})^{k-l-1}\mathbb{E}\left[\left\|-\int_{l\gamma}^{k\gamma}\nabla\bar{U}(\mathsf{X}_s)\mathrm{d}s + \sqrt{\frac{2}{b}}(\mathsf{W}_{k\gamma} - \mathsf{W}_{l\gamma})\right\|^2\right] \\
&\leq 2\gamma q\sum_{l=0}^{k-1}(k-l)(1-q_{\mathrm{c}})^{k-l-1}\left(\int_{l\gamma}^{k\gamma}\mathbb{E}\left[\left\|\nabla\bar{U}(\mathsf{X}_s)\right\|^2\right]\mathrm{d}s + \frac{2d}{b}\right)\,.
\end{aligned} \tag{131}$$

Using Dalalyan (2017, Lemma 2) with $s \in \mathbb{R}_+$, we obtain

$$\mathbb{E}\left[\|\nabla\bar{U}(\mathsf{X}_s)\|^2\right] \leq dL/b\,.$$

Using by convention that $\sum_{l=1}^{0} = 0$, for any $k \in \mathbb{N}$ and $x \neq 1$ we have

$$\sum_{l=1}^{k} l^2 x^{l-1} = (1-x)^{-3}\left(1 + x - x^k\left[2x + kx(1-x) + (k+1)(1+k(1-x))(1-x)\right]\right)\,.$$

Thus, setting $x = 1-q$ inside the last shows that

$$\sum_{l=1}^{k} l^2(1-q_{\mathrm{c}})^{l-1} \leq 2/q_{\mathrm{c}}^3\,.$$

Hence, the above line combined with $\sum_{l=1}^{k} l(1-q_{\mathrm{c}})^{l-1} = q^{-2}\left[1-(1+kq)(1-q_{\mathrm{c}})^k\right]$ and (131) yield the following upper bound

$$\begin{aligned}
\mathbb{E}\left[\|\mathsf{X}_{k\gamma} - \mathsf{X}_{t_k\gamma}\|^2\right] &\leq \frac{2\gamma dq}{b}\sum_{l=0}^{k-1}\left[(k-l)(1-q_{\mathrm{c}})^{k-l-1}\left(2 + (k-l)\gamma L\right)\right] \\
&\leq \frac{2\gamma dq}{b}\sum_{l=0}^{k-1}\left[(k-l)(1-q_{\mathrm{c}})^{k-l-1}\left(2 + (k-l)\gamma L\right)\right] \\
&\leq \frac{4\gamma d}{bq_{\mathrm{c}}}\left(1 + \frac{\gamma L}{q_{\mathrm{c}}}\right)\,.
\end{aligned} \tag{132}$$

In addition, by definition (129) of t_k, we immediately get for any $k \geq 1$, that

$$\mathbb{E}\left[\|\mathsf{X}_{t_k\gamma} - X_{t_k}\|^2\right] = \sum_{l=0}^{k-1}\mathbb{P}\left(t_k = l\right)\mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right]$$

$$= q \sum_{l=0}^{k-1} (1-q_{\mathrm{c}})^{k-l-1} \mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right] .$$

Combining (128), (130) with (132), for any $k \geq 1$ we obtain

$$\mathbb{E}\left[\operatorname{Var}^{\mathcal{F}_k}\left(\frac{1}{b}\sum_{i=1}^{b} G_k^i\right)\right] \leq 4\omega \mathbb{E}\left[\|X_k - \mathsf{X}_{k\gamma}\|^2\right] + 4\omega q_{\mathrm{c}} \sum_{l=0}^{k-1} (1-q_{\mathrm{c}})^{k-l-1} \mathbb{E}\left[\|\mathsf{X}_{l\gamma} - X_l\|^2\right] \\ + 4\omega \mathbb{E}\left[V_k\right] + \frac{16\gamma\omega d}{b q_{\mathrm{c}}}\left(1 + \frac{\gamma L}{q_{\mathrm{c}}}\right) . \tag{133}$$

Since $Y_0 = b^{-1}\sum_{i=1} X_0^i$, we have $\operatorname{Var}^{\mathcal{F}_k}(b^{-1}\sum_{i=1}^{b} G_k^i) \leq \omega V_k$ and therefore the above inequality also holds for $k = 0$. Lastly, using Proposition 4 gives

$$\mathbb{E}^{\mathcal{F}_k}\left[\|\mathsf{X}_{(k+1)\gamma} - X_{k+1}\|^2\right] \leq \left[1 - \gamma m\left(1 - 3\gamma L\right)\right] \|\mathsf{X}_{k\gamma} - X_k\|^2 + \gamma\left(\frac{2L^2}{m} + 3\gamma L^2\right) V_k \\ + \left(\frac{2}{\gamma m}\left\|\mathbb{E}^{\mathcal{F}_k}\left[I_k\right]\right\|^2 + 3\mathbb{E}^{\mathcal{F}_k}\left[\|I_k\|^2\right]\right) + \gamma^2 \operatorname{Var}^{\mathcal{F}_k}\left(\frac{1}{b}\sum_{i=1}^{b} G_k^i\right) .$$

Hence, plugging (133) in the above inequality yields the expected result. □

Based on Lemma 27, for any $\gamma > 0$ introduce the following notations

$$\alpha_0 = (1-q_{\mathrm{c}})^{-1}\left[1 - \gamma m + \gamma^2\left(3mL + 4\omega\right)\right] , \qquad \alpha_1 = \frac{4\gamma^2\omega q}{(1-q_{\mathrm{c}})^2} , \tag{134}$$

$$\alpha_2 = \frac{\gamma}{1-q_{\mathrm{c}}}\left(\frac{2L^2}{m} + 3\gamma L^2 + 4\gamma\omega\right) , \qquad \alpha_3 = 0 ,$$

$$\alpha_4 = (1-q_{\mathrm{c}})^{-1}\left(\frac{2\sup_{l\in\mathbb{N}}}{\gamma m}\mathbb{E}\left[\left\|\mathbb{E}^{\mathcal{F}_l}\left[I_l\right]\right\|^2\right] + 3\sup_{l\in\mathbb{N}}\mathbb{E}\left[\|I_l\|^2\right] + \frac{16\gamma^3\omega d}{b q_{\mathrm{c}}}\left(1 + \frac{\gamma L}{q_{\mathrm{c}}}\right)\right) .$$

For ease of reading, we also introduce the time step-size γ_2 defined by

$$\gamma_2 \leq \frac{q_{\mathrm{c}}}{L} \wedge \frac{q_{\mathrm{c}}}{2m} \wedge \frac{1}{6(L + 4m^{-1}\omega)} .$$

Theorem 28. *Assume* **A**1, **A**4, **A**6, **A**7 *and let* $\gamma \in (0, \gamma_1 \wedge \gamma_2)$. *Then, for any initial probability measure* $\mu_0^{(\mathrm{Vr}\star)} \in \mathcal{P}_2(\mathbb{R}^d)$, $k \in \mathbb{N}$, *we have*

$$\mathbf{W}_2^2\left(\mu_k^{(\mathrm{Vr}\star)}, \pi\right) \leq \left(1 - \frac{\gamma m}{2}\right)^k \mathbf{W}_2^2\left(\mu_0^{(\mathrm{Vr}\star)}, \pi\right) + \left(1 - \frac{\gamma m}{8}\right)^k \frac{3L^2}{m^2} v_1 + \frac{6L^2}{m^2} v_2 + \frac{6\gamma d}{b m^2}\kappa_I + \frac{32\gamma^2\omega d}{b m q} ,$$

where v_1, v_2 *are defined in* (127) *and* $\kappa_I = L^2(1 + \gamma L^2/m)$. *If in addition we suppose* **HX**1, *set* $\kappa_I = 2\gamma(L^3 + d\tilde{L}^2/b)$.

