Refined Analysis of Constant Step Size Federated Averaging

Anonymous Author
Anonymous Institution

Abstract

In this paper, we present a novel analysis of FedAvg with constant step size, relying on the Markov property of the underlying process. We demonstrate that the global iterates of the algorithm converge to a stationary distribution and analyze its resulting bias and variance relative to the problem's solution. We provide a first-order expansion of the bias in both homogeneous and heterogeneous settings. Interestingly, this bias decomposes into two distinct components: one that depends solely on stochastic gradient noise and another on client heterogeneity. Finally, we introduce a new algorithm based on the Richardson-Romberg extrapolation technique to mitigate this bias.

1 INTRODUCTION

Federated averaging (FedAvg) [McMahan et al., 2017] has become a cornerstone of federated learning. It allows multiple clients to collaborate on a shared optimization problem without having to exchange their local data directly. While FedAvg has proven practical efficiency in many federated learning scenarios, its convergence can be significantly affected by the heterogeneity of clients. In fact, FedAvg performs several local updates to speed up the training process and reduce communication costs. However, this leads to the *local drift* phenomenon [Karimireddy et al., 2020]: as the number of local steps increases, each client tends to converge to an optimum that matches its local data, rather than the global optimum of the entire coalition, leading to biases in the resulting conclusions.

Several methods have been proposed to mitigate the bias of FedAvg caused by the heterogeneity across clients. These approaches typically fall into two categories: control variates-based methods [Karimireddy et al., 2020, Mishchenko et al., 2022, Malinovsky et al., 2022] and primal-dual proximal approaches [Sadiev et al., 2022, Grudzień et al., 2023]. These techniques allow for more local steps while complying with lower bounds on the number of communications required for federated learning [Arjevani and Shamir, 2015].

Recently, it was found that FedAvg suffers from a second type of bias known as *iterate bias.* This bias appeared in multiple analyse of federated averaging Khaled et al. [2020], Glasgow et al. [2022], Wang et al. [2024], as an additional term that scales with the variance of the gradients and the number of local steps. This bias arises from the use of local stochastic gradients, similar to what was observed in previous work on SGD [Pflug, 1986, Dieuleveut et al., 2020]. In this paper, we propose a new analysis of FedAvg for strongly convex and smooth local objective functions, which provides new insights into the convergence and allows us to design a simple mechanism that reduces the bias. Our main contributions are as follows:

- We start with a refined analysis of FedAvg, with more than one local step, in the deterministic setting, where the local gradients are exact. We show that under these conditions and in the presence of client heterogeneity, FedAvg suffers from a bias: it does not converge to the global optimum of the coalition, but rather to a point that lies in an $O(\gamma H)$ neighborhood of this optimum, where γ is the step size and H the number of local updates. We derive a first-order expansion in γH of this bias, showing that the local drift phenomenon is not due to stochasticity.

- Second, we extend this analysis to FedAvg with *stochastic* gradients. Exploiting the Markov property of FedAvg's iterates and assumptions similar to those in Dieuleveut et al. [2020], we show that this sequence admits a unique stationary distribution and converges exponentially fast in the second-order Wasserstein distance. This allows us to provide a sharp analysis of FedAvg, establish-

Preliminary work. Under review by AISTATS 2025. Do not distribute.

- ing an explicit first-order expansion of its bias in $O(\gamma H)$. We show that the bias can be decomposed into two terms: one depending solely on the covariance of the stochastic gradients, and another one depending solely on client heterogeneity. The scaling of these terms is influenced by both *gradient* and *Hessian* dissimilarity, which extends existing results. Furthermore, our analysis does not rely on restrictive assumptions on the gradient noise, allowing for a polynomially growing gradient variance.
- We suggest a novel approach for mitigating bias, addressing both heterogeneity and stochastic noise using the Richardson-Romberg extrapolation procedure. In contrast to SCAFFOLD, this method does not use control variates, and thus does not incur additional memory cost at the client level. To the best of our knowledge, this is the first method capable of reducing the stochastic bias inherent in FEDAVG. We validate this approach numerically, demonstrating that it outperforms existing bias-correction techniques, such as SCAFFOLD, particularly in scenarios where gradient variance is substantial.

Notation. In this paper, we denote by $\langle\cdot,\cdot\rangle$ the euclidean dot product, and $\|\cdot\|$ the associated norm. Vectors are column vectors, and we denote Id the identity matrix, and $\mathbf{1}_n$ the vector of size n with all components equal to 1. For a three times differentiable function f and $i \in \{1,2,3\}$ we denote $\nabla^i f$ its i-th order derivatives. For a sequence of matrices $M_1, \dots M_k$, we denote the product by $\prod_{\ell=1}^{k} M_\ell = M_k M_{k-1} \cdots M_1$. For two matrices A, B, we denote $A \otimes B$ the linear operator $M \mapsto AMB$, where A, B and M are matrices of compatible sizes. Furthermore, we denote $M^{\otimes k}$ the k^{th} tensor power of a tensor M. Let $\mathcal{B}(\mathbb{R}^d)$ be the Borel σ-field of $\mathbb{R}^d$. For two probability measures λ, ν over $\mathbb{R}^d$ with finite second moment, we define the second-order Wasserstein distance as $\mathbf{W}_2^2(\lambda,\nu) = \inf_{\xi \in \Pi(\lambda,\nu)} \int \|\theta - \vartheta\|^2 \xi(\mathrm{d}\theta, \mathrm{d}\vartheta)$, where $\Pi(\lambda,\nu)$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ such that $\xi(\mathsf{A} \times \mathbb{R}^d) = \lambda(\mathsf{A})$ and $\xi(\mathbb{R}^d \times \mathsf{A}) = \nu(\mathsf{A})$ for all $\mathsf{A} \in \mathcal{B}(\mathbb{R}^d)$.

2 PRELIMINARIES

Algorithm 1 FEDAVG

Input: step size $\gamma > 0$, initial $\theta_0 \in \mathbb{R}^d$, number of rounds $T > 0$, number of clients $N > 0$, number of local steps $H > 0$

```
for $t = 0$ to $T - 1$ do
    for $c = 1$ to $N$ do
        Initialize $\theta_{c,t}^0 = \theta_t$
        for $h = 0$ to $H - 1$ do
            Receive random state $Z_{c,t}^{h+1}$
            Set $\theta_{c,t}^{h+1} = \theta_{c,t}^h - \gamma \nabla F_c^{Z_{c,t}^{h+1}}(\theta_{c,t}^h)$
        end for
    end for
    Average: $\theta_{t+1} = \frac{1}{N} \sum_{c=1}^{N} \theta_{c,t}^H$
end for
Return: $\theta_T$
```

Federated Averaging. We study the federated stochastic optimization problem

$$\theta^\star \in \min_{\theta \in \mathbb{R}^d} f(\theta) = \frac{1}{N} \sum_{c=1}^{N} f_c(\theta) \;,\; f_c(\theta) = \mathbb{E}\left[F_c^{Z_c}(\theta)\right] \;, \quad (1)$$

where for each $c \in \{1, \dots, N\}$, Z_c is a random variable with the distribution ξ_c, which takes on values in a measurable set $(\mathsf{Z}, \mathcal{Z})$, and $(z, \theta) \mapsto F_c^z(\theta)$ are measurable functions. To solve (1), we consider N clients indexed by $c \in 1, \dots, N$, and assume that each client c has access to its own function f_c through stochastic sampling of $F_c^{Z_c}$. In this case, FEDAVG solves the problem (1) by performing local stochastic gradient updates on each client. These local iterations are sent at regular intervals to a central server, which aggregates them by calculating the average and sends this updated estimate back to the clients. The clients then restart their local updates based on this new estimate. Starting from a common initial point θ_0 shared by all clients and the server, in each round $t \in \mathbb{N}^*$ the server sends its current estimate θ_t to each client $c \in 1, \dots, N$. Then each client c starts with this updated value and sets $\theta_{c,t}^0 = \theta_t$, and performs $H \in \mathbb{N}^*$ local updates: for $h \in \{0, \dots, H-1\}$,

$$\theta_{c,t}^{h+1} = \theta_{c,t}^h - \gamma \nabla F_c^{Z_{c,t}^h}(\theta_{c,t}^h) \;,$$

where $\gamma > 0$ is a common step size shared by the clients, and $\{Z_{\tilde{c},\tilde{t}}^{\tilde{h}} : \tilde{c} \in \{1, \dots, N\}, \tilde{h} \in \{0, \dots, H-1\}, \tilde{t} \in \mathbb{N}\}$ are independent random variables, so that for each $\tilde{c} \in \{1, \dots, N\}$, $\tilde{h} \in \{0, \dots, H-1\}$ and $\tilde{t} \in \mathbb{N}$, $Z_{\tilde{c},\tilde{t}}^{\tilde{h}}$ has distribution ξ_c. Once the local updates are complete, each client sends its last iteration $\theta_{c,t}^H$ to the central server, which updates the global parameters:

$$\theta_{t+1} = N^{-1} \sum_{c=1}^{N} \theta_{c,t}^H \;. \quad (2)$$

We give the pseudocode of FEDAVG in Algorithm 1. The main challenge with this algorithm is that using local updates introduces bias when the clients' local functions are heterogeneous, a phenomenon that we formally characterize in Section 4 and Section 5.

Assumptions. Throughout this paper, we consider the following assumptions.

A 1 (Regularity)**.** *For every $c \in \{1, \dots, N\}$, the function f_c is three times differentiable. In addition, suppose that for every $c \in \{1, \dots, N\}$:*

(a) *The function f_c is μ-strongly convex with $\mu > 0$, that is $\nabla^2 f_c(\theta) \succcurlyeq \mu \mathrm{Id}$.*

(b) *The function F_c is L-expected-smooth, that is, there exists $L \geq 0$ such that for all $\theta, \vartheta \in \mathbb{R}^d$,*

$$\mathbb{E}[\|\nabla F_c^{Z_c}(\theta) - \nabla F_c^{Z_c}(\vartheta)^2\|] \leq \\ 2L\langle \theta - \vartheta, \nabla f_c(\theta) - \nabla f_c(\vartheta)\rangle \ ,$$

where Z_c has a distribution ξ_c.

(c) *It holds that $\nabla^2 f_c(\theta) \preccurlyeq L\mathrm{Id}$.*

Note that under A 1, $N^{-1}\sum_{c=1}^N f_c$ is μ-strongly convex and therefore has a unique minimizer $\theta^\star$, and the operator $\mathrm{Id} \otimes \nabla^2 f(\theta^\star) + \nabla^2 f(\theta^\star) \otimes \mathrm{Id}$ is invertible.

A 2 (Heterogeneity Measure)**.** *There exist $\zeta_{\star,1}, \zeta_{\star,2} > 0$ such that for any $c \in \{1, \dots, N\}$, with $\theta^\star$ as in* (1)*,*

$$\frac{1}{N}\sum_{c=1}^{N}\|\nabla^i f_c(\theta^\star) - \nabla^i f_c(\theta^\star)\|^2 \leq \zeta_{\star,i}^2 \ \textit{ for } i \in \{1, 2\} \ ,$$

where we recall that $\nabla f(\theta^\star) = 0$.

Note that when the solution of (1) is unique, which is notably the case under A1, this assumption also holds.

3 RELATED WORK

Analysis of Federated Averaging. FedAvg was first introduced by McMahan et al. [2017]. Since then, numerous analyses have been developed. Initial studies primarily relied on assumptions of homogeneity [Stich, 2019, Wang and Joshi, 2018, Haddadpour and Mahdavi, 2019, Yu et al., 2019b, Wang and Joshi, 2018, Li et al., 2019]. Over time, various heterogeneity measures have been proposed to derive upper bounds on the error of FedAvg. Among the most common assumptions is *bounded gradient dissimilarity* [Yu et al., 2019a, Khaled et al., 2020, Karimireddy et al., 2020, Reddi et al., 2021, Zindari et al., 2023, Crawshaw et al., 2024]. Other measures include second-order similarity [Arjevani and Shamir, 2015, Khaled et al., 2020], relaxed first-order heterogeneity [Glasgow et al., 2022], and average drift at the optimum [Wang et al., 2024, Patel et al., 2023]. It has also been demonstrated that FedAvg can achieve linear speed-up in the number of clients [Yang et al., 2021, Qu et al., 2021, 2023].

Correcting Heterogeneity Bias. A first approach for addressing heterogeneity is based on control variates, pioneered by the Scaffold algorithm [Karimireddy et al., 2020]. Mishchenko et al. [2022] later demonstrated that Scaffold effectively accelerates training, and since then, other control variates schemes have been developed [Condat and Richtárik, 2022, Malinovsky et al., 2022, Condat et al., 2022, Grudzień et al., 2023, Mangold et al., 2024]. In addition, a class of algorithms relying on dual-primal approaches has been proposed to address heterogeneity [Sadiev et al., 2022, Grudzień et al., 2023]. While both approaches allow for more local training steps and effectively correct heterogeneity bias, they do not address the bias caused by stochasticity when using fixed steps ize.