Proof. We know that **H**1 is satisfied since for any $i \in [b], x \in \mathbb{R}^d$ the stochastic gradient $H^i(x, \xi^i)$ is unbiased. Lemma 26 proves that **H**2 holds with $\alpha_v = 1 - A_d/4$ and v_1, v_2 defined in (127). Lemma 27 implies that **HX**2 holds with the choice of $(\alpha_i)_{i=0}^4$ detailed in (134). Finally, since **HX**2 and **H**2 hold, we can apply Theorem 8 to show that

$$\mathbf{W}_2^2\left(\mu_k^{(\mathrm{Vr}\star)}, \pi\right) \leq (1 + \alpha_0 + \delta)^k (1-q_{\mathrm{c}})^k \mathbf{W}_2^2\left(\mu_0^{(\mathrm{Vr}\star)}, \pi\right) \\ + (1-q_{\mathrm{c}}) v_1 \left(\alpha_2 + \frac{\alpha_3}{\alpha_0 + \delta}\right) \frac{\alpha_v^k - (1 + \alpha_0 + \delta)^k (1-q_{\mathrm{c}})^k}{\alpha_v - (1 + \alpha_0 + \delta)(1-q_{\mathrm{c}})} \\ + \frac{1-q_{\mathrm{c}}}{q_{\mathrm{c}} - (1-q_{\mathrm{c}})(\alpha_0 + \delta)}\left[\left(\alpha_2 + \frac{\alpha_3}{\alpha_0 + \delta}\right) v_2 + \alpha_4\right] , \tag{135}$$

where $\delta = 2^{-1}(\sqrt{(\alpha_0-1)^2+4\alpha_1}-1-\alpha_0)$ is defined in (45). Using for any $a>0, b\geq 0$, that $\sqrt{a+b}\leq\sqrt{a}+b/(2\sqrt{a})$, we obtain

$$\begin{aligned}\alpha_0+\sqrt{(\alpha_0-1)^2+4\alpha_1} &= 1+(\alpha_0-1)\left(1+\sqrt{1+\frac{4\alpha_1}{(\alpha_0-1)^2}}\right)\\ &\leq 1+2\left(\alpha_0-1\right)\left(1+\frac{\alpha_1}{(\alpha_0-1)^2}\right)=2\alpha_0-1+\frac{2\alpha_1}{\alpha_0-1}\,.\end{aligned}$$

Since $\gamma\leq\gamma_2\leq q(2m)^{-1}\wedge\{6(L+4m^{-1}\omega)\}^{-1}$, the previous line implies that

$$\begin{aligned}2\left(1-q_{\mathrm{c}}\right)\left(1+\alpha_0+\delta\right) &= \left(1-q_{\mathrm{c}}\right)\left(1+\alpha_0+\sqrt{(\alpha_0-1)^2+4\alpha_1}\right)\\ &\leq 2(1-q_{\mathrm{c}})\left(\alpha_0+\frac{\alpha_1}{\alpha_0-1}\right)\\ &= 2\left(1-\gamma m+\gamma^2\left(3mL+4\omega+\frac{4q\omega}{q_{\mathrm{c}}-\gamma m+\gamma^2\left(3mL+4\omega\right)}\right)\right)\\ &\leq 2\left(1-\gamma m/2\right)\,. \end{aligned} \tag{136}$$

This upper bound gives that

$$(1-q_{\mathrm{c}})(\alpha_0+\delta)=\left(1-q_{\mathrm{c}}\right)\left(1+\alpha_0+\delta\right)+q-1\leq q-\gamma m/2\,.$$

Thus, we deduce that

$$\frac{1}{q_{\mathrm{c}}-(1-q_{\mathrm{c}})(\alpha_0+\delta)}\leq\frac{2}{\gamma m}\,. \tag{137}$$

Further, using $\gamma\leq\gamma_2$ combined with the definitions of $\alpha_0,\alpha_2,\alpha_3,\alpha_v$ and δ show that

$$\begin{aligned}\frac{\alpha_v^k-\left(1+\alpha_0+\delta\right)^k\left(1-q_{\mathrm{c}}\right)^k}{\alpha_v-\left(1+\alpha_0+\delta\right)\left(1-q_{\mathrm{c}}\right)} &\leq\frac{8}{3\gamma m}\left(1-\frac{\gamma m}{8}\right)^k\,,\\ \alpha_2+\frac{\alpha_3}{\alpha_0+\delta}=\frac{\gamma}{1-q_{\mathrm{c}}}\left(\frac{2L^2}{m}+3\gamma L^2+4\gamma\omega\right) &\leq\frac{3\gamma L^2}{(1-q_{\mathrm{c}})m}\,.\end{aligned} \tag{138}$$

Lastly, plugging (136), (137) and (138) in (135) yields

$$\mathbf{W}_2^2\left(\mu_k^{(\mathrm{Vr}\star)},\pi\right)\leq\left(1-\frac{\gamma m}{2}\right)^k\mathbf{W}_2^2\left(\mu_0^{(\mathrm{Vr}\star)},\pi\right)+\left(1-\frac{\gamma m}{8}\right)^k\frac{3L^2}{m^2}v_1+\frac{6L^2}{m^2}v_2+\frac{2(1-q_{\mathrm{c}})\alpha_4}{\gamma m}\,. \tag{139}$$

In addition, following the lines provided in the proof of Theorem 20, we deduce

$$\frac{2(1-q_{\mathrm{c}})\alpha_4}{\gamma m}\leq\frac{6\gamma dL^2}{bm^2}\left(1+\frac{19\gamma L^2}{36m}\right)+\frac{32\gamma^2\omega d}{bmq}\,. \tag{140}$$

If in addition we suppose **HX1**, then we obtain

$$\frac{2(1-q_{\mathrm{c}})\alpha_4}{\gamma m}\leq\gamma mL^2\left(1+\frac{\gamma L^2}{2m}+\frac{\gamma^2L^2}{12}\right)+\frac{4\gamma}{9}\left(L^3+\frac{d\tilde{L}^2}{b}\right)+\frac{32\gamma^2\omega d}{bmq}\,. \tag{141}$$

Finally, plugging (140) or (141) if **HX1** holds inside (139) combined with $\gamma\leq qL^{-1}$ lead to the expected result. □

Now, consider the time stepsizes γ_3 and $\gamma_\star$ defined by

$$\gamma_3=\frac{p_{\mathrm{c}}m}{3L^2+16\hat{L}^2+p_{\mathrm{c}}\tilde{\omega}}\,,\qquad\qquad\gamma_\star=\gamma_1\wedge\gamma_2\wedge\gamma_3\,.$$

From the previous result, the next corollary controls the asymptotic bias obtained by Algorithm 3.

Corollary 29. *Assume* ***A**1*, ***A**4*, ***A**6*, ***A**7* *and let* $\gamma \in (0, \gamma_\star)$ *with* $\tau = 1$. *Then, for any initial probability measure* $\mu_0^{(\mathrm{Vr}\star)} \in \mathcal{P}_2(\mathbb{R}^d)$, $k \in \mathbb{N}$, *we have*

$$\frac{9^{-9} b}{\gamma d} \limsup_{k\to\infty} \mathbf{W}_2^2\left(\mu_k^{(\mathrm{Vr}\star)}, \pi\right) \leq \frac{\kappa_I}{m^2} + \frac{\gamma\omega}{mq} \\ + \frac{(1-p_{\mathrm{c}})\gamma L^2}{p_{\mathrm{c}}^2 m^5}\left(L^2 + \hat{L}^2 + p_{\mathrm{c}}\tilde{\omega}\right)\left(1 + \frac{\gamma}{bd}\mathbb{E}\left[\left\|H(x_\star, \xi)\right\|^2\right]\right)\left(L^2 + \frac{q_{\mathrm{c}}}{p_{\mathrm{c}}}\hat{L}^2\right),$$

where $\tilde{\kappa}_I = L^2(1 + \gamma L^2 m^{-1})$ *and if we suppose* ***HX**1*, $\tilde{\kappa}_I = \gamma(L^3 + d\tilde{L}^2 b^{-1})$.

Proof. Applying Theorem 28 with $\gamma \in (0, \gamma_1 \wedge \gamma_2)$ shows that

$$\limsup_{k\to\infty} \mathbf{W}_2^2\left(\mu_k^{(\mathrm{Vr}\star)}, \pi\right) \leq \frac{6L^2}{m^2} v_2 + \frac{6\gamma d}{bm^2}\kappa_I + \frac{32\gamma^2 \omega d}{bmq} \\ \leq \frac{6L^2 \mathrm{C}_\delta^\gamma}{m^2} + \frac{12 L^2 \mathrm{C}_r^\gamma \tilde{v}_2}{A_d m^2} + \frac{6\gamma d}{bm^2}\kappa_I + \frac{32\gamma^2\omega d}{bmq}. \tag{142}$$

Plugging the definitions of $\tilde{v}_1, \tilde{v}_2$ provided in (123) combined with the previous inequality, we obtain

$$\limsup_{k\to\infty} \mathbf{W}_2^2\left(\mu_k^{(\mathrm{Vr}\star)}, \pi\right) \leq \frac{6L^2\mathrm{C}_\delta^\gamma}{m^2} + \frac{24 L^2 \mathrm{C}_r^\gamma \alpha_d D_d}{A_d^2 m^2} + \frac{48 L^2 \mathrm{C}_r^\gamma\,(1-\tau)\,(b-1)\,\gamma d}{bA_d^2 m^2} + \frac{6\gamma d}{bm^2}\kappa_I + \frac{32\gamma^2\omega d}{bmq}.$$

Further, recall that $A_d, B, \bar{B}, D, \bar{D}, D_d$ are provided in (121) and α_d is defined in (122) by

$$\alpha_d = \frac{4\gamma^2}{p_{\mathrm{c}} A_d}\max\left\{p_{\mathrm{c}} B + 3\bar{B}, \frac{4B_\sigma}{A_\sigma}\left(p_{\mathrm{c}} C + 3\bar{C}\right)\right\} \\ = \frac{24\gamma}{p_{\mathrm{c}} m}\max\left\{3L^2 + p_{\mathrm{c}}\tilde{\omega}, 8(p_{\mathrm{c}}+3)\hat{L}^2\right\} \leq \frac{768\gamma}{p_{\mathrm{c}} m}\left(L^2 + \hat{L}^2 + p_{\mathrm{c}}\tilde{\omega}\right).$$

Moreover, $\mathrm{C}_\delta^\gamma, \mathrm{C}_r^\gamma$ are defined in (124) by

$$\mathrm{C}_\delta^\gamma = \frac{\mathrm{C}^\gamma D_d}{A_d}\left(1 + \frac{2B_d B_\sigma}{A_d(A_\sigma - A_d)}\right) + \frac{8\,(1-\tau)\,(b-1)\,\gamma d}{bp_{\mathrm{c}}} \\ = \frac{10\mathrm{C}^\gamma}{m}\left(1 + \frac{64\gamma q\hat{L}^2}{(2q - \gamma m)m}\right)\left(5\gamma\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] + \frac{d}{b}\right) + \frac{8\,(1-\tau)\,(b-1)\,\gamma d}{bp_{\mathrm{c}}} \\ \leq \frac{360(1-p_{\mathrm{c}})\gamma^2}{mp_{\mathrm{c}}^2}\left(3L^2 + 11\hat{L}^2 + p_{\mathrm{c}}\tilde{\omega}\right)\left(5\gamma\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] + \frac{d}{b}\right) + \frac{8\,(1-\tau)\,(b-1)\,\gamma d}{bp_{\mathrm{c}}}, \tag{143}$$

$$\mathrm{C}_r^\gamma = \frac{9\gamma^2\,(1-p_{\mathrm{c}})\,C_\sigma}{p_{\mathrm{c}} - 4A_d}\left(C + \frac{2+p_{\mathrm{c}}}{p_{\mathrm{c}}}\bar{C}\right) + 3\mathrm{C}^\gamma\left(C_d + \frac{B_d C_\sigma}{A_\sigma - A_d}\right) \\ \leq \frac{144\gamma^2(1-p_{\mathrm{c}})}{p_{\mathrm{c}}^2}\left[3q\hat{L}^2 + \gamma\left(\frac{L^2}{m} + 8\gamma\hat{L}^2\right)\left(p_{\mathrm{c}}\tilde{\omega} + 3L^2 + 16\hat{L}^2\right)\right] \leq \frac{432\gamma^2(1-p_{\mathrm{c}})}{p_{\mathrm{c}}^2}\left(p_{\mathrm{c}} L^2 + q\hat{L}^2\right)$$

Eventually, for the specific choice $\tau = 1$ combined with (142) and (143), it yields that

$$\limsup_{k\to\infty} \mathbf{W}_2^2\left(\mu_k^{(\mathrm{Vr}\star)}, \pi\right) \leq \frac{6\gamma d}{bm^2}\kappa_I + \frac{32\gamma^2\omega d}{bmq} + \frac{18432\gamma\mathrm{C}_r^\gamma D_d L^2}{A_d^2 m^3 p_{\mathrm{c}}}\left(L^2 + \hat{L}^2 + p_{\mathrm{c}}\tilde{\omega}\right) \\ + \frac{2160(1-p_{\mathrm{c}})\gamma^2 L^2}{p_{\mathrm{c}}^2 m^3}\left(3L^2 + 11\hat{L}^2 + p_{\mathrm{c}}\tilde{\omega}\right)\left(5\gamma\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] + \frac{d}{b}\right). \tag{144}$$

Therefore, using (143) and (144) we can finally conclude that

$$9^9 \limsup_{k\to\infty} \mathbf{W}_2^2\left(\mu_k^{(\mathrm{Vr}\star)}, \pi\right) \leq \frac{\gamma d}{bm^2}\kappa_I + \frac{\gamma^2\omega d}{bmq} \\ + \frac{(1-p_{\mathrm{c}})\gamma^2 L^2}{p_{\mathrm{c}}^2 m^5}\left(L^2 + \hat{L}^2 + p_{\mathrm{c}}\tilde{\omega}\right)\left(\gamma\mathbb{E}\left[\left\|\bar{H}(x_\star, \xi)\right\|^2\right] + \frac{d}{b}\right)\left(L^2 + \frac{q_{\mathrm{c}}}{p_{\mathrm{c}}}\hat{L}^2\right).$$

□

The single client case corresponds to $b = p_{\mathrm{c}} = 1$ and leads for $k \geq 0$ to $V_k = 0$. Moreover, the assumption **H**2 holds with $v_1 = v_2 = 0$. Thus, we obtain a convergence bound for SVRG-LD from Theorem 28.