Stochastic Bias. Even in the single-client setting, SGD with fixed step size have been shown to exhibit bias [Lan, 2012, Défossez and Bach, 2015, Dieuleveut and Bach, 2016, Chee and Toulis, 2017]. Dieuleveut et al. [2020] proposed framing SGD iterates with a constant step size as a Markov chain, drawing connections to established results in stochastic processes [Pflug, 1986]. Stochastic bias has also been observed in the analysis of federated learning methods. For instance, Khaled et al. [2020] identified this bias in their bounds on client drift, and similar observations were made in the convergence analyses of Glasgow et al. [2022], Wang et al. [2024], which compared SGD's iterates to those of deterministic gradient descent. In this work, we investigate the iterate bias of FedAvg, demonstrating that the stationary distribution of SGD's iterates is inherently biased.

Richardson-Romberg. The Richardson-Romberg extrapolation technique, originally introduced by Richardson [1911], is a classical method in numerical analysis. This approach has been widely applied across various fields, including time-varying autoregressive processes [Moulines et al., 2005], data science [Bach, 2021], and many others [Stoer and Bulirsch, 2013]. Specifically, it has been utilized in the context of SGD by Dieuleveut et al. [2020] and Sheshukova et al. [2024]. In this work, we extend these ideas to the federated learning setting, demonstrating that this form of extrapolation effectively mitigates both heterogeneity and stochastic bias.

4 DETERMINISTIC FEDAVG

In this section, we present a new analysis of FedAvg with deterministic gradients (FedAvg-D), where $F_c^z = f_c$ for all $c \in \{1, \dots, N\}$ and $z \in \mathsf{Z}$. This analysis highlights the core philosophy of the method developed in this paper. Unlike previous analyses, we demonstrate that FedAvg-D converges to a point $\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}$ that differs from the optimal solution $\theta^\star$. We then provide an explicit expression for the distance between these two points, allowing us to establish tight upper bounds on the bias of FedAvg-D.

In the FedAvg-D setting, we define the local updates of the client c by induction, starting from the point $\theta \in \mathbb{R}^d$:

$$\mathsf{T}_c^{(\gamma,h+1)}(\theta) \triangleq (\mathrm{Id}-\gamma\nabla f^{(c)})(\mathsf{T}_c^{(\gamma,h)}(\theta)) \ , \ \mathsf{T}_c^{(\gamma,0)}(\theta) \triangleq \theta \ ,$$

where $h \in \{0, \ldots, H-1\}$. The global updates from (2) can thus be rewritten as

$$\mathsf{T}^{(\gamma,H)}(\theta) \triangleq \frac{1}{N}\sum_{c=1}^{N} \mathsf{T}_c^{(\gamma,H)}(\theta) \ ,$$

or, equivalently, we can write $\mathsf{T}^{(\gamma,H)}(\theta) = \theta - \gamma \mathsf{g}^{(\gamma,H)}(\theta)$, with the pseudo-gradient

$$\mathsf{g}^{(\gamma,H)}(\theta) \triangleq \frac{1}{N}\sum_{c=1}^{N}\sum_{h=0}^{H-1} \nabla f_c(\mathsf{T}_c^{(\gamma,h)}(\theta)) \ .$$

First, we show that FedAvg-D with deterministic updates converges to a fixed point of $\mathsf{T}^{(\gamma,H)}$.

Proposition 1 (Stationary Point of FedAvg-D)**.** *Assume A 1. Then for all $H > 0$ and $\gamma \le 1/L$,* FedAvg-D *converges to a unique point $\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}$ that satisfies $\mathsf{T}^{(\gamma,H)}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) = \bar{\theta}_{\mathrm{det}}^{(\gamma,H)}$ and $\mathsf{g}^{(\gamma,H)}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) = 0$. Moreover, the iterates of* FedAvg-D *satisfy*

$$\|\theta_t - \bar{\theta}_{\mathrm{det}}^{(\gamma,H)}\|^2 \le (1-\gamma\mu)^{Ht}\|\theta_0 - \bar{\theta}_{\mathrm{det}}^{(\gamma,H)}\|^2 \ .$$

To prove Proposition 1, we use the fact that local updates are contractive; see details in Appendix A.

To the best of our knowledge, this is the first result to establish the convergence of FedAvg-D to a *stationary point*, rather than merely converging to a neighborhood of $\theta^\star$. To characterize this stationary point, we derive an explicit expression for the bias $\bar{\theta}_{\mathrm{det}}^{(\gamma,H)} - \theta^\star$ of FedAvg. We rely on the following matrices:

$$\bar{D}_c^{(\bar{\theta}_{c,h},\theta^\star)} = \int_0^1 \nabla^2 f_c(u\,\bar{\theta}_{c,h} + (1-u)\theta^\star)\mathrm{d}u \ , \qquad (3)$$

where $\bar{\theta}_{c,h} = \mathsf{T}_c^{(\gamma,h)}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)})$. Based on (3), we define the following matrices, that express the update of the error when starting from the point $\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}$:

$$\mathcal{F}_c^\star = \prod_{\ell=0}^{H-1}\Big(\mathrm{Id} - \gamma \bar{D}_c^{(\bar{\theta}_{c,h},\theta^\star)}\Big) \ , \ \mathcal{F}^\star = \frac{1}{N}\sum_{c=1}^{N}\mathcal{F}_c^\star \ . \qquad (4)$$

We now provide an expression and an upper bound on the bias of FedAvg-D.

Proposition 2 (Bias of FedAvg-D)**.** *Assume A 1 and A 2. Then for all $H > 0$ and $\gamma \le 1/L$, we have*

$$\bar{\theta}_{\mathrm{det}}^{(\gamma,H)} - \theta^\star = \frac{1}{N}\sum_{c=1}^{N} \Upsilon_{\mathrm{het}}^c \nabla^2 f_c(\theta^\star)^{-1}\nabla f_c(\theta^\star) \ ,$$

where $\Upsilon_{\mathrm{het}}^c = (\mathrm{Id} - \mathcal{F}^\star)^{-1}(\mathrm{Id} - \mathcal{F}_c^\star)$ and $\mathcal{F}_c^\star, \mathcal{F}^\star$ are defined in (4). Furthermore, if $\gamma\mu H \le 1$, then

$$\|\bar{\theta}_{\mathrm{det}}^{(\gamma,H)} - \theta^\star\| \le \gamma(H-1)\mathrm{C}_1 \ , \ \textit{with } \mathrm{C}_1 = L\zeta_{\star,1}/\mu \ .$$

We prove Proposition 2 in Appendix A, using the fact that $\mathsf{T}^{(\gamma,H)}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) = \bar{\theta}_{\mathrm{det}}^{(\gamma,H)}$ from Proposition 1. Importantly, when $H = 1$, the bias of FedAvg completely vanishes, recovering the fact that gradient descent converges. Based on Proposition 2, we further propose a first-order expansion of the bias of FedAvg-D. This highlights that (i) the bias of FedAvg-D solely depends on heterogeneity, and (ii) the convergence bound derived in Proposition 2 is sharp for small values of the product γH.

Theorem 1 (First-Order Bias of FedAvg)**.** *Assume A 1 and A 2. Then for all $H > 0$ and $\gamma \le 1/L \wedge 1/H$, we have*

$$\bar{\theta}_{\mathrm{det}}^{(\gamma,H)} - \theta^\star = \frac{\gamma(H-1)}{2}\mathrm{b_h} + O(\gamma^2 H^2) \ ,$$

where and the heterogeneity bias $\mathrm{b_h}$ is given by

$$\mathrm{b_h} = \frac{1}{N}\sum_{c=1}^{N}\nabla^2 f(\theta^\star)^{-1}(\nabla^2 f_c(\theta^\star) - \nabla^2 f(\theta^\star))\nabla f_c(\theta^\star) \ ,$$

and the explicit expression of the reminder term $O(\gamma^2 H^2)$ is given in Appendix A.

The proof of Theorem 1 is given in Appendix A. This statement shows that the scale of $\bar{\theta}_{\mathrm{det}}^{(\gamma,H)} - \theta^\star$ depends on the scale of local gradients at $\theta^\star$, but *also on the difference of Hessians at the solution.* Furthermore, as a byproduct of Propositions 1 and 2, we obtain the following corollary, establishing the convergence of FedAvg-D to a neighborhood of $\theta^\star$.

Corollary 1 (Convergence Rate of Deterministic FedAvg-D)**.** *Assume A 1 and A 2. Let $H > 0$ and $\gamma \le 1/L$ such that $\gamma\mu H \le 1$. Then the global iterates of* FedAvg-D *satisfy*

$$\|\theta_t - \theta^\star\|^2 \le 2(1-\gamma\mu)^{Ht}\|\theta_0 - \bar{\theta}_{\mathrm{det}}^{(\gamma,H)}\|^2 + 2\gamma^2(H-1)^2\mathrm{C}_1^2 \, .$$

This result shows that the iterates of FedAvg-D converge linearly to a neighborhood of the solution $\theta^\star$. The radius of this neighborhood is determined by the level of heterogeneity among the clients, quantified by $\zeta_{\star,1}$. Although this result may seem close to existing analyses (e.g., Wang et al., 2024), we stress that our analysis technique is completely novel.

5 STOCHASTIC FEDAVG

In this section, we present our main findings, including the first-order expansion of the bias in FedAvg

Assumption	Stochastic Bias	Heterogeneity Bias
Deterministic (Thm. 1)	N/A	$\frac{\gamma(H-1)}{2N}\nabla^2 f(\theta^\star)^{-1}\sum_{c=1}^N(\nabla^2 f_c(\theta^\star)-\nabla^2 f(\theta^\star))\nabla f_c(\theta^\star)$
Quadratic (Thm. 2)	0	$\frac{\gamma(H-1)}{2N}\nabla^2 f(\theta^\star)^{-1}\sum_{c=1}^N(\nabla^2 f_c(\theta^\star)-\nabla^2 f(\theta^\star))\nabla f_c(\theta^\star)$
Homogeneous (Thm. 3)	$\frac{\gamma}{2N}\nabla^2 f(\theta^\star)^{-1}\nabla^3 f(\theta^\star)\mathbf{A}\mathcal{C}(\theta^\star)$	0
Heterogeneous (Thm. 4)	$\frac{\gamma}{2N}\nabla^2 f(\theta^\star)^{-1}\nabla^3 f(\theta^\star)\mathbf{A}\mathcal{C}(\theta^\star)$	$\frac{\gamma(H-1)}{2N}\nabla^2 f(\theta^\star)^{-1}\sum_{c=1}^N(\nabla^2 f_c(\theta^\star)-\nabla^2 f(\theta^\star))\nabla f_c(\theta^\star)$

Table 1: Summary of our main results. Each row indicates, for one of our four possible setups, which biases FedAvg suffers from, and the leading term in the expansion of the bias value for small values of γH.

when using stochastic gradients. We demonstrate that FedAvg is affected by *two types of bias*: one due to *heterogeneity* and the other one due to *stochasticity.* Our analysis is structured into three scenarios, with progressive complexity:

- First, when the functions f_c are quadratic, we show that, similar to the single-client setting, there is no stochastic bias, but only a bias due to heterogeneity;
- Second, assuming homogeneous functions, we show that the bias in FedAvg only arises from the use of stochastic gradients;
- Finally, in the general heterogeneous case, both sources of bias are observed.

A summary of our results can be found in Table 1. For our analysis, we introduce the following assumption, which provides an upper bound on the variance of the stochastic gradient. This bound is expressed as the variance at the solution $\theta^\star$, along with an additional polynomial term. For all $z \in \mathsf{Z}$ and $\theta \in \mathbb{R}^d$, we denote the centered stochastic gradient by

$$\varepsilon_c^z(\theta) \triangleq \nabla F_c^z(\theta) - \nabla f_c(\theta) \ . \tag{5}$$

A 3 (Gradient's Variance)**.** *There exist constants $M_\epsilon, k_\epsilon \geq 0$ such that for any $\theta \in \mathbb{R}^d$, and $c \in \{1, \dots, N\}$, it holds with a random variable Z_c with distribution ξ_c and $\varepsilon_c^z(\theta)$ as in (5), that*

$$\mathbb{E}\big[\|\varepsilon_c^{Z_c}(\theta)\|^4\big] \leq M_\epsilon \left\{1 + \|\theta - \theta^\star\|^{k_\epsilon}\right\} \ .$$

In particular, we have $\|\mathbb{E}[\varepsilon_c^{Z_c}(\theta^\star)^{\otimes 2}]\| \leq M_\epsilon$.