Theorem 30. *Assume **A**1, **A**4, **A**6, **A**7 and let $\gamma \in (0, \gamma_1 \wedge \gamma_2)$. Then, for any initial probability measure $\mu_0^{(\mathrm{Vr}\star)} \in \mathcal{P}_2(\mathbb{R}^d)$, $k \in \mathbb{N}$, we have*

$$\mathbf{W}_2^2\left(\mu_k^{(\mathrm{Vr}\star)}, \pi\right) \leq \left(1 - \frac{\gamma m}{2}\right)^k \mathbf{W}_2^2\left(\mu_0^{(\mathrm{Vr}\star)}, \pi\right) + \frac{6\gamma d}{bm^2}\kappa_I + \frac{32\gamma^2 \omega d}{mq} ,$$

*where $\kappa_I = L^2(1 + \gamma L^2/2m + \gamma^2 L^2/12)$. If in addition we suppose **HX**1, set $\kappa_I = 3\gamma(L^3 + d\tilde{L}^2/b)$.*

Remark 31.

- *The constants obtained in this result can be refined by directly using that $\mathbb{E}[V_k] = 0$ in the proof of Lemma 27 and by simplifying the calculations detailed in Theorem 28.*

- *The proof given in Chatterji et al. (2018, Theorem 4.2-Option 2) on the convergence of SVRG-LD seems to have some gaps since the authors use Grönwall's inequality (Clark, 1987) as if $\spadesuit = \tau^2\left(8\delta d + 4M\delta^2 d + 4\delta^2 M \Omega_1\right)$ were constant, which is not the case because $\Omega_1 = \langle \nabla f(y_k) - \nabla f(x_k), y_k - x_k \rangle$ depends on the iteration k. If we denote $\spadesuit_k$ instead of $\spadesuit$ and adopt their other notation (we also correct a typography in the right-hand term), we obtain*

$$\mathbb{E}\left[\|x_k - \tilde{x}\|_2^2\right] \leq \spadesuit_k + \sum_{j=\tau s}^{k-1} \mathbb{E}\left[\|x_j - \tilde{x}\|_2^2\right] . \tag{145}$$

Then, it is claimed in the proof of Chatterji et al. (2018, Theorem 4.2-Option 2) that (145) implies $\mathbb{E}[\|x_k - \mathsf{X}_k\|^2] \leq \spadesuit_k \exp(\tau\rho)$. But this inequality cannot hold in all generalities, for example if we consider : $\tau s = 0$, for $j < k$, $\spadesuit_j = 1$, $x_j = \tilde{x} + \sqrt{2^j/d} \cdot \mathbf{1}$ and $\spadesuit_k = 0$, $x_k = \tilde{x} + \mathbf{1}/\sqrt{d}$, then (145) holds for $j \in [k]$ but $\mathbb{E}[\|x_k - \mathsf{X}_k\|^2] = 1$ whereas $\spadesuit_k \exp(\tau\rho) = 0$.

8 Lower bound on the heterogeneity in a Gaussian case

In this section, we want to illustrate the heterogeneity problem by lower bounding the Wasserstein distance $\mathbf{W}_2$ in a simple case. To simplify the calculations, we assume that each client performs 2 local iterations following the FALD update before communicating its local parameter to the central server. More specifically, take $(\mu_1, \mu_2, \sigma_1, \sigma_2) \in \mathbb{R}^2 \times (\mathbb{R}_+^*)^2$ and define the potentials $U^1 : x \in \mathbb{R}^d \mapsto \sigma_1^{-2}(x - \mu_1)^2$, $U^2 : x \in \mathbb{R}^d \mapsto \sigma_2^{-2}(x - \mu_2)^2$. Thus, the global posterior distribution π is Gaussian with mean $\bar{\mathrm{m}}$ and variance $\bar{\sigma}^2$ given by

$$\bar{\mathrm{m}} = \frac{\mu_1 \sigma_2^2 + \mu_2 \sigma_1^2}{\sigma_1^2 + \sigma_2^2} \qquad\qquad \bar{\sigma} = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1/2} . \tag{146}$$

The objective is to illustrate the problem of heterogeneity in the basic version of FALD. To do so, we first show that this algorithm generates samples targeting a distribution $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$ where the distance $\mathbf{W}_2(\pi, \pi_\gamma)$ is lower bounded by a heterogeneity term. To this end, we introduce the Markov kernel, which for each $\gamma > 0, \mathsf{B} \in \mathcal{B}(\mathbb{R}^d)$ is given by

$$P_\gamma(x, \mathsf{B}) = \int_{\mathsf{B}} \exp\left(-\frac{\left\|x' - \left(1 - \frac{\gamma}{\bar{\sigma}^2} + \frac{\gamma^2}{2}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right)\right)x - \frac{\gamma\bar{\mathrm{m}}}{\bar{\sigma}^2} + \frac{\gamma^2}{2}\left(\frac{\mu_1}{\sigma_1^4} + \frac{\mu_2}{\sigma_2^4}\right)\right\|^2}{2\gamma\left(1 + \left(1 - \frac{\gamma}{2\bar{\sigma}^2}\right)^2\right)}\right)\frac{\mathrm{d}x'}{(2\pi)^{d/2}} ,$$

and we define the stochastic processes $(A_k, \tilde{A}_k)_{k \geq 0}$ on $\mathbb{R}^d \times \mathbb{R}^d$ starting from $(X_0, X_0) = (x, \tilde{x})$ and following the recursion for $k \geq 0$,

$$\begin{aligned} A_{k+1} &= A_k - \frac{\gamma}{\bar{\sigma}^2}(A_k - \bar{\mathrm{m}}) + \frac{\gamma^2}{2}\left(\frac{A_k - \mu_1}{\sigma_1^4} + \frac{A_k - \mu_2}{\sigma_2^4}\right) + \sqrt{\gamma}\left[\left(1 - \frac{\gamma}{2\bar{\sigma}^2}\right)Z_{k+1} + Z_{k+2}\right] , \\ \tilde{A}_{k+1} &= \tilde{A}_k - \frac{\gamma}{\bar{\sigma}^2}\left(\tilde{A}_k - \bar{\mathrm{m}}\right) + \frac{\gamma^2}{2}\left(\frac{\tilde{A}_k - \mu_1}{\sigma_1^4} + \frac{\tilde{A}_k - \mu_2}{\sigma_2^4}\right) + \sqrt{\gamma}\left[\left(1 - \frac{\gamma}{2\bar{\sigma}^2}\right)Z_{k+1} + Z_{k+2}\right] . \end{aligned} \tag{147}$$

It is possible to verify that $(A_k, \tilde{A}_k)$ is distributed according to $(\delta_x P_\gamma^k, \delta_{\tilde{x}} P_\gamma^k)$.