FedAvg **Generating Operators.** Now we extend the methodology described in the deterministic case to FedAvg with stochastic gradients. For a vector $Z_{1:N}^{1:H} = \{Z_{\tilde{c}}^{\tilde{h}} : \tilde{c} \in \{1, \dots, N\}, \tilde{h} \in \{1, \dots, H\}\}$, and any $c \in \{1, \dots, N\}$, we recursively define $\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})$ as an operator generating the local updates of FedAvg starting form θ. That is, we set $\widetilde{\mathsf{T}}_c^{(\gamma,0)} = \mathrm{Id}$, and for $h \geq 0$, we define

$$\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) = \Big(\mathrm{Id} - \gamma \nabla F_c^{Z_c^{h+1}}\Big)\Big(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})\Big) \ .$$

We then define $\widetilde{\mathsf{T}}^{(\gamma,H)}(\theta; Z_{1:N}^{1:H})$ as an operator generating the global updates of FedAvg. That is, for $\theta \in \mathbb{R}^d$, we let

$$\widetilde{\mathsf{T}}^{(\gamma,H)}\left(\theta; Z_{1:N}^{1:H}\right) \triangleq \frac{1}{N}\sum_{c=1}^N \widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_{1:H}^c) \ . \tag{6}$$

Note that (6) can also be written as $\widetilde{\mathsf{T}}^{(\gamma,H)}\left(\theta; Z_{1:N}^{1:H}\right) = \left(\mathrm{Id} - \gamma \mathsf{G}^{(\gamma,H)}(\cdot; Z_{1:N}^{1:H})\right)\theta$, where

$$\mathsf{G}^{(\gamma,H)}(\theta; Z_{1:N}^{1:H}) \triangleq \frac{1}{N}\sum_{c=1}^N \sum_{h=0}^{H-1} \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \ .$$

With the notations above, we have that the iterates defined in (2) can be written, for any $t \geq 0$, as

$$\theta_{t+1} = \widetilde{\mathsf{T}}^{(\gamma,H)}\left(\theta_t; Z_{1:H,t}^{1:N}\right) \ , \tag{7}$$

with $Z_{1:H,t}^{1:N}$ the random states at global iteration t. We now study the properties of the sequence $\{\theta_t\}_{t\in\mathbb{N}}$.

Properties of $\{\theta_t\}_{t\in\mathbb{N}}$ as a Markov chain. Equation (7) shows that FedAvg's global iterates define a time-homogeneous Markov chain with the corresponding Markov kernel κ on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$ defined as

$$\kappa(\theta, \mathsf{B}) = \mathbb{E}[\mathbb{1}_{\mathsf{B}}(\widetilde{\mathsf{T}}^{(\gamma,H)}(\theta, Z_{1:N}^{1:H}))] \ , \quad \mathsf{B} \in \mathcal{B}(\mathbb{R}^d), \theta \in \mathbb{R}^d \ .$$

Next we define, for $t \geq 1$, the iterates of κ as $\kappa^1 = \kappa$, and, with $\mathsf{B} \in \mathcal{B}(\mathbb{R}^d), \theta \in \mathbb{R}^d$,

$$\kappa^{t+1}(\theta, \mathsf{B}) = \int \kappa^t(\theta, \mathrm{d}\vartheta)\kappa(\vartheta, \mathsf{B}) \ .$$

For any probability measure ρ on $\mathcal{B}(\mathbb{R}^d)$ and $t \in \mathbb{N}^*$, $\rho\kappa^t$ is the distribution of the iterates θ_t of FedAvg when started from $\theta_0 \sim \rho$. The iterates of FedAvg converge to a unique stationary distribution, similar to Corollary 2 in stochastic regime.

Proposition 3. *Assume A 1 and let $\gamma \leq 1/L$. Then the iterates of* FedAvg *converge to a unique stationary distribution $\pi^{(\gamma,H)}$, admitting a finite second moment. Furthermore, for any initial distribution ρ and $t \in \mathbb{N}^*$,*

$$\mathbf{W}_2^2(\rho\kappa^t, \pi^{(\gamma,H)}) \leq (1 - \gamma\mu)^{Ht}\mathbf{W}_2^2(\rho, \pi^{(\gamma,H)}) \ .$$

The proof is postponed to Appendix B.1. Proposition 3 shows that the Markov kernel κ is geometrically ergodic in 2-Wasserstein distance. Moreover, the distribution of θ_t converges to the limiting distribution $\pi^{(\gamma,H)}$ at a linear rate $(1-\mu/L)$, for a step size of $1/L$, with the exponent given by the number of *effective* steps $H \times t$.

Under the conditions of Proposition 3, we define the mean and covariance matrix of θ_t under the invariant distribution $\pi^{(\gamma,H)}$, that is,

$$\begin{aligned} \bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} &\triangleq \int \vartheta \pi^{(\gamma,H)}(\mathrm{d}\vartheta) \ , \\ \bar{\boldsymbol{\Sigma}}^{(\gamma,H)}_{\mathrm{sto}} &\triangleq \int \{\vartheta - \theta^\star\}^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\vartheta) \ . \end{aligned} \tag{8}$$

In the remainder of this section, we derive expansions in γ and γH for the bias $\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star$ and $\bar{\boldsymbol{\Sigma}}^{(\gamma,H)}_{\mathrm{sto}}$. To this end, we define for $c \in \{1, \dots, N\}$ the matrices $\Gamma_c^{\star,H}$ and $\Gamma^{\star,H}$:

$$\Gamma_c^\star \triangleq \left(\mathrm{Id} - \gamma \nabla^2 f_c(\theta^\star)\right)^H \ , \ \Gamma^\star \triangleq \frac{1}{N} \sum_{c=1}^N \Gamma_c^\star \ . \tag{9}$$

Note that $\Gamma_c^\star$ and $\Gamma^\star$ are analogous to the matrices introduced in (4),but, contrarily to (4), we use the Hessian of f_c at $\theta^\star$. We also define the following quantities, that will appear in our analysis of bias and variance of $\pi^{(\gamma,H)}$:

$$\begin{aligned} \mathbf{A} &\triangleq (\mathrm{Id} \otimes \nabla^2 f(\theta^\star) + \nabla^2 f(\theta^\star) \otimes \mathrm{Id})^{-1} \ , \\ \mathcal{C}(\theta^\star) &\triangleq \mathbb{E}\Big[\frac{1}{N} \sum_{c=1}^N \varepsilon_1^1(\theta^\star)^{\otimes 2}\Big] \ . \end{aligned} \tag{10}$$

Quadratic Functions. When the functions f_c are quadratic, we show that FedAvg's bias only comes from heterogeneity.

A 4. *Assume that for $c \in \{1, \dots, N\}$ it holds*

$$f_c(\theta) = \tfrac{1}{2} \|(\bar{A}_c)^{1/2}(\theta - \theta_c^\star)\|^2 \ ,$$

where $\bar{A}_c \in \mathbb{R}^{d \times d}$ is a positive semi-definite matrix, and $\theta_c^\star \in \mathbb{R}^d$.

Note that $\theta^\star$ generally differ from $N^{-1} \sum_{c=1}^N \theta_c^\star$ when not all the $\theta_c^\star$'s or the $\bar{A}_c$'s are equal.

Theorem 2. *Assume A 1, A 2, A 3, and A 4. Then, using notations from (9), the bias of* FedAvg *is*

$$\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star = \frac{1}{N} \sum_{c=1}^N (\mathrm{Id} - \Gamma^\star)^{-1} (\mathrm{Id} - \Gamma_c^\star)(\theta^\star - \theta_c^\star) \ .$$

Furthermore, when $\gamma \mu H \leq 1$, it holds that

$$\|\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star\| \leq \gamma (H-1) \zeta_{\star,2} \zeta_{\star,1} / \mu \ ,$$

and the following expansion holds, using notations from (8),

$$\begin{aligned} \bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star &= \frac{\gamma(H-1)}{2} \mathrm{b_h} + O(\gamma^2 H^2) \ , \\ \bar{\boldsymbol{\Sigma}}^{(\gamma,H)}_{\mathrm{sto}} &= \frac{\gamma}{N} \mathbf{A} \mathcal{C}(\theta^\star) + O(\gamma^2 H^2 + \gamma^{3/2} H^{1/2}) \ , \end{aligned}$$

where $\mathbf{A}$ and $\mathcal{C}(\theta^\star)$ are defined in (10) and the heterogeneity bias $\mathrm{b_h}$ is given in Theorem 1.

The proof is given in Appendix B.3, together with an explicit expression of the remainders. This result demonstrates that for quadratic problems, the bias of FedAvg is *solely driven by heterogeneity.* Furthermore, its magnitude is bounded above by the product of the gradient and Hessian heterogeneities: there is no bias if either of these terms is zero. This refines previous bounds in the quadratic setting [Wang et al., 2024, Mangold et al., 2024]. Additionally, we confirm that there is no bias when $H = 1$, that is if only a single local step is performed. It also reveals that the variance of FedAvg's stationary distribution scales as $\frac{1}{N}$, ensuring a linear speed-up with the number of clients — a critical feature for federated learning.

Homogeneous Functions. When the functions f_c are homogeneous, we demonstrate that FedAvg remains biased, with the bias arising solely from the stochasticity of the gradients. Namely, we consider the following assumption.

A 5 (Homogeneity)**.** *Assume that all functions are equal, that is, $f_c = f$ for all $c \in \{1, \dots, N\}$.*

Under this assumption, the following theorem holds.

Theorem 3. *Assume A 1, A 3 and A 5. Let $\gamma \leq 1/L \wedge 1/H$, then the bias and variance of* FedAvg*, as per (8), under the stationary distribution $\pi^{(\gamma,H)}$ are*

$$\begin{aligned} \bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star &= \frac{\gamma}{2N} \mathrm{b_s} + O(\gamma^2 H + \gamma^{3/2}) \ , \\ \bar{\boldsymbol{\Sigma}}^{(\gamma,H)}_{\mathrm{sto}} &= \frac{\gamma}{N} \mathbf{A} \mathcal{C}(\theta^\star) + O(\gamma^2 H + \gamma^{3/2}) \ , \end{aligned}$$

where $\mathbf{A}$ and $\mathcal{C}(\theta^\star)$ are defined in (10), and the stochasticity bias $\mathrm{b_s}$ is given by

$$\mathrm{b_s} \triangleq \nabla^2 f(\theta^\star)^{-1} \nabla^3 f(\theta^\star) \mathbf{A} \mathcal{C}(\theta^\star) \ .$$

The proof of Theorem 3 is given, together with an explicit expression of the remainders, in Appendix B.4. Theorem 3 shows that FedAvg is bias whenever the function f is not quadratic. This bias is proportional to the third-order derivative of f and the variance of gradients at the solution. Crucially, this bias exists even when clients are homogeneous. It closely resembles the bias of SGD given in Dieuleveut et al. [2020]

for $N = 1$, and arises from the fact that the third derivative of f_c is non-zero. Remarkably, Theorem 3 guarantees that, as long as γH is small enough, both the bias and variance of FedAvg decrease inversely proportional to the number of clients N, leading to the desired linear speed-up property.

It is worth noting that FedAvg's bias in homogeneous settings has been previously identified as *iterate bias*. Khaled et al. [2020], Wang et al. [2024] showed that this iterate bias scales with a uniform bound on the gradient variance, and Glasgow et al. [2022] provided a refined upper bound using constraints on the third-order derivative of f. Our contribution goes beyond these results and offers a precise first-order expansion of this bias. Importantly, our estimate scales with the variance at $\theta^\star$, and does not require a uniform bound on gradient variance.

Heterogeneous Functions. Finally, we present the bias of FedAvg in the general case, encompassing non-quadratic and heterogeneous functions.

Theorem 4. *Assume A 1, A 2 and A 3. Let $\gamma \leq 1/L \wedge 1/H$, then the bias and variance of* FedAvg*, as defined in* (8)*, are*

$$\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star = \frac{\gamma}{2N}\mathrm{b_s} + \frac{\gamma(H-1)}{2}\mathrm{b_h} + O(\gamma^2 H^2) \ ,$$
$$\bar{\boldsymbol{\Sigma}}^{(\gamma,H)}_{\mathrm{sto}} = \frac{\gamma}{N}\mathbf{A}\mathcal{C}(\theta^\star) + O(\gamma^2 H^2 + \gamma^{3/2} H^{1/2}) \ ,$$

where $\mathbf{A}$ *and* $\mathcal{C}(\theta^\star)$ *are defined in* (10)*, and* $\mathrm{b_h}$ *and* $\mathrm{b_s}$ *are defined in Theorems 2 and 3 respectively.*

The proof of Theorem 4, as well as an explicit expression of the remainders, is given in Appendix B.5. This result shows that the bias of FedAvg with heterogeneous clients consists of two terms: one due to heterogeneity, which exactly matches the bias of FedAvg in quadratic settings, and one due to stochasticity, which exactly matches the bias of FedAvg for homogeneous functions. Again, in this result, we show that when H is of order $O(1/N)$, FedAvg achieves the linear speed-up with respect to the number of clients N.

6 RICHARDSON-ROMBERG FOR FEDERATED AVERAGING

In this section, we outline the application of Richardson-Romberg extrapolation to FedAvg in the context of stochastic gradients and heterogeneous clients. This approach builds upon the bias expression derived from Theorems 2 to 4 to define new estimators by running FedAvg twice, using different step sizes or varying the number of local updates, and then combining the resulting iterates. Specifically, for $t \in \{0, \dots, T\}$, let $\theta_t^{(\gamma,H)}$ represent the iterates of FedAvg with parameters γ and H, and $\theta_t^{(2\gamma,H)}$ represent the iterates with parameters 2γ and H. Using these, we can define

$$\vartheta_t^{(\gamma,H)} = 2\theta_t^{(\gamma,H)} - \theta_t^{(2\gamma,H)} \ , \quad \bar{\vartheta}_T^{(\gamma,H)} = \frac{1}{T}\sum_{t=0}^{T-1} \vartheta_t^{(\gamma,H)} \ .$$

Then, we obtain the following result from Proposition 1 and Theorem 4.