Lemma 32. *Let $\gamma \in \left(0, 2(\sigma_1\sigma_2)^4[\bar{\sigma}^2(\sigma_1^4+\sigma_2^4)]^{-1}\right)$. Then, there exists $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$ such that for any distribution $\pi^0 \in \mathcal{P}_2(\mathbb{R}^d)$, the sequence $(\pi^0 P_\gamma^k)_{k\in\mathbb{N}}$ converges to π_γ in $\mathcal{P}_2(\mathbb{R}^d)$.*

Proof. Let $k \in \mathbb{N}$ and consider the stochastic processes $(A_l, \tilde{A}_l)_{l\in\mathbb{N}}$ defined in (147), subtracting the two recursions we obtain

$$A_{k+1} - \tilde{A}_{k+1} = \left(1 - \frac{\gamma}{\bar{\sigma}^2} + \frac{\gamma^2}{2}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right)\right)\left(A_k - \tilde{A}_k\right) .$$

Since $0 < \gamma < 2(\sigma_1\sigma_2)^4[\bar{\sigma}^2(\sigma_1^4+\sigma_2^4)]^{-1}$, taking the norm in the previous inequality implies that

$$\|A_{k+1} - \tilde{A}_{k+1}\| = \left(1 - \frac{\gamma}{\bar{\sigma}^2} + \frac{\gamma^2}{2}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right)\right)\|A_k - \tilde{A}_k\| . \tag{148}$$

Finally, combining (148) with Douc et al. (2018, Lemma 20.3.2), we deduce that the c-Dobrushin coefficient of P_γ is upper bounded by $1 - \gamma/\bar{\sigma}^2 + \gamma^2/2\left(1/\sigma_1^4 + 1/\sigma_2^4\right)$. Hence, applying Douc et al. (2018, Theorem 20.3.4) we deduce the existence and uniqueness of a stationary distribution $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$ for the Markov Kernel P_γ such that $\mathbf{W}_2(\pi^0 P_\gamma^k, \pi) \leq \left(1 - \gamma/\bar{\sigma}^2 + \gamma^2/2\left(1/\sigma_1^4 + 1/\sigma_2^4\right)\right)^k \mathbf{W}_2(\pi^0, \pi_\gamma)$. □

Lemma 32 shows the existence of a invariant distribution $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^d)$ for P_γ and the next lemma specifies this distribution of π_γ.

Lemma 33. *Assume $\gamma \in \left(0, 2(\sigma_1\sigma_2)^4[\bar{\sigma}^2(\sigma_1^4+\sigma_2^4)]^{-1}\right)$. Then, the stationarity distribution π_γ is Gaussian with parameters given by*

$$\mathrm{m}_{(\gamma)} = \frac{\bar{\mathrm{m}} - \frac{\gamma\bar{\sigma}^2}{2}\left(\frac{\mu_1}{\sigma_1^4} + \frac{\mu_2}{\sigma_2^4}\right)}{1 - \frac{\gamma\bar{\sigma}^2}{2}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right)} , \qquad \sigma_{(\gamma)}^2 = \frac{\bar{\sigma}^2 - \frac{\gamma}{2} + \frac{\gamma^2}{8\sigma^2}}{1 - \frac{\gamma}{2}\left(\frac{\bar{\sigma}^2}{\sigma_1^4} + \frac{\bar{\sigma}^2}{\sigma_2^4}\right) - \frac{\gamma}{2}\left(\frac{1}{\bar{\sigma}} - \frac{\gamma}{2}\left(\frac{\bar{\sigma}}{\sigma_1^4} + \frac{\bar{\sigma}}{\sigma_2^4}\right)\right)^2} .$$

Proof. First, let $k \in \mathbb{N}$ be fixed and introduce

$$\alpha = 1 - \frac{\gamma}{\bar{\sigma}^2} + \frac{\gamma^2}{2}\left(\frac{1}{\sigma_1^4} + \frac{1}{\sigma_2^4}\right) , \qquad \beta = \frac{\gamma\bar{\mathrm{m}}}{\bar{\sigma}^2} - \frac{\gamma^2}{2}\left(\frac{\mu_1}{\sigma_1^4} + \frac{\mu_2}{\sigma_2^4}\right) ,$$

$$\tilde{Z}_k = \left(1 - \frac{\gamma}{2\bar{\sigma}^2}\right) Z_{2k-1} + Z_{2k} .$$

Moreover, consider $(A_l)_{l\in\mathbb{N}}$ the stochastic process following (147) and initialized at π_γ. By induction, we know that

$$A_k = \alpha^k A_0 + \beta\sum_{l=0}^{k-1}\alpha^l + \sqrt{\gamma}\sum_{l=0}^{k-1}\alpha^{k-l-1}\tilde{Z}_l . \tag{149}$$

Since A_k is distributed according to $\pi_\gamma P_\gamma^k$, we have that A_k follows π_γ. Denote ν_γ^k the distribution of $\sqrt{\gamma}\sum_{l=0}^{k-1}\alpha^{k-l-1}\tilde{Z}_l - \beta\sum_{l=0}^{k-1}\alpha^l$, combining (149) with the definition of the Wasserstein, we have

$$\mathbf{W}_2^2\left(\pi_\gamma, \nu_\gamma^k\right) \leq \mathbb{E}\left[\left\|A_k - \sqrt{\gamma}\sum_{l=0}^{k-1}\alpha^{k-l-1}\tilde{Z}_l - \beta\sum_{l=0}^{k-1}\alpha^l\right\|^2\right] = \alpha^{2k}\mathbb{E}\left[\|A_0\|^2\right] . \tag{150}$$

Since A_0 is distributed according to π_γ belonging to $\mathcal{P}_2(\mathbb{R}^d)$, we deduce that $\mathbb{E}[\|A_0\|^2] < \infty$. Consequently, (150) implies that $(\nu_\gamma^k)_{k\in\mathbb{N}}$ converges to π_γ, but using the fact that $(\nu_\gamma^k)_{k\in\mathbb{N}}$ converges to a Gaussian distribution, we obtain by uniqueness of the limit in metric space $(\mathcal{P}_2(\mathbb{R}^d), \mathbf{W}_2)$ that π_γ is a Gaussian distribution. Recalling that $\mathrm{m}_{(\gamma)}$ denotes the expectation of the random variable distributed according to π_γ, using (147) at stationarity yields

$$\mathrm{m}_{(\gamma)} = \mathrm{m}_{(\gamma)} - \frac{\gamma}{\bar{\sigma}^2}\left(\mathrm{m}_{(\gamma)} - \bar{\mathrm{m}}\right) + \frac{\gamma^2}{2}\left(\frac{\mathrm{m}_{(\gamma)} - \mu_1}{\sigma_1^4} - \frac{\mathrm{m}_{(\gamma)} - \mu_2}{\sigma_2^4}\right)$$

Thus, we deduce that

$$\mathrm{m}_{(\gamma)} = \frac{\bar{\mathrm{m}} - (\gamma\bar{\sigma}^2/2)\left(\mu_1/\sigma_1^4 + \mu_2/\sigma_2^4\right)}{1 - (\gamma\bar{\sigma}^2/2)\left(1/\sigma_1^4 + 1/\sigma_2^4\right)} .$$