Theorem 5. *Assume A 1, A 2 and A 3. Let $\gamma \leq 2^{-1}[1/L \wedge 1/H]$, then the iterates $\{\bar{\vartheta}_T^{(\gamma,H)}\}_{T\geq 1}$ converge in L^2 to $\bar{\vartheta}^{(\gamma,H)}_{\mathrm{sto}} = 2\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \bar{\theta}^{(2\gamma,H)}_{\mathrm{sto}}$:*

$$\lim_{T\to\infty} \mathbb{E}\left[\|\bar{\vartheta}_T^{(\gamma,H)} - \bar{\vartheta}^{(\gamma,H)}_{\mathrm{sto}}\|^2\right] = 0 \ .$$

In addition, $\bar{\vartheta}^{(\gamma,H)}_{\mathrm{sto}}$ satisfies

$$\bar{\vartheta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star = O(\gamma^2 H^2 + \gamma^{3/2} H^{1/2}) \ .$$

Consequently, this implies that, when γH is small, the averaged iterates of FedAvg with Richardson-Romberg extrapolation have a smaller bias than vanilla FedAvg. Remarkably, this procedure is able to reduce the bias of FedAvg without requiring the clients to store an additional variable. This method is thus very well suited for use with devices that have limited computational resources.

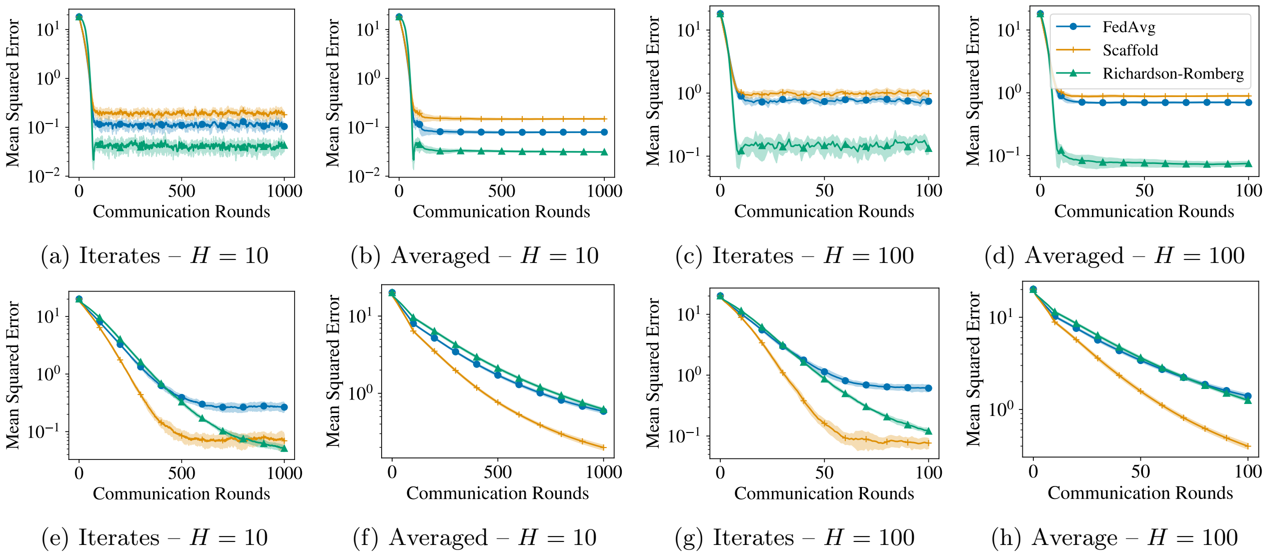
Note that, in contrast to Dieuleveut et al. [2020], we do not deal with the variance of FedAvg and therefore its Richardson-Romberg approximation, i.e., we do not quantify the rate of convergence to 0 of $\mathbb{E}[\|T^{-1}\sum_{t=0}^{T-1}\vartheta_t^{(\gamma,H)} - \bar{\theta}^{(\gamma,H)}_{\mathrm{sto}}\|^2]$. Solving this question is an interesting direction for future work.

Remark 1. *When $H > 1$, one could define a Richardson-Romberg estimator by varying the number of local steps, resulting in the construction $\omega_t^{(\gamma,H)} = (2H-1)/(H-1)\theta_t^{(\gamma,H)} - \theta_t^{(2\gamma,H)}$ and $\bar{\omega}_T^{(\gamma,H)} = T^{-1}\sum_{t=0}^{T-1}\omega_t^{(\gamma,H)}$. The sequence $\{\bar{\omega}_T^{(\gamma,H)}\}_{T\geq 1}$ converges to $(2H-1)/(H-1)\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \bar{\theta}^{(\gamma,2H)}_{\mathrm{sto}} = \gamma\mathrm{b_s}/(2N) + O(\gamma^2 H^2 + \gamma^{3/2} H^{1/2})$, removing heterogeneity bias but not stochasticity bias. The iterates obtained through this procedure therefore have a bias close to the one of the homogeneous setting.*

7 NUMERICAL EXPERIMENTS

This section illustrates our theoretical findings using regularized logistic regression problems. This problem can be formulated as (1), using $z = (x, y)$ where x and y are respectively the data features and label, and $\lambda > 0$ is a regularization parameter

$$f_c(\theta) = \mathbb{E}\left[\log(1 + \exp(1 - y_c x_c^\top \theta)) + \lambda/2\|\theta\|^2\right] \ ,$$



(a) Iterates – $H = 10$ (b) Averaged – $H = 10$ (c) Iterates – $H = 100$ (d) Averaged – $H = 100$

(e) Iterates – $H = 10$ (f) Averaged – $H = 10$ (g) Iterates – $H = 100$ (h) Average – $H = 100$

Figure 1: Mean squared error on the `synthetic noisy` (first line) and on the `synthetic heterogeneous` dataset (second line), as a function of the number of communications, for $H \in \{10, 100\}$. In Figures 1a, 1c, 1e and 1g (labelled *Iterates*), we plot the MSE for global iterates of the three methods, while in Figures 1b, 1d, 1f and 1h (labelled *Averaged*), we plot the MSE for first 10% of iterates, and then plot the MSE of the averaged iterates for the last 90% of the iterates. We plot the average over 10 runs, with standard deviation.

where, for each $c \in \{1, \dots, N\}$, the sample $z_c = (x_c, y_c)$ is drawn from client c's local distribution.

We evaluate our approach on two synthetic datasets with $N = 10$ clients. The first dataset, coined **`synthetic noisy`**, is made of two blobs with large variance, split uniformly among clients. It is thus homogeneous, but contains very noisy data. On the opposite, the second dataset, coined **`synthetic heterogeneous`**, is made of 2 blobs with small variance. Half of the clients receive part of the observations directly, while the other half receive perturbed records with shuffled labels. In this second dataset, data is very heterogeneous but has little noise.

We evaluate three algorithms on these datasets: (i) vanilla FedAvg, (ii) FedAvg with Richardson-Romberg extrapolation, as described in Section 6, and (iii) Scaffold [Karimireddy et al., 2020]. For all experiments, we use $N = 10$ and run the algorithm for a total of $TH = 10,000$ estimation of the full gradient, using batch size one and step size $\gamma = 0.01$.

We plot the results of these experiments in Figure 1. In all results, FedAvg with Richardson-Romberg performs at least as good as vanilla FedAvg, which is in line with our theory. However, in non-noisy, stochastic settings (second line of Figure 1), it can only partly remove the bias due to heterogeneity; on the opposite, Scaffold handles this bias successfully. More remarkably, when clients are homogeneous, but have noisy data (first line of Figure 1), FedAvg with Richardson-Romberg can reduce the bias, while Scaffold fails. This further confirms our theory, and highlights the fact that FedAvg with Richardson-Romberg extrapolation effectively reduces FedAvg's bias due to stochasticity.

8 CONCLUSION

In this paper, we introduced a novel perspective on FedAvg, centered on the idea that the global iterates of the algorithm converge to a stationary distribution. We conducted a detailed analysis of this distribution, deriving an exact first-order expression for both the bias and variance of FedAvg's iterates. Notably, our results demonstrate that, as long as the number of local steps is not excessively large, the bias of FedAvg decreases at a rate of $1/N$. Moreover, we established that FedAvg's bias consists of two distinct components: one arising purely from data heterogeneity and the other from the stochastic nature of the gradients. Crucially, this proves that FedAvg remains biased even in perfectly homogeneous settings. Building on this key insight, we applied the Richardson-Romberg extrapolation technique to introduce a new method for mitigating FedAvg's bias. Unlike existing approaches, our method can reduce *both sources of bias*—heterogeneity bias and gradient stochasticity bias—offering a more comprehensive solution. This opens novel perspectives for the design of federated learning methods with local training.

REFERENCES

Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. *Advances in neural information processing systems*, 28, 2015.

Francis Bach. On the effectiveness of richardson extrapolation in data science. *SIAM Journal on Mathematics of Data Science*, 3(4):1251–1277, 2021.

Jerry Chee and Panos Toulis. Convergence diagnostics for stochastic gradient descent with constant step size. *arXiv preprint arXiv:1710.06382*, 2017.

Laurent Condat and Peter Richtárik. Randprox: Primal-dual optimization algorithms with randomized proximal updates. *arXiv preprint arXiv:2207.12891*, 2022.

Laurent Condat, Ivan Agarskỳ, and Peter Richtárik. Provably doubly accelerated federated learning: The first theoretically successful combination of local training and communication compression. *arXiv preprint arXiv:2210.13277*, 2022.

Michael Crawshaw, Yajie Bao, and Mingrui Liu. Federated learning with client subsampling, data heterogeneity, and unbounded smoothness: A new algorithm and lower bounds. *Advances in Neural Information Processing Systems*, 36, 2024.

Alexandre Défossez and Francis Bach. Averaged least-mean-squares: Bias-variance trade-offs and optimal sampling distributions. In *Artificial Intelligence and Statistics*, pages 205–213. PMLR, 2015.

Aymeric Dieuleveut and Francis Bach. Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, 44(4):1363 – 1399, 2016. doi: 10.1214/15-AOS1391. URL https://doi.org/10.1214/15-AOS1391.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348 – 1382, 2020. doi: 10.1214/19-AOS1850. URL https://doi.org/10.1214/19-AOS1850.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Probability and moment inequalities for additive functionals of geometrically ergodic markov chains. *Journal of Theoretical Probability*, pages 1–50, 2024.

Margalit R Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR, 2022.

Michał Grudzień, Grigory Malinovsky, and Peter Richtárik. Can 5th generation local training methods support client sampling? yes! In *International Conference on Artificial Intelligence and Statistics*, pages 1055–1092. PMLR, 2023.

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143. PMLR, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

Xiang Li, Wenhao Yang, Shusen Wang, and Zhihua Zhang. Communication-efficient local decentralized sgd methods. *arXiv preprint arXiv:1910.09126*, 2019.

Grigory Malinovsky, Kai Yi, and Peter Richtárik. Variance reduced proxskip: Algorithm, theory and application to federated learning. *Advances in Neural Information Processing Systems*, 35:15176–15189, 2022.

Paul Mangold, Sergey Samsonov, Safwan Labbi, Ilya Levin, Reda Alami, Alexey Naumov, and Eric Moulines. SCAFFLSA: Quantifying and Eliminating Heterogeneity Bias in Federated Linear Stochastic Approximation and Temporal Difference Learning. *arXiv preprint arXiv:2402.04114*, 2024.

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

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pages 15750–15769. PMLR, 2022.

Eric Moulines, Pierre Priouret, and François Roueff. On recursive estimation for time varying autoregressive processes. *The Annals of Statistics*, 33(6):2610 – 2654, 2005. doi: 10.1214/009053605000000624. URL https://doi.org/10.1214/009053605000000624.

Kumar Kshitij Patel, Margalit Glasgow, Lingxiao Wang, Nirmit Joshi, and Nathan Srebro. On the still unreasonable effectiveness of federated averaging for heterogeneous distributed learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

Georg Ch Pflug. Stochastic minimization with constant step-size: asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986.

Zhaonan Qu, Kaixiang Lin, Zhaojian Li, and Jiayu Zhou. Federated learning's blessing: Fedavg has linear speedup. In *ICLR 2021-Workshop on Distributed and Private Machine Learning (DPML)*, 2021.

Zhaonan Qu, Kaixiang Lin, Zhaojian Li, Jiayu Zhou, and Zhengyuan Zhou. A unified linear speedup analysis of federated averaging and nesterov fedavg. *Journal of Artificial Intelligence Research*, 78:1143–1200, 2023.

Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021.

Lewis Fry Richardson. Ix. the approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 210(459-470):307–357, 1911.