In addition, we can obtain the standard deviation $\sigma_{(\gamma)}$ of π_γ since we have

$$\begin{aligned}\operatorname{Var}\left(\beta\sum_{l=0}^{k-1}\alpha^l+\sqrt{\gamma}\sum_{l=0}^{k-1}\alpha^{k-l-1}\tilde{Z}_l\right)&=\gamma\operatorname{Var}\left(\sum_{l=0}^{k-1}\alpha^{k-l-1}\tilde{Z}_l\right)=\frac{\gamma(1-\alpha^{2k})}{1-\alpha^2}\operatorname{Var}(\tilde{Z}_0)\\&\xrightarrow[k\to\infty]{}\frac{\gamma\operatorname{Var}(\tilde{Z}_0)}{1-\alpha^2}\\&=\frac{\gamma\left(2-\frac{\gamma}{\bar{\sigma}^2}+\frac{\gamma^2}{4\bar{\sigma}^4}\right)}{1-\left(1-\frac{\gamma}{\bar{\sigma}^2}+\frac{\gamma^2}{2}\left(\frac{1}{\sigma_1^4}+\frac{1}{\sigma_2^4}\right)\right)^2}\\&=\frac{1-\frac{\gamma}{2\bar{\sigma}^2}+\frac{\gamma^2}{8\bar{\sigma}^4}}{\frac{1}{\bar{\sigma}^2}-\frac{\gamma}{2}\left(\frac{1}{\sigma_1^4}+\frac{1}{\sigma_2^4}\right)-\frac{\gamma}{2}\left(\frac{1}{\bar{\sigma}^2}-\frac{\gamma}{2}\left(\frac{1}{\sigma_1^4}+\frac{1}{\sigma_2^4}\right)\right)^2}\,.\end{aligned}$$

□

Theorem 34. *Assume* $\gamma\in\left(0,2(\sigma_1\sigma_2)^4[\bar{\sigma}^2(\sigma_1^4+\sigma_2^4)]^{-1}\right)$. *Then, the Wasserstein distance between the stationnary distribution* π_γ *and the target* π *of* FALD *is lower bounded as*

$$\mathbf{W}_2\left(\pi_\gamma,\pi\right)\geq\frac{\gamma}{2}\left|\mu_1-\mu_2\right|\left|\frac{\bar{\sigma}^2}{\sigma_1^2}-\frac{\bar{\sigma}^2}{\sigma_2^2}\right|\,.$$

Proof. Based on Lemma 33, we know that π_γ is Gaussian with parameters $(\mathrm{m}_{(\gamma)},\sigma^2_{(\gamma)})$ and using that π is Gaussian too with parameters $(\bar{\mathrm{m}},\bar{\sigma}^2)$ given in (146), we have that

$$\begin{aligned}\mathbf{W}_2^2\left(\pi_\gamma,\pi\right)=\left(\mathrm{m}_{(\gamma)}-\bar{\mathrm{m}}\right)^2+\left(\sigma_{(\gamma)}-\bar{\sigma}\right)^2&\geq\frac{\gamma^2\bar{\sigma}^4}{4}\left|\left(\frac{1}{\sigma_1^4}+\frac{1}{\sigma_2^4}\right)\bar{\mathrm{m}}-\frac{\mu_1}{\sigma_1^4}-\frac{\mu_2}{\sigma_2^4}\right|^2\\&=\frac{\gamma^2\bar{\sigma}^4\left(\mu_1-\mu_2\right)^2}{4}\left(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}\right)^2\,.\end{aligned}$$

□

9 Analysis of the complexity and communication cost

In this section, we study the optimal choices of k,γ when p_{c} is fixed. For $c_0,c_1,c_2\geq 0$ fixed, we consider the following optimization problem:

$$\left\{\begin{array}{l}\min_{k\in\mathbb{N}^\star,\gamma>0}\left\{k\right\}\\\text{Subject to }\left\{c_0\exp\left(-8k\gamma/m\right)+c_1\gamma+c_2\gamma^2\leq\epsilon^2\right\}\end{array}\right.\,.$$

Using that the constraint must be saturated at the optimum (which can be proved), we can write k as a function of γ. Hence, the problem becomes

$$\left\{\begin{array}{l}\min_{k,\gamma}\left\{\frac{8}{\gamma m}\log\left(\frac{c_0}{\epsilon^2-c_1\gamma-c_2\gamma^2}\right)\right\}\\\text{Subject to }0<\gamma\text{ and }\epsilon^2-c_1\gamma-c_2\gamma^2>0\end{array}\right.\,.\tag{151}$$

Let us introduce $x\in\mathbb{R}_+^*$, defined by $x=\epsilon^{-2}\gamma$ and let $\tilde{c}_2=\epsilon^2c_2$. We can rewrite (151) as

$$\left\{\begin{array}{l}\min_{k,x}\left\{\frac{8}{\epsilon^2mx}\log\left(\frac{c_0}{\epsilon^2(1-c_1x-\tilde{c}_2x^2)}\right)\right\}\\\text{Subject to }0<x\text{ and }1-c_1x-\tilde{c}_2x^2>0\end{array}\right.\,.\tag{152}$$

Consider $\mu=-c_1/(2\tilde{c}_2)$, $\sigma=\sqrt{c_1^2/(4\tilde{c}_2^2)+1/\tilde{c}_2}$, and denote $z=(x-\mu)/\sigma$. Since $x=\mu+z\sigma$, we can verify that $1-c_1x-\tilde{c}_2x^2=\tilde{c}_2\sigma^2(1-z^2)$. Hence, (153) is equivalent to

$$\left\{\begin{array}{l}\min_{k,\gamma}\left\{\frac{8}{\epsilon^2m(\mu+z\sigma)}\log\left(\frac{c_0}{\epsilon^2\tilde{c}_2\sigma^2(1-z^2)}\right)\right\}\\\text{Subject to }-\mu/\sigma<z<1\end{array}\right.\,.\tag{153}$$

According to the intermediate value theorem, we have the existence of z_ϵ (not necessarily unique, but we can consider one of the solutions) such that

$$z_\epsilon = \underset{-\mu/\sigma<z<1}{\arg\max} \left\{ \frac{\log(1-z^2)}{\mu + z\sigma} \right\} .$$

Thus, the solution is

$$\begin{aligned} \gamma_\epsilon &= \epsilon^2 \times \frac{z_\epsilon^2 + (4\epsilon^2 c_2)^{-1}(z_\epsilon^2 - 1)c_1^2}{c_1/2 + z_\epsilon\sqrt{4^{-1}\, c_1^2 + \epsilon^2 c_2}} , \\ K_\epsilon &= \frac{8\big(c_1/2 + z_\epsilon\sqrt{4^{-1}\, c_1^2 + \epsilon^2 c_2}\big)}{\epsilon^2 m \,(z_\epsilon^2 + (4\epsilon^2 c_2)^{-1}(z_\epsilon^2 - 1)c_1^2)} \log\left(\frac{c_0}{\epsilon^2 (c_1^2/4 + \epsilon^2 c_2)^{1/2}(1 - z_\epsilon^2)} \right) . \end{aligned}$$

FALD. According to the Theorem 1, we have

$$\begin{cases} c_0 = \mathsf{I}(\mu_0) \\ c_1 = \mathsf{V}_\pi + (1 - 1_{\mathbf{HX}1})\,\mathsf{J}/b + (1-\tau)(1 - b^{-1})d/p_{\mathrm{c}} \\ c_2 = 1_{\mathbf{HX}1}\mathsf{J}/b + (1 - p_{\mathrm{c}})\left\{\mathsf{H} + p_{\mathrm{c}}\mathsf{V}_\epsilon + d/b\right\}/p_{\mathrm{c}}^2 \end{cases} .$$