Abdurakhmon Sadiev, Dmitry Kovalev, and Peter Richtárik. Communication acceleration of local gradient methods via an accelerated primal-dual algorithm with an inexact prox. *Advances in Neural Information Processing Systems*, 35:21777–21791, 2022.

Marina Sheshukova, Denis Belomestny, Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Nonasymptotic analysis of stochastic gradient descent with the richardson-romberg extrapolation. *arXiv preprint arXiv:2410.05106*, 2024.

Sebastian U Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations*, 2019.

Josef Stoer and Roland Bulirsch. *Introduction to numerical analysis*. Springer Science & Business Media, 2013.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *Trans. Mach. Learn. Res.*, 2024, 2024.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019a.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5693–5700, 2019b.

Ali Zindari, Ruichen Luo, and Sebastian U Stich. On the convergence of local sgd under third-order smoothness and hessian similarity. In *OPT 2023: Optimization for Machine Learning*, 2023.

Checklist

1. For all models and algorithms presented, check if you include:
   - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes, in Section 2, Section 5.**
   - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes, in Section 4, Section 5 and Section 6.**
   - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes, and we provide code in supplementary.**
2. For any theoretical claim, check if you include:
   - (a) Statements of the full set of assumptions of all theoretical results. **Yes, we state all assumptions used in every theorem, and state the assumption in Section 2 and beginning of Section 5 so that it is easy to find them.**
   - (b) Complete proofs of all theoretical results. **Yes, all proofs are provided in appendix.**
   - (c) Clear explanations of any assumptions. **Yes, we describe every assumption and the rationale behind it.**
3. For all figures and tables that present empirical results, check if you include:
   - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes, we provide code as supplementary and will release it upon publication of the paper.**
   - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes, in Section 7.**
   - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes, we say that error bars represent standard deviation over multiple runs of our algorithms.**
   - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes, experiments we run on a laptop.**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   - (a) Citations of the creator If your work uses existing assets. **Yes, we put appropriate reference for all datasets used.**
   - (b) The license information of the assets, if applicable. **Yes.**
   - (c) New assets either in the supplemental material or as a URL, if applicable. **Not applicable.**
   - (d) Information about consent from data providers/curators. **Not applicable.**
   - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not applicable.**
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   - (a) The full text of instructions given to participants and screenshots. **Not applicable.**
   - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not applicable.**
   - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not applicable.**

Supplementary Materials

A Refined Analysis of FedAvg

A.1 Convergence and Bias

To study the convergence of FedAvg-D, we first recall the notations introduced in Section 4. Namely, we recall that the local updates of FedAvg-D for $\theta \in \mathbb{R}^d$ and $0 \le h \le H-1$ are denoted as

$$\begin{aligned} \mathsf{T}_c^{(\gamma,0)}(\theta) &\triangleq \theta \ , \\ \mathsf{T}_c^{(\gamma,h+1)}(\theta) &\triangleq \mathsf{T}_c^{(\gamma,h)}(\theta) - \gamma \nabla f_c(\mathsf{T}_c^{(\gamma,h)}(\theta)) \ . \end{aligned}$$

Additionally, we recall that $\mathsf{T}^{(\gamma,H)} = \frac{1}{N}\sum_{c=1}^{N} \mathsf{T}_c^{(\gamma,H)}$. First, we show that the local operators are contractions.

Lemma 1 (Contraction of FedAvg-D's Local Iterates)**.** *Assume A 1. Then, for any $\gamma \le 1/(2L)$, $\theta, \vartheta \in \mathbb{R}^d$, and $c \in \{1, \ldots, N\}$, it holds that*

$$\|(\theta - \gamma \nabla f_c(\theta)) - (\vartheta - \gamma \nabla f_c(\vartheta))\|^2 \le (1-\gamma\mu)\|\theta - \vartheta\|^2 \ . \tag{11}$$

Proof. Using strong convexity and co-coercivity, we have for any $c \in \{1, \ldots, N\}$, that

$$\begin{aligned} \|(\theta - \gamma \nabla f_c(\theta)) - (\vartheta - \gamma \nabla f_c(\vartheta))\|^2 &= \|\theta - \vartheta\|^2 + \gamma^2 \|\nabla f_c(\theta) - \nabla f_c(\vartheta)\|^2 - 2\gamma\langle \theta - \vartheta, \nabla f_c(\theta) - \nabla f_c(\vartheta \quad)\rangle \quad (12) \\ &\le \|\theta - \vartheta\|^2 - 2\gamma(1-\gamma L)\langle \theta - \vartheta, \nabla f_c(\theta) - \nabla f_c(\vartheta)\rangle \quad (13) \\ &\le \|\theta - \vartheta\|^2 - 2\gamma\mu(1-\gamma L)\|\theta - \vartheta\|^2 \ . \quad (14) \end{aligned}$$

To conclude, it remains to note that $\gamma \le 1/(2L)$. □

Lemma 2 (Contraction of FedAvg-D's Global Iterates)**.** *Assume A 1. Then for any $H > 0$, $\gamma \le 1/(2L)$, and $\theta, \vartheta \in \mathbb{R}^d$, the operator $\mathsf{T}^{(\gamma,H)}$ satisfies*

$$\|\mathsf{T}^{(\gamma,H)}(\theta) - \mathsf{T}^{(\gamma,H)}(\vartheta)\|^2 \le (1-\gamma\mu)^H \|\theta - \vartheta\|^2 \ . \tag{15}$$

Proof. First, we show that $\mathsf{T}_c^{(\gamma,h)}$ is a strict contraction for any $h \in \{1, \ldots, H\}$. Note that for any $\theta, \vartheta \in \mathbb{R}^d$,

$$\mathsf{T}_c^{(\gamma,h+1)}(\theta) - \mathsf{T}_c^{(\gamma,h+1)}(\vartheta) = (\mathsf{T}_c^{(\gamma,h)}(\theta) - \gamma \nabla f_c(\mathsf{T}_c^{(\gamma,h)}(\theta))) - (\mathsf{T}_c^{(\gamma,h)}(\vartheta) - \gamma \nabla f_c(\mathsf{T}_c^{(\gamma,h)}(\vartheta))) \ . \tag{16}$$

Thus, it follows from Lemma 1 that

$$\|\mathsf{T}_c^{(\gamma,h+1)}(\theta) - \mathsf{T}_c^{(\gamma,h+1)}(\vartheta)\|^2 \le (1-\gamma\mu)\|\mathsf{T}_c^{(\gamma,h)}(\theta) - \mathsf{T}_c^{(\gamma,h)}(\vartheta)\|^2 \ . \tag{17}$$

Using Jensen's inequality and applying (17) recursively, we obtain

$$\|\mathsf{T}^{(\gamma,H)}(\theta) - \mathsf{T}^{(\gamma,H)}(\vartheta)\|^2 \le \frac{1}{N}\sum_{c=1}^{N} \|\mathsf{T}_c^{(\gamma,H)}(\theta) - \mathsf{T}_c^{(\gamma,H)}(\vartheta)\|^2 \le (1-\gamma\mu)^H \|\theta - \vartheta\|^2 \ , \tag{18}$$

which concludes the proof. □

Corollary 2 (Stationary Point of FedAvg-D)**.** *Assume A 1. Then for any $H > 0$ and $\gamma \le 1/(2L)$, deterministic FedAvg converges to a unique point $\bar{\theta}_{\text{det}}^{(\gamma,H)}$ which satisfies $\mathsf{T}^{(\gamma,H)}(\bar{\theta}_{\text{det}}^{(\gamma,H)}) = \bar{\theta}_{\text{det}}^{(\gamma,H)}$ and $\mathsf{g}^{(\gamma,H)}(\bar{\theta}_{\text{det}}^{(\gamma,H)}) = 0$.*

Proof. By Lemma 2, $\mathsf{T}^{(\gamma,H)}$ is a contraction mapping. Thus the result follows from Banach fixed point theorem. □

Theorem 6 (Convergence Rate of FedAvg-D)**.** *Assume A 1. Let $H > 0$, $\gamma \leq 1/L$, and define $\bar{\theta}^{(\gamma,H)}_{\text{det}}$ as the unique point such that $\mathsf{T}^{(\gamma,H)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) = \bar{\theta}^{(\gamma,H)}_{\text{det}}$. Then, the iterates of* FedAvg *satisfy*

$$\|\theta_t - \bar{\theta}^{(\gamma,H)}_{\text{det}}\|^2 \leq (1-\gamma\mu)^{Ht}\|\theta_0 - \bar{\theta}^{(\gamma,H)}_{\text{det}}\|^2 \ . \tag{19}$$

Proof. Let $t > 0$, and θ_{t+1} be the $(t+1)$-th global iterate of FedAvg. Since $\mathsf{T}^{(\gamma,H)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) = \bar{\theta}^{(\gamma,H)}_{\text{det}}$, we write

$$\theta_{t+1} - \bar{\theta}^{(\gamma,H)}_{\text{det}} = \mathsf{T}^{(\gamma,H)}(\theta_t) - \mathsf{T}^{(\gamma,H)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) \ . \tag{20}$$

Thus, by Lemma 2, we have

$$\|\theta_{t+1} - \bar{\theta}^{(\gamma,H)}_{\text{det}}\|^2 = \|\mathsf{T}^{(\gamma,H)}(\theta_t) - \mathsf{T}^{(\gamma,H)}(\bar{\theta}^{(\gamma,H)}_{\text{det}})\|^2 \leq (1-\gamma\mu)^H\|\theta_t - \bar{\theta}^{(\gamma,H)}_{\text{det}}\|^2 \ , \tag{21}$$

and the result follows by induction. □

Theorem 7 (Bias of FedAvg)**.** *Assume A 1 and A 2. Let $H > 0$, $\gamma \leq 1/(2L)$ such that $\gamma\mu H \leq 1$, then we have*

$$\|\bar{\theta}^{(\gamma,H)}_{\text{det}} - \theta^\star\|^2 \leq \frac{\gamma^2 L^2 (H-1)^2}{\mu^2}\zeta_{\star,1} \ . \tag{22}$$

Consequently, it holds that $\|\bar{\theta}^{(\gamma,H)}_{\text{det}} - \theta^\star\| = O(\gamma H)$ and $\|\mathsf{T}^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) - \theta^\star\| = O(\gamma H)$.

Proof. Starting from $\bar{\theta}^{(\gamma,H)}_{\text{det}}$, we write

$$\begin{aligned}
\mathsf{T}_c^{(\gamma,h+1)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) &= \mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) - \gamma\nabla f_c(\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\text{det}})) \qquad (23)\\
&= \mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) - \gamma(\nabla f_c(\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\text{det}})) - \nabla f_c(\theta^\star)) - \gamma\nabla f_c(\theta^\star) \ . \qquad (24)
\end{aligned}$$

Using the hessian matrix of f_c, we write the previous identity as

$$\mathsf{T}_c^{(\gamma,h+1)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) = \mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) - \gamma\bar{D}_c^{(\theta^\star_{c,h},\theta^\star)}(\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) - \theta^\star) - \gamma\nabla f_c(\theta^\star) \ , \tag{25}$$

where $\bar{D}_c^{(\theta^\star_{c,h},\theta^\star)} = \int_0^1 \nabla^2 f_c(t\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) + (1-t)\theta^\star)\mathrm{d}t$, and, for $\ell \in \{0,\ldots,H\}$, $\theta^\star_{c,\ell}$ is obtained after ℓ local iterates of agent c, starting from $\bar{\theta}^{(\gamma,H)}_{\text{det}}$, that is,

$$\theta^\star_{c,\ell} = \mathsf{T}_c^{(\gamma,\ell)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) \ .$$

Consequently, we have

$$\mathsf{T}_c^{(\gamma,H)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) - \theta^\star = \mathsf{F}_c^{\star,1:H}(\bar{\theta}^{(\gamma,H)}_{\text{det}} - \theta^\star) - \gamma\sum_{h=1}^{H}\mathsf{F}_c^{\star,h+1:H}\nabla f_c(\theta^\star) \ , \tag{26}$$

where we set, for $h \in \{1,\ldots,H\}$, the quantity

$$\mathsf{F}_c^{\star,h:H} = \prod_{\ell=h}^{H-1}\left(\mathrm{Id} - \gamma\bar{D}_c^{(\theta^\star_{c,\ell},\theta^\star)}\right) \ . \tag{27}$$

Averaging over all clients, we obtain

$$\mathsf{T}^{(\gamma,H)}(\bar{\theta}^{(\gamma,H)}_{\text{det}}) - \theta^\star = \mathsf{F}^\star(\bar{\theta}^{(\gamma,H)}_{\text{det}} - \theta^\star) - \frac{\gamma}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}\mathsf{F}_c^{\star,h+1:H}\nabla f_c(\theta^\star) \ . \tag{28}$$

We now use the fact that $\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}$ is the fixed point of $\mathsf{T}^{(\gamma,H)}$, i.e., $\mathsf{T}^{(\gamma,H)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) = \bar{\theta}^{(\gamma,H)}_{\mathrm{det}}$, and subtract $F^{\star}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}} - \theta^{\star})$ on both sides to obtain

$$(\mathrm{Id} - F^{\star})(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}} - \theta^{\star}) = -\frac{\gamma}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} F_c^{\star,h+1:H} \nabla f_c(\theta^{\star}) \ . \tag{29}$$

Now we introduce an additional notation for

$$F_{\mathrm{avg}}^{\star,h:H} = \prod_{\ell=h}^{H-1} \left(\mathrm{Id} - \frac{\gamma}{N} \sum_{c=1}^{N} \bar{D}_c^{(\theta^{\star}_{c,\ell},\theta^{\star})} \right) \ . \tag{30}$$

With $F_{\mathrm{avg}}^{\star,h:H}$ defined in (30), we get the following identity:

$$\begin{aligned}
\bar{\theta}^{(\gamma,H)}_{\mathrm{det}} - \theta^{\star} &= -\frac{\gamma}{N} (\mathrm{Id} - F^{\star})^{-1} \sum_{c=1}^{N} \sum_{h=1}^{H} F_c^{\star,h+1:H} \nabla f_c(\theta^{\star}) && (31) \\
&\overset{(a)}{=} \frac{\gamma}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} (\mathrm{Id} - F^{\star})^{-1} (F_{\mathrm{avg}}^{\star,h+1:H} - F_c^{\star,h+1:H}) \nabla f_c(\theta^{\star}) && (32) \\
&\overset{(b)}{=} \frac{\gamma}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} \sum_{k=0}^{\infty} (F^{\star})^k (F_{\mathrm{avg}}^{\star,h+1:H} - F_c^{\star,h+1:H}) \nabla f_c(\theta^{\star}) \ , && (33)
\end{aligned}$$

where (a) comes from $\sum_{c=1}^{N} \nabla f_c(\theta^{\star}) = 0$, and (b) is the Neumann series. Note that

$$\begin{aligned}
\left\| F_{\mathrm{avg}}^{\star,h+1:H} - F_c^{\star,h+1:H} \right\| &= \left\| \sum_{\ell=h+1}^{H} F_{\mathrm{avg}}^{\star,h+1:\ell-1} (\gamma \bar{D}_c^{(\theta^{\star}_{c,\ell},\theta^{\star})} - \tfrac{\gamma}{N} \textstyle\sum_{c'=1}^{N} \bar{D}_{c'}^{(\theta^{\star}_{c',\ell},\theta^{\star})}) F_{\mathrm{avg}}^{\star,\ell+1:H} \right\| && (34) \\
&\leq \gamma \sum_{\ell=h+1}^{H} \left\| \bar{D}_c^{(\theta^{\star}_{c,\ell},\theta^{\star})} - \tfrac{1}{N} \textstyle\sum_{c'=1}^{N} \bar{D}_{c'}^{(\theta^{\star}_{c',\ell},\theta^{\star})} \right\| \ . && (35)
\end{aligned}$$

Thus, we have $\left\| F_{\mathrm{avg}}^{\star,h+1:H} - F_c^{\star,h+1:H} \right\| \leq 2\gamma(H-h)L$. This gives

$$\begin{aligned}
\|\bar{\theta}^{(\gamma,H)}_{\mathrm{det}} - \theta^{\star}\| &\leq \frac{\gamma}{N} \sum_{k=0}^{\infty} \sum_{c=1}^{N} \sum_{h=1}^{H} \|(F^{\star})^k\| \left\| F_{\mathrm{avg}}^{\star,h+1:H} - F_c^{\star,h+1:H} \right\| \|\nabla f_c(\theta^{\star})\| && (36) \\
&\leq \frac{2\gamma^2 L}{N} \sum_{k=0}^{\infty} (1-\gamma\mu)^{Hk} \sum_{c=1}^{N} \sum_{h=1}^{H} (H-h) \|\nabla f_c(\theta^{\star})\| \ , && (37)
\end{aligned}$$

where we also used that $\|F^{\star}\| \leq (1-\gamma\mu)^H$. Consequently, when $\gamma\mu H \leq 1$, we obtain

$$\|\bar{\theta}^{(\gamma,H)}_{\mathrm{det}} - \theta^{\star}\| \leq \frac{\gamma^2 L H (H-1)}{1-(1-\gamma\mu)^H} \frac{1}{N} \sum_{c=1}^{N} \|\nabla f_c(\theta^{\star})\| \leq \frac{\gamma L (H-1)}{\mu} \frac{1}{N} \sum_{c=1}^{N} \|\nabla f_c(\theta^{\star})\| \leq \frac{\gamma L (H-1)}{\mu} \zeta_{\star,1} \ , \tag{38}$$

which is the first part of the result. From (38), it holds that $\|\bar{\theta}^{(\gamma,H)}_{\mathrm{det}} - \theta^{\star}\| = O(\gamma H)$. We now prove that the same result holds for the local iterates $\mathsf{T}^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}})$. Let $h \in \{0, \ldots, H-1\}$. Then, using the triangle inequality and the fact that $\nabla f(\theta^{\star}) = 0$, we obtain

$$\begin{aligned}
&\|\mathsf{T}_c^{(\gamma,h+1)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) - \theta^{\star}\| && (39) \\
&\quad = \|\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) - \gamma \nabla f_c(\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}})) - (\theta^{\star} - \gamma \nabla f_c(\theta^{\star})) + \gamma(\nabla f_c(\theta^{\star}) - \nabla f(\theta^{\star}))\| && (40) \\
&\quad \leq \|\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) - \gamma \nabla f_c(\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}})) - (\theta^{\star} - \gamma \nabla f_c(\theta^{\star}))\| + \gamma \|\nabla f_c(\theta^{\star}) - \nabla f(\theta^{\star})\| \ . && (41)
\end{aligned}$$

Applying Lemma 1 and (41) recursively, then A 2, we obtain

$$\|\mathsf{T}_c^{(\gamma,h+1)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) - \theta^{\star}\| \leq \|\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) - \theta^{\star}\| + \gamma \|\nabla f_c(\theta^{\star}) - \nabla f(\theta^{\star})\| \leq \|\bar{\theta}^{(\gamma,H)}_{\mathrm{det}} - \theta^{\star}\| + \gamma H \zeta_{\star,1} = O(\gamma H) \ ,$$

which proves the second part of the result. □

Corollary 3 (Convergence Rate of FedAvg)**.** *Let $H > 0$ and $\gamma \leq 1/L$. Then the global iterates of* FedAvg *satisfy*

$$\|\theta_t - \theta^\star\|^2 \leq 2(1-\gamma\mu)^{Ht}\|\theta_0 - \bar{\theta}^{(\gamma,H)}_{\mathrm{det}}\|^2 + \frac{2\gamma^2 L^2 (H-1)^2 \zeta^2_{\star,1}}{\mu^2} \ . \tag{42}$$

Proof. We start with the upper bound

$$\|\theta_t - \theta^\star\|^2 \leq 2\|\theta_t - \bar{\theta}^{(\gamma,H)}_{\mathrm{det}}\|^2 + 2\|\bar{\theta}^{(\gamma,H)}_{\mathrm{det}} - \theta^\star\|^2 \ . \tag{43}$$

Then, we apply Theorem 6 to bound the first term, and Theorem 7 to bound the second term. □

A.2 Expansion of the Bias

Proposition 4 (Expansion of FedAvg-D's Bias)**.** *Assume A1, A2. Let $H > 0$, $\gamma \leq 1/(2L)$ such that $\gamma\mu H \leq 1$, then the bias of* FedAvg-D *can be expanded as*

$$\bar{\theta}^{(\gamma,H)}_{\mathrm{det}} - \theta^\star = \frac{\gamma(H-1)}{2N}\nabla^2 f(\theta^\star)^{-1}\sum_{c=1}^{N}(\nabla^2 f_c(\theta^\star) - \nabla^2 f(\theta^\star))\nabla f_c(\theta^\star) + O(\gamma^2 H^2) \ . \tag{44}$$

Proof. Starting from (32), we have

$$\bar{\theta}^{(\gamma,H)}_{\mathrm{det}} - \theta^\star = \frac{\gamma}{N}\sum_{c=1}^{N}\sum_{h=1}^{H}(\mathrm{Id} - \varGamma^\star)^{-1}(\varGamma^{\star,h+1:H}_{\mathrm{avg}} - \varGamma^{\star,h+1:H}_{c})\nabla f_c(\theta^\star) \ . \tag{45}$$

We start by writing the expansion of $\bar{D}_c^{(\theta^\star_{c,h},\theta^\star)}$. Note that, for $t \in (0,1)$, we can write

$$t\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) + (1-t)\theta^\star = \theta^\star + t(\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) - \theta^\star) \ .$$

Thus, we can expand the Hessian

$$\nabla^2 f_c(t\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) + (1-t)\theta^\star) = \nabla^2 f_c(\theta^\star) + \mathsf{r}^c_{1,h,t}(\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}})) \ ,$$

where $\mathsf{r}^c_{1,h,t} : \mathbb{R}^d \to \mathbb{R}^d$ is such that $\sup_{\vartheta\in\mathbb{R}^d}\|\mathsf{r}^c_{1,h,t}(\vartheta)\|/\|\vartheta - \theta^\star\| < +\infty$. Hence, combining this bound and the definition of $\bar{D}_c^{(\theta^\star_{c,h},\theta^\star)}$, we obtain

$$\bar{D}_c^{(\theta^\star_{c,h},\theta^\star)} = \int_0^1 \left\{\nabla^2 f_c(\theta^\star) + \mathsf{r}^c_{1,h,t}(\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}))\right\}\mathrm{d}t = \nabla^2 f_c(\theta^\star) + \mathsf{r}^c_{1,h}(\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}})) \ , \tag{46}$$

where $\mathsf{r}^c_{1,h} : \vartheta \mapsto \int_0^1 \left\{\mathsf{r}^c_{1,h,t}(\vartheta - \theta^\star)\right\}\mathrm{d}t$ is such that

$$\sup_{\vartheta\in\mathbb{R}^d}\|\mathsf{r}^c_{1,h}(\vartheta)\|/\|\vartheta - \theta^\star\| < +\infty \ . \tag{47}$$

Using (47) and Theorem 7, we can expand $\varGamma^{\star,h+1:H}_{c} = \prod_{\ell=h}^{H-1}\left(\mathrm{Id} - \gamma\bar{D}_c^{(\theta^\star_{c,\ell},\theta^\star)}\right)$ and $(\mathrm{Id} - \Gamma^\star)^{-1}$ as

$$\varGamma^{\star,h+1:H}_{c} = \mathrm{Id} - \gamma(H-h-1)\nabla^2 f_c(\theta^\star) + \gamma H\mathcal{R}^c_{1,h}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) \ , \tag{48}$$

$$\varGamma^{\star,h+1:H}_{\mathrm{avg}} = \mathrm{Id} - \gamma(H-h-1)\nabla^2 f(\theta^\star) + \gamma H\mathcal{R}_{1,h}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}}) \ , \tag{49}$$

$$(\mathrm{Id} - \Gamma^\star)^{-1} = (\gamma H\nabla^2 f(\theta^\star))^{-1} + \mathcal{R}_1(\mathsf{T}_c^{(\gamma,h)}(\bar{\theta}^{(\gamma,H)}_{\mathrm{det}})) \ , \tag{50}$$

where $\mathcal{R}^c_{1,h} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$, $\mathcal{R}_{1,h} = \frac{1}{N}\sum_{c=1}^{N}\mathcal{R}^c_{1,h}$, and $\mathcal{R}_1 : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ are such that

$$\sup_{\vartheta\in\mathbb{R}^d}\|\mathcal{R}^c_{1,h}(\vartheta)\|/\|\vartheta - \theta^\star\| < +\infty \ , \quad \text{and} \ \sup_{\vartheta\in\mathbb{R}^d}\|\mathcal{R}_1(\vartheta)\|/\|\vartheta - \theta^\star\| < +\infty \ . \tag{51}$$

Plugging the two identities above in (45), we obtain

$$
\begin{aligned}
\bar{\theta}_{\mathrm{det}}^{(\gamma,H)} - \theta^\star &= \frac{\gamma}{N} \sum_{c=1}^{N} \sum_{h=1}^{H} \left\{ (\gamma H \nabla^2 f(\theta^\star))^{-1} + \mathcal{R}_1(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) \right\} \qquad (52) \\
&\qquad \times \left\{ \gamma(H-h-1)(\nabla^2 f_c(\theta^\star) - \nabla^2 f_{\mathrm{avg}}(\theta^\star)) + \gamma H (\mathcal{R}_{1,h}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) - \mathcal{R}_{1,h}^{c}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}))) \right\} \nabla f_c(\theta^\star) \\
&= \frac{\gamma}{NH} \sum_{c=1}^{N} \sum_{h=1}^{H} (H-h-1) \nabla^2 f(\theta^\star)^{-1} (\nabla^2 f_c(\theta^\star) - \nabla^2 f(\theta^\star)) \nabla f_c(\theta^\star) + \gamma H \mathcal{R}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) \ , \qquad (53)
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{R}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) &= \frac{1}{NH} \sum_{c=1}^{N} \sum_{h=1}^{H} \nabla^2 f(\theta^\star)^{-1} (\mathcal{R}_{1,h}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) - \mathcal{R}_{1,h}^{c}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)})) \nabla f_c(\theta^\star) \qquad (54) \\
&\quad + \frac{1}{NH} \sum_{c=1}^{N} \sum_{h=1}^{H} \gamma (H-h-1) \mathcal{R}_1(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) (\nabla^2 f_c(\theta^\star) - \nabla^2 f(\theta^\star)) \nabla f_c(\theta^\star) \\
&\quad + \frac{1}{NH} \sum_{c=1}^{N} \sum_{h=1}^{H} \gamma H \mathcal{R}_1(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) (\mathcal{R}_{1,h}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) - \mathcal{R}_{1,h}^{c}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)})) \nabla f_c(\theta^\star) \ .
\end{aligned}
$$