If $c_1 > 0$, define $w = \epsilon^2 c_2/c_1^2$. For $\epsilon \in (0, c_1/\sqrt{2c_2}]$, we have $0 < w \leq 1/2$. Consider $z = 1 - w$, we get that

$$\left(\frac{\mu}{\sigma}\right)^2 = \frac{1}{1 + 4\epsilon^2 c_2/c_1^2} < \frac{1}{1 + 2w} \leq 1 - w \leq 1 - 2w + w^2 = z^2 < 1 .$$

Hence, the previous inequalities show that $-\mu/\sigma < z < 1$, and for this choice

$$\frac{c_1/2 + z\sqrt{4^{-1}c_1^2 + \epsilon^2 c_2}}{z^2 + (4\epsilon^2 c_2)^{-1}(z^2 - 1)c_1^2} \leq \frac{c_1 + \epsilon(1-w)\sqrt{c_2}}{7/8 + (w - 2 + 1/64)w} .$$

Thus, for any $\epsilon \in (0, c_1(2\sqrt{c_2})^{-1}]$, we deduce that $w < 1/4$. Therefore, we have shown that $K_\epsilon = \tilde{\mathrm{O}}((\epsilon^2 m)^{-1}(c_1 + \epsilon\sqrt{c_2}))$. Moreover, this result is immediately valid when $c_1 = 0$ since $z_\epsilon = \arg\max_{0<z<1}\{z^{-1}\log(1 - z^2)\}$. Furthermore, when $p_{\mathrm{c},\epsilon} \downarrow 0^+$, $p_{\mathrm{c},\epsilon}K_\epsilon = \tilde{\mathrm{O}}((\epsilon m)^{-1}\sqrt{b^{-1}\mathsf{J}})$ as stressed in the main paper.

VR-FALD*. Using Theorem 3, we obtain

$$\begin{cases} c_0 = \mathsf{I}^{\mathrm{Vr}\star}(\mu_0) \\ c_1 = (1 - 1_{\mathbf{HX}1})\,\mathsf{J}/b + (1-\tau)(1 - b^{-1})d/p_{\mathrm{c}} \\ c_2 = 1_{\mathbf{HX}1}\mathsf{J}/b + (1 - p_{\mathrm{c}})\left\{p_{\mathrm{c}}\mathsf{V}_\epsilon + d/b\right\}/p_{\mathrm{c}}^2 \end{cases} .$$

When assuming **HX**1 and $\tau = 1$, we have $c_1 = 0$. Hence, $z_\epsilon = \arg\max_{0<z<1}\{z^{-1}\log(1 - z^2)\}$ and therefore

$$K_\epsilon = \frac{8\sqrt{c_2}}{\epsilon m z_\epsilon} \log\left(\frac{c_0}{\epsilon^3\sqrt{c_2}(1 - z_\epsilon^2)} \right) .$$

When $p_{\mathrm{c},\epsilon} \downarrow 0^+$, the minimum number of communications becomes $p_{\mathrm{c},\epsilon}K_\epsilon = \tilde{O}(\epsilon^{-1}\sqrt{b^{-1}d})$. Finally, setting $p_{\mathrm{c},\epsilon} = 1$ gives $K_\epsilon = \tilde{O}(\epsilon^{-1}\sqrt{b^{-1}\mathsf{J} + b^{-1}\omega d})$.

Table 5: Complexity and communication settings of Figure 1.

PARAMETER	d	m	ω	H	J	V_π	$\mathsf{V}_\star$
VALUE	10	1	10	100	20	10	30

10 Numerical experiments

10.1 Gaussian example

In this first experiment, we consider $b = 100$ clients associated with potentials: $\forall i \in [b]$, $U^i : x \in \mathbb{R}^d \mapsto (1/2)(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i)$ in dimension $d = 20$. In this particular case, we know, that the posterior distribution $\pi \propto \exp(-\sum_{i=1}^b U^i)$ is Gaussian with mean $x_\star = \sum_{i=1}^b (\Sigma_\star \Sigma_i^{-1} \mu_i)$ and covariance $\Sigma_\star = (\sum_{i=1}^b \Sigma_i^{-1})^{-1}$. Also, we have a close formula to calculate $\int \|x - x_\star\|^2 \mathrm{d}\pi(x)$, since this quantity is equal to $\mathrm{Tr}(\Sigma_\star)$. To speed up the calculations, we initialize all chains at $x_\star$, we discard the first 10% of the samples and keep all others. Moreover, we consider the step size $\bar{\gamma} = 2[\lambda_{\text{min}}(\Sigma_\star^{-1}) + \lambda_{\text{max}}(\Sigma_\star^{-1})]^{-1}$ for Langevin Monte Carlo (Dalalyan and Karagulyan, 2019; Durmus and Moulines, 2019), and we run the algorithms for the step sizes $\gamma \in \{\frac{p_\mathrm{c}\bar{\gamma}}{2}, \frac{p_\mathrm{c}\bar{\gamma}}{5}, \frac{p_\mathrm{c}\bar{\gamma}}{10}\}$ associated with $p_\mathrm{c} \in \{\frac{1}{5}, \frac{1}{10}, \frac{1}{20}\}$. We set the probability of updating the control variates $q_\mathrm{c} = p_\mathrm{c}$ so as not to increase the communication cost too much. We also consider the two extreme values of the parameter $\tau \in \{0, 1\}$ to determine whether it is preferable to have independent Gaussian noise on each client or if it is better to have a common one.

10.2 Bayesian Logistic Regression

The second experiment is performed on the Titanic dataset, which is in the public domain and licensed under the Commons Public Domain Dedication License (PDDL-1.0). We distribute this dataset heterogeneously across $b = 10$ clients by drawing a Dirichlet random variable for each label on the standard $b - 1$ simplex. Since the sum of the coordinates of these random variables equals 1, each coordinate indicates the fraction of labels to be distributed to each client. To have access to ground truth, we also implement Langevin Stochastic Dynamics (LSD). We compute $K = 250000$ iterations, each time considering a burn-in period of length 10% initialized with a warm start provided by SGD. The ith client uses its local dataset $\{(z_{ij}, o_{ij}) \in \mathbb{R}^4 \times \{0, 1\} : j \in [N_i]\}$ to calculate the local potential $U^i(x) = \sum_{j=1}^{N_i} [o_{ij} \log(1 + \exp(-z_{ij}^T x)) + (1 - o_{ij}) \log(1 + \exp(z_{ij}^T x))] + \lambda \|x\|^2$, where $\lambda = 1$ is associated with the Gaussian prior. Denote $\mathrm{Z}_{\text{train}}$ the matrix whose lines are the covariates z_{ij}^T, and write $\Sigma = \mathrm{Z}_{\text{train}}^T \mathrm{Z}_{\text{train}}$. We run the algorithms with mini-batches of size $n_i = 1$; a step size $\gamma = 2[\lambda_{\text{min}}(\Sigma) + \lambda_{\text{max}}(\Sigma)]^{-1}$ for FALD, VR-FALD* and equal to γ/b for LSD with thinning inversely proportional to the step size. Moreover, we consider a communication probability of $p_\mathrm{c} = 1/20$ and clients update their control variables with probability $q_\mathrm{c} = p_\mathrm{c}$. Finally, to evaluate the obtained results, we consider the accuracy, agreement, and total variation, as well as the calibration results such as ECE, BS, and NLL, which are described below.