Since $\sum_{h=1}^{H} h = \frac{H(H-1)}{2}$, we obtain from above inequalities that

$$
\bar{\theta}_{\mathrm{det}}^{(\gamma,H)} - \theta^\star = \frac{\gamma(H-1)}{2N} \sum_{c=1}^{N} \nabla^2 f(\theta^\star)^{-1} (\nabla^2 f_c(\theta^\star) - \nabla^2 f(\theta^\star)) \nabla f_c(\theta^\star) + \gamma H \mathcal{R}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) \ . \qquad (55)
$$

The result follows from (51), which ensures that $\sup_{\vartheta \in \mathbb{R}^d} \|\mathcal{R}(\vartheta)\| / \|\vartheta - \theta^\star\| < +\infty$, and Theorem 7, which gives $\|\bar{\theta}_{\mathrm{det}}^{(\gamma,H)} - \theta^\star\| = O(\gamma H)$ and thus the upper bound on the remainder $\gamma H \mathcal{R}(\bar{\theta}_{\mathrm{det}}^{(\gamma,H)}) = O(\gamma^2 H^2)$. □

B Analysis of Stochastic FedAvg

B.1 Convergence to a Stationary Distribution

In the stochastic setting, we recall the following operators that generate the iterates of FedAvg. That is, for $\theta \in \mathbb{R}^d$, we let

$$
\begin{aligned}
\widetilde{\mathsf{T}}_c^{(\gamma,0)}(\theta) &\triangleq \theta \ , \\
\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) &\triangleq \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \gamma \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \ ,
\end{aligned}
$$

and define the global update

$$
\widetilde{\mathsf{T}}^{(\gamma,H)}\left(\theta; Z_{1:N}^{1:H}\right) \triangleq \frac{1}{N} \sum_{c=1}^{N} \widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_{1:H}^{c}) \ .
$$

Here $Z_{1:N}^{1:H} = \{Z_{\tilde{c}}^{\tilde{h}} : \tilde{c} \in \{1, \ldots, N\}, \tilde{h} \in \{1, \ldots, H\}\}$ is a sequence of independent random variable, such that $Z_{\tilde{c}}^{\tilde{h}}$ has distribution $\xi_{\tilde{c}}$. Additionally, FedAvg's global updates are of the form $\theta_{t+1} = \theta_t - \gamma \mathsf{G}^{(\gamma,H)}(\theta_t; Z_{1:N}^{1:H})$, where

$$
\mathsf{G}^{(\gamma,H)}(\theta; Z_{1:N}^{1:H}) = \frac{1}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \nabla F_c^{Z_{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \ , \qquad (56)
$$

where $\theta_{c,0}(Z), \theta_{c,1}(Z), \ldots, \theta_{c,H}(Z)$ is the sequence obtained using the stochastic local update rule, and $Z = (Z_1, \ldots, Z_H)$ is a sequence of i.i.d. random variables.

Contrarily to FedAvg-D, the stochastic variant of FedAvg does not converge to a single point. Thus, we rather study the convergence of its global iterates to a stationary distribution. To this end, we start with the following two lemma, that are analogous to Lemma 1 and Lemma 2 in the stochastic setting.

Lemma 3 (Contraction of FedAvg's Local Iterates)**.** *Assume A 1. Let θ, ϑ be random vectors, $\mathcal{F}$ be a σ-algebra, such that θ, ϑ are $\mathcal{F}$-measurable. Moreover, let $c \in \{1, \ldots, N\}$ and $Z_c \sim \xi_c$ be independent of $\mathcal{F}$. Then for any $\gamma \leq 1/(2L)$, it holds that*

$$\mathbb{E}\left[\|(\theta - \gamma \nabla F_c^{Z_c}(\theta)) - (\vartheta - \gamma \nabla F_c^{Z_c}(\vartheta))\|^2\right] \leq (1 - \gamma\mu)\mathbb{E}\left[\|\theta - \vartheta\|^2\right] \ . \tag{57}$$

Proof. We start by expanding the norm as

$$\begin{aligned}&\|(\theta - \gamma \nabla F_c^{Z_c}(\theta)) - (\vartheta - \gamma \nabla F_c^{Z_c}(\vartheta))\|^2 \\ &\quad = \|\theta - \vartheta\|^2 + \gamma^2 \|\nabla F_c^{Z_c}(\theta) - \nabla F_c^{Z_c}(\vartheta)\|^2 - 2\gamma\langle \theta - \vartheta, \nabla F_c^{Z_c}(\theta) - \nabla F_c^{Z_c}(\vartheta)\rangle \ .\end{aligned}$$

By the expected smoothness property (see A 1), we have

$$\mathbb{E}\left[\gamma^2 \|\nabla F_c^{Z_c}(\theta) - \nabla F_c^{Z_c}(\vartheta)\|^2 \;\middle|\; \mathcal{F}\right] \leq 2L\gamma^2 \langle \theta - \vartheta, \nabla f_c(\theta) - \nabla f_c(\vartheta)\rangle \ .$$

Then, strong convexity gives

$$\mathbb{E}\left[-\gamma\langle \theta - \vartheta, \nabla F_c^{Z_c}(\theta) - \nabla F_c^{Z_c}(\vartheta)\rangle \;\middle|\; \mathcal{F}\right] = -\gamma\langle \theta - \vartheta, \nabla f_c(\theta) - \nabla f_c(\vartheta)\rangle \leq -\gamma\mu\|\theta - \vartheta\|^2 \ .$$

Combining the above inequalities, we obtain

$$\mathbb{E}\left[\|(\theta - \gamma \nabla F_c^{Z_c}(\theta)) - (\vartheta - \gamma \nabla F_c^{Z_c}(\vartheta))\|^2 \;\middle|\; \mathcal{F}\right] \leq (1 - \gamma\mu)\|\theta - \vartheta\|^2 - \gamma(1 - 2L\gamma)\langle \theta - \vartheta, \nabla F_c^{Z_c}(\theta) - \nabla F_c^{Z_c}(\vartheta)\rangle \ ,$$

and the result follows from $\gamma \leq 1/2L$ and the tower property of conditional expectations. □

Lemma 4 (Contraction of FedAvg's Global Updates)**.** *Assume A 1. Let $H > 0$ and $Z_{1:N}^{1:H} = \{Z_{\tilde{c}}^{\tilde{h}} : \tilde{c} \in \{1, \ldots, N\}, \tilde{h} \in \{1, \ldots, H\}\}$ be a sequence of independent random variable, such that $Z_{\tilde{c}}^{\tilde{h}}$ has distribution $\xi_{\tilde{c}}$. Let $\mathcal{F}$ be a sub-σ-algebra and $\theta, \vartheta \in \mathbb{R}^d$ be two $\mathcal{F}$-measurable random variables. Then for the operator $\widetilde{\mathsf{T}}_c^{(\gamma,H)}(\cdot; Z_{1:N}^{1:H})$ it holds, for $\gamma \leq 1/(2L)$, that*

$$\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_{1:N}^{1:H}) - \widetilde{\mathsf{T}}_c^{(\gamma,H)}(\vartheta; Z_{1:N}^{1:H})\|^2\right] \leq (1 - \gamma\mu)^H \mathbb{E}\left[\|\theta - \vartheta\|^2\right] \ . \tag{58}$$

Proof. First, remark that

$$\begin{aligned}&\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\vartheta; Z_c^{1:h+1}) \\ &\quad = (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \gamma(\nabla F_c^{Z_h}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))) - (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\vartheta; Z_c^{1:h}) - \gamma \nabla F_c^{Z_h}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\vartheta; Z_c^{1:h}))) \ .\end{aligned}$$

Therefore, by Lemma 3, we have

$$\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\vartheta; Z_c^{1:h+1})\|^2\right] \leq (1 - \gamma\mu)\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\vartheta; Z_c^{1:h})\|^2\right] \ .$$

Thus, using this inequality H times recursively, together with Jensen's inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathsf{T}}^{(\gamma,H)}(\theta; Z_{1:N}^{1:H}) - \widetilde{\mathsf{T}}^{(\gamma,H)}(\vartheta; Z_{1:N}^{1:H})\|^2\right] &\leq \frac{1}{N}\sum_{c=1}^{N} \mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_c^{1:H}) - \widetilde{\mathsf{T}}_c^{(\gamma,H)}(\vartheta; Z_c^{1:H})\|^2\right] \\ &\leq (1 - \gamma\mu)^H \mathbb{E}\left[\|\theta - \vartheta\|^2\right] \ ,\end{aligned}$$

which implies the statement. □

We now use the above lemma to show that the iterates of FedAvg converge to a stationary distribution.

Proposition 5 (Proposition 3 in the main text)**.** *Assume A 1 and let $\gamma \leq 1/(2L)$. Then the iterates of* FedAvg *converge to a unique stationary distribution $\pi^{(\gamma,H)}$ with finite second moment. Furthermore, for any probability measure λ on $\mathbb{R}^d$, it holds that*

$$\mathbf{W}_2^2(\lambda\kappa^t, \pi^{(\gamma,H)}) \leq (1 - \gamma\mu)^{Ht}\mathbf{W}_2^2(\lambda, \pi^{(\gamma,H)}) \ . \tag{59}$$

Proof. The proof is inspired by Dieuleveut et al. [2020, Proposition 2]. Let λ_1, λ_2 be two probability measures on $\mathbb{R}^d$. By Villani et al. [2009], Theorem 4.1, there exists two random variables θ_0 and ϑ_0 such that

$$\mathbf{W}_2^2(\lambda_1, \lambda_2) = \mathbb{E}\left[\|\theta_0 - \vartheta_0\|^2\right] \quad . \tag{60}$$

For $t \geq 0$, let $Z^{1:H}_{1:N,t} = \{Z^{\tilde{h}}_{\tilde{c},t} : \tilde{c} \in \{1, \ldots, N\}, \tilde{h} \in \{1, \ldots, H\}, \}$ is a sequence of independent random variables, such that $Z^{\tilde{h}}_{\tilde{c},t}$ has distribution $\xi_{\tilde{c}}$, and define recursively the two sequences for $t \geq 0$,

$$\theta_{t+1} = \widetilde{\mathsf{T}}^{(\gamma,H)}(\theta_t; Z^{1:H}_{1:N,t}) \ , \qquad \vartheta_{t+1} = \widetilde{\mathsf{T}}^{(\gamma,H)}(\vartheta_t; Z^{1:H}_{1:N,t}) \ , \tag{61}$$

corresponding to two trajectories of FedAvg, sampled with the same noise but with different initializations. In the following, we use the filtration $\mathcal{F}_t = \sigma\{Z^{1:H}_{1:N,s} : s \leq t\}$. By the definition of the Wasserstein distance, and using Lemma 4, we obtain, for any $k \geq 0$,

$$\begin{aligned}
\mathbf{W}_2^2(\lambda_1 \kappa^t, \lambda_2 \kappa^t) &\leq \mathbb{E}\left[\|\theta_t - \vartheta_t\|^2\right] &(62)\\
&= \mathbb{E}\left[\mathbb{E}\left[\|\widetilde{\mathsf{T}}^{(\gamma,H)}(\theta_{t-1}; Z^{1:H}_{1:N,t}) - \widetilde{\mathsf{T}}^{(\gamma,H)}(\vartheta_{t-1}; Z^{1:H}_{1:N,t-1})\|^2 \,\middle|\, \mathcal{F}_{t-1}\right]\right] &(63)\\
&\leq (1 - \gamma\mu)^H \mathbb{E}\left[\|\theta_{t-1} - \vartheta_{t-1}\|^2\right] \quad . &(64)
\end{aligned}$$

Applying Lemma 4 resursively, we obtain

$$\mathbf{W}_2^2(\lambda_1 \kappa^t, \lambda_2 \kappa^t) \leq (1 - \gamma\mu)^{Ht} \|\theta_0 - \vartheta_0\|^2 = (1 - \gamma\mu)^{Ht} \mathbf{W}_2^2(\lambda_1, \lambda_2) \quad . \tag{65}$$

The rest of the proof follows the lines of Dieuleveut et al. [2020]. □

B.2 Crude Bounds on FedAvg's Convergence

In this section, we give crude bounds on the moments of FedAvg's stationary distribution, that will be used to bound higher-order terms in the expansions below.