Accuracy. Based on samples from the approximate posterior distribution, we compute the minimum mean squared estimator (*i.e.*, which corresponds to the posterior mean) and use it to make predictions for the test dataset. The *Accuracy* metric corresponds to the percentage of well-predicted labels.

Agreement. Let p_{ref} and p denote the predictive densities associated with 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 proportion of test data points for which the top-1 predictions of p_{ref} and p, *i.e.*

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

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

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

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 each $m \in [M]$ by $\mathrm{B}_{\mathrm{m}} = \{(x, y) \in \mathrm{D}_{\text{test}} : p(y_{\text{pred}}(x)|x) \in$

$](m-1)/M, m/M]\}$. As in the previous work of Ovadia et al. (2019), we denote the model accuracy on $\mathrm{B_m}$ by

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

and define the confidence on $\mathrm{B_m}$ by

$$\operatorname{conf}(\mathrm{B_m}) = \frac{1}{|\mathrm{B_m}|} \sum_{(x,y)\in \mathrm{B_m}} p(y_{\mathrm{pred}}(x)|x)\,.$$

As emphasized in Guo et al. (2017), for any $m \in [M]$ the accuracy $\operatorname{acc}(\mathrm{B_m})$ 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 is defined by

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

and is an estimator of

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

Thus, the 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 confidence of the model in its predictions and is defined by

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

Normalized 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 ability of the model to predict good labels with high probability.

Highest posterior density (HPD). Under the Bayesian paradigm, we are interested in quantifying uncertainty by estimating the regions of high probability. For all $\alpha \in (0,1)$, we run each algorithm to estimate $\eta_\alpha^{\mathrm{algo}} > 0$ such that $\int_{R_\alpha} \pi(x)\mathrm{d}x = 1-\alpha$, where $\mathcal{R}_\alpha = \{x \in \mathbb{R}^d : \pi(x) \geq \exp(-\eta_\alpha^{\mathrm{algo}})\}$. Then we define the relative HPD error as $|\eta_\alpha^{\mathrm{algo}}/\eta_\alpha^{\mathrm{LSD}} - 1|$, where $\eta_\alpha^{\mathrm{LSD}}$ is estimated based on the samples drawn with the Langevin Stochastic Dynamics method.

10.3 Bayesian Neural Network: MNIST

To investigate the behavior of the proposed algorithms in a highly non-convex setting, we perform a first Deep Learning experiment on the MNIST dataset (Deng, 2012), which can be publicly downloaded using the torchvision package and is available under the Creative Commons Attribution-Share Alike 3.0 license. To this end, we distribute the entire dataset across $b = 20$ clients in a highly heterogeneous manner to train the LeNet5 neural network (LeCun et al., 1998). The MNIST real-world dataset consists of 70000 grayscale images of size 28×28 associated with the 10 digits. This dataset is divided into two subsets: the training set, which contains 60000 images, and the test set, which consists of the remaining 10000 images. We report the median of the scores with their associated hyperparameters in Table 6. The burn-in corresponds to the number of steps performed before we start storing the samples, and the thinning is the frequency with which we keep the samples. We also consider a Gaussian prior which corresponds to a squared norm regularizer with weight decay. We initialized FSGLD (El Mekkaoui et al., 2021) with a global SGD warm start combined with local SWAG (Maddox et al., 2019) to learn Gaussian conducive gradients.

METHOD	SGLD	pSGLD	FALD	VR-FALD*	FSGLD
Accuracy	99.1 ± 0.1	99.2 ± 0.1	99.1 ± 0.1	99.2 ± 0.1	98.5 ± 0.2
$10^3\times$ECE	6.88 ± 27.07	21.6 ± 11.1	4.07 ± 0.80	4.34 ± 1.26	6.34 ± 1.90
$10^2\times$BS	1.66 ± 1.76	1.45 ± 0.12	1.47 ± 0.45	1.39 ± 0.07	2.39 ± 1.72
$10^2\times$nNLL	3.53 ± 5.08	4.24 ± 1.14	3.06 ± 0.43	3.43 ± 0.37	4.87 ± 0.51
Weight Decay	5	5	5	5	5
Batch Size	64	64	8	8	64
Learning rate	1e-07	1e-08	1e-07	1e-07	1e-08
Local steps	N/A	N/A	20	20	20
Burn-in	100epch.	100epch.	1e04	1e04	1e04
Thinning	1	1	1e03	1e03	1e03
Training	1e03epch.	1e03epch.	1e05it.	1e05it.	1e05it.

Table 6: Performance of Bayesian FL algorithms on MNIST.

METHOD	HMC	SGD	DEEP ENS.	SGLD	FALD	VR-FALD*
Accuracy	89.6 ± 0.25	91.57 ± 0.34	91.68 ± 0.17	89.96 ± 0.72	$\mathbf{92.54} \pm 0.04$	92.03 ± 0.09
Agreement	94.0 ± 0.25	90.99 ± 0.35	91.03 ± 0.43	$\mathbf{92.43} \pm 0.03$	91.53 ± 0.39	91.12 ± 0.39
$10\times$ TV	0.74 ± 0.03	1.45 ± 0.05	1.49 ± 0.05	$\mathbf{1.03} \pm 0.03$	1.42 ± 0.01	1.39 ± 0.01
$10^2\times$ECE	5.9±NA	4.71 ± 1.35	5.44 ± 0.67	4.41 ± 0.37	3.79 ± 0.11	$\mathbf{3.26} \pm 0.09$
$10\times$BS	1.4±NA	1.69 ± 0.11	1.45 ± 0.10	1.53 ± 0.10	$\mathbf{1.16} \pm 0.03$	1.20 ± 0.03
$10\times$nNLL	3.07±NA	3.35 ± 0.70	3.81 ± 0.51	3.15 ± 0.21	2.75 ± 0.04	$\mathbf{2.63} \pm 0.04$

Table 7: Performance of Bayesian FL algorithms on CIFAR10.

10.4 Bayesian Neural Network: CIFAR10

In this last experiment, we consider the more challenging dataset CIFAR10 (Krizhevsky, 2009), which is available under license MIT and contains images of size $(3, 32, 32)$. We used different approaches to sample the weights for the ResNet-20 model (He et al., 2016), which is publicly available in the pytorchcv library. We initialized the algorithms with 10 different parameters using SGD (400 epochs) trained with a OneCycleLR scheduler (Smith and Topin, 2019), and we also use data augmentation with a mini-batch of size 128 and a learning rate of 2e-7. Based on these initializations, we ran 10 chains in parallel for SGLD, FALD, and VR-FALD* with step sizes of 1e-7, 2e-8, 1e-8. We considered 1e4 iterations with only one stored sample every 1e3 iterations (we did not keep the initial weights obtained by SGD to make the predictions). For each chain, we can see that Bayesian model averaging increases the accuracy. To compare the behavior of the mentioned algorithms, we compute the accuracy, the agreement, i.e., the percentage of time the top-1 prediction of an algorithm matches that given by the HMC, and the total variation (TV) between the predictive distribution given by an algorithm with the one associated with the HMC sampler. We also give some classical calibration scores (Guo et al., 2017), such as the expected calibration error (ECE), the Brier score (BS), and the negative log-likelihood (nNLL).