B.2.1 Homogeneous Functions

For homogeneous functions, we can prove that the errors of FedAvg's global and local iterates at stationarity are of order $O(\gamma)$. This is stated in the next lemma, whose proof follows the lines of classical analysis of SGD, but only uses the fact that gradients of $\nabla f_c()$'s at solution have the same expectation.

Lemma 5 (Crude Bound, Homogeneous Functions)**.** *Assume A 1, A 3, and let A 2 holds with $\zeta_{\star,1} = 0$. Let $\gamma \leq 1/(2L)$, and $\gamma\mu H \leq 1$, then*

$$\mathbb{E}[\|\theta_t - \theta^\star\|^2] \leq (1 - 2\gamma\mu(1 - \gamma L))^{Ht} \mathbb{E}[\|\theta - \theta^\star\|^2] + \frac{\gamma}{\mu(1 - \gamma L)} M_\epsilon^{1/2} \quad .$$

This implies that, for $\theta \sim \pi^{(\gamma,H)}$, where $\pi^{(\gamma,H)}$ is the stationary distribution of FedAvg *with step size γ and H local updates, it holds that*

$$\int \|\theta - \theta^\star\|^2 \pi^{(\gamma,H)}(\mathrm{d}\theta) = O(\gamma) \ , \quad \textit{and} \ \int \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 \pi^{(\gamma,H)}(\mathrm{d}\theta) = O(\gamma) \quad .$$

Remark 2. *Lemma 5 only assumes that $\nabla f_c(\theta^\star) = 0$ for all $c \in \{1, \ldots, N\}$. This notably holds under A 5, but is in fact a stronger result.*

Proof. First, we rewrite the local updates of FedAvg, for $c \in \{1, \ldots, N\}$ and $h \in \{0, \ldots, H-1\}$,

$$\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) = \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \gamma \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \quad .$$

Thus, we have

$$\begin{aligned}
&\|\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2 \\
&\quad = \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 - 2\gamma\langle \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle + \|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|^2 \quad .
\end{aligned}$$

Decomposing the gradient of $\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))$ using the fact that, since $\zeta_{\star,1} = 0$, the functions f_c's satisfy $\nabla f_c(\theta^\star) = 0$, we obtain

$$\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) = \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star) + \nabla F_c^{Z_c^{h+1}}(\theta^\star) - \nabla f_c(\theta^\star) \ ,$$

and using Young's inequality, we obtain

$$\begin{aligned}\|\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2 \le \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 - 2\gamma\langle \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle \\ + 2\gamma^2\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^2 + 2\gamma^2\|\nabla F_c^{Z_c^{h+1}}(\theta^\star) - \nabla f_c(\theta^\star)\|^2 \ .\end{aligned}$$

Now, we define the filtration $\mathcal{F}_{h,c} = \sigma(Z_c^\ell : \ell \le h)$, and take the conditional expectation to obtain

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2 \;\middle|\; \mathcal{F}_{h,c}\right] \le \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 - 2\gamma\langle \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle \\ + 2\gamma^2\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^2 \;\middle|\; \mathcal{F}_{h,c}\right] \\ + 2\gamma^2\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\theta^\star) - \nabla f_c(\theta^\star)\|^2 \;\middle|\; \mathcal{F}_{h,c}\right] \ .\end{aligned}$$

By A 1-(a), A 1-(b), and using that $\nabla f_c(\theta^\star) = 0$, we have

$$\begin{aligned}&\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2 \;\middle|\; \mathcal{F}_{h,c}\right] \\ &\quad\le \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 - 2\gamma(1-\gamma L)\langle \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle \\ &\qquad + 2\gamma^2\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\theta^\star) - \nabla f_c(\theta^\star)\|^2 \;\middle|\; \mathcal{F}_{h,c}\right] \\ &\quad\le (1 - 2\gamma\mu(1-\gamma L))\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 + 2\gamma^2\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\theta^\star) - \nabla f_c(\theta^\star)\|^2 \;\middle|\; \mathcal{F}_{h,c}\right] \ . \qquad (66)\end{aligned}$$

Using the definition of (5), taking the expectation and unrolling the above inequality we obtain

$$\mathbb{E}[\|\widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_c^{1:H}) - \theta^\star\|^2] \le (1 - 2\gamma\mu(1-\gamma L))^H\mathbb{E}[\|\theta - \theta^\star\|^2] + 2\gamma^2 H\mathbb{E}[\|\varepsilon_c^{Z_c^{h+1}}(\theta^\star)\|^2 \ .$$

Therefore, using Jensen's inequality, A 2 and A 3, we obtain the following bound:

$$\mathbb{E}[\|\widetilde{\mathsf{T}}^{(\gamma,H)}(\theta; Z_{1:N}^{1:H}) - \theta^\star\|^2] \le (1 - 2\gamma\mu(1-\gamma L))^H\mathbb{E}[\|\theta - \theta^\star\|^2] + 4\gamma^2 H(M_\epsilon^{1/2} + \zeta_{\star,1}^2) \ . \qquad (67)$$

Denoting θ_t the global iterates of FedAvg, and using (67) recursively, we obtain

$$\mathbb{E}[\|\theta_t - \theta^\star\|^2] \le (1 - 2\gamma\mu(1-\gamma L))^{Ht}\mathbb{E}[\|\theta - \theta^\star\|^2] + \frac{2\gamma}{\mu(1-\gamma L)}M_\epsilon^{1/2} \ ,$$

which is the first part of the result. Taking $\theta \sim \pi^{(\gamma,H)}$ and using the fact that $\pi^{(\gamma,H)}$ is the stationary distribution of FedAvg's global iterates, θ_t and θ are identically distributed, then taking the limit as $t \to +\infty$ gives the second part of the result. Finally, using (66) we obtain

$$\mathbb{E}[\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2] \le \mathbb{E}[\|\theta - \theta^\star\|^2] + 2\gamma^2 h(M_\epsilon^{1/2} + \zeta_{\star,1}^2) = O(\gamma + \gamma^2 h) = O(\gamma) \ ,$$

since $\gamma h = O(1)$, which gives the last part of the result. □

Lemma 6. *Assume A 1, A 3, and let A 2 holds with $\zeta_{\star,1} = 0$. Furthermore, assume that $\mathbb{E}^{1/3}\left[\|\varepsilon_c^{Z_c}(\theta^\star)\|^6\right] \le \tau_6^2$. Let $\gamma \le 1/8L$, and $\gamma\mu H \le 1$ then there exist a universal constant $\beta > 0$ such that*

$$\mathbb{E}^{1/3}\left[\|\theta_t - \theta^\star\|^6\right] \le (1 - \gamma\mu/3)^{Ht}\mathbb{E}^{1/3}[\|\theta - \theta^\star\|^6] + \frac{3\beta\gamma}{\mu}\tau_6^2 \ .$$

Moreover, for $\theta \sim \pi^{(\gamma,H)}$, where $\pi^{(\gamma,H)}$ is the stationary distribution of FedAvg *with step size γ and H local updates, it holds that, for $p \in \{2, 3\}$,*

$$\int \|\theta - \theta^\star\|^{2p}\pi^{(\gamma,H)}(\mathrm{d}\theta) = O(\gamma^p) \ , \quad \textit{and} \ \int \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{2p}\pi^{(\gamma,H)}(\mathrm{d}\theta) = O(\gamma^p) \ .$$

Proof. We now extend the results of Lemma 5 to higher moments of $\|\theta - \theta^\star\|^2$, with $\theta \sim \pi^{(\gamma,H)}$. First, we prove a bound on the moment of order 6. To this end, we start by deriving an upper bound for local updates, decomposing the update between a contraction and an additive term due to stochasticity. Starting from a point $\theta \in \mathbb{R}^d$, we first expand the squared norm, as in the proof of Lemma 5, as

$$\begin{aligned}
&\|\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2 \\
&\quad = \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 - 2\gamma\langle \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle + \|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|^2 \enspace .
\end{aligned}$$

To reach the sixth power, we take this equation at the power three. We use the fact that, for $a, b, c \in \mathbb{R}$, it holds that $(a+b+c)^3 = a^3 + 3a^2b + 3ab^2 + b^3 + 3a^2c + 6abc + 3b^2c + 3ac^2 + 3bc^2 + c^3$. Thus,

$$(a^2 - 2\gamma b + \gamma^2 c^2)^3 = a^6 - 6\gamma a^4 b + 3\gamma^2 a^4 c^2 + 12\gamma^2 a^2 b^2 - 6\gamma^3 a^2 b c^2 + 3\gamma^4 a^2 c^4 - 8\gamma^3 b^3 + 12\gamma^4 b^2 c^2 - 6\gamma^5 b c^4 + \gamma^6 c^6 \enspace .$$

If a, b, c satisfy $|b| \le ac$, we have

$$\begin{aligned}
&(a^2 - 2\gamma b + \gamma^2 c^2)^3 \\
&\le a^6 - 6\gamma a^4 b + 3\gamma^2 a^4 c^2 + 12\gamma^2 a^4 c^2 + 6\gamma^3 a^3 c^3 + 3\gamma^4 a^2 c^4 + 8\gamma^3 a^3 c^3 + 12\gamma^4 a^2 c^4 + 6\gamma^5 a c^5 + \gamma^6 c^6 \\
&= a^6 - 6\gamma a^4 b + 15\gamma^2 a^4 c^2 + 14\gamma^3 a^3 c^3 + 15\gamma^4 a^2 c^4 + 6\gamma^5 a c^5 + \gamma^6 c^6 \enspace . \qquad (68)
\end{aligned}$$

Now, we take $a = \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|$, $b = \langle \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle$, and $c = \|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|$. Note that we indeed have $b \le ac$ using the Cauchy-Schwarz inequality.

At this point, we have the following bound, for $2 \le k \le 6$,

$$\begin{aligned}
\mathbb{E}\left[c^k \;\middle|\; \mathcal{F}_c^h\right] &= \mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|^k \;\middle|\; \mathcal{F}_c^h\right] \\
&\le 2^{k-1}\left\{\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^k \;\middle|\; \mathcal{F}_c^h\right] + \mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|^k \;\middle|\; \mathcal{F}_c^h\right]\right\} \\
&\le 2^{k-1}\left\{\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^k \;\middle|\; \mathcal{F}_c^h\right] + \tau_6^k\right\} \enspace .
\end{aligned}$$

Then, by A 1, and since $\nabla f_c(\theta^\star) = 0$, we have

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^k \;\middle|\; \mathcal{F}_c^h\right] \\
&\quad \le L^{k-2}\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{k-2}\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^k \;\middle|\; \mathcal{F}_c^h\right] \\
&\quad \le L^{k-1}\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{k-2}\langle \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle \enspace .
\end{aligned}$$

This guarantees that

$$\begin{aligned}
&\mathbb{E}\left[c^k \;\middle|\; \mathcal{F}_c^h\right] \\
&\quad \le 2^{k-1}L^{k-1}\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{k-2}\langle \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle + 2^{k-1}\tau_6^k \enspace .
\end{aligned}$$

Which in turn proves that

$$\begin{aligned}
&\mathbb{E}\left[\gamma^k a^{6-k} c^k \;\middle|\; \mathcal{F}_c^h\right] \\
&\le 2^{k-1}\gamma^k L^{k-1}\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{6-k+k-2}\langle \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle \\
&\quad + 2^{k-1}\gamma^k\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{6-k}\tau_6^k \\
&= 2^{k-1}\gamma^k L^{k-1}\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^4\langle \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle \\
&\quad + 2^{k-1}\gamma^k\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{6-k}\tau_6^k \enspace .
\end{aligned}$$

Then, we remark that

$$\mathbb{E}\left[-6\gamma a^4 b \;\middle|\; \mathcal{F}_c^h\right] \le -6\gamma\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^4\langle \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle \enspace .$$

Plugging this in the conditional expectation of (68), we obtain

$$\begin{aligned}(a^2 - 2\gamma b + \gamma^2 c^2)^3 \le a^6 + \Big(&- 6\gamma + 2\cdot 15\gamma^2 L + 4\cdot 14\gamma^3 L^2 + 8\cdot 15\gamma^4 L^3 + 16\cdot 6\gamma^5 L^4 + 32\gamma^6 L^5\Big) \\ &\times \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^4 \langle \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle \\ &+ 15 \sum_{k=2}^{6} 2^{k-1}\gamma^k \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{6-k}\tau_6^k \enspace .\end{aligned}$$

Taking $\gamma L \le 1/8$, we have $2\cdot 15\gamma^2 L + 4\cdot 14\gamma^3 L^2 + 8\cdot 15\gamma^4 L^3 + 16\cdot 6\gamma^5 L^4 + 32\gamma^6 L^5 \le 5\gamma$, which, combined with the following inequality, which holds by A 1,

$$-\gamma\langle \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})), \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\rangle \le -\gamma\mu \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 \enspace ,$$

ensures that

$$\begin{aligned}&\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^6 \,\middle|\, \mathcal{F}_c^h\right] \\ &\le (1-\gamma\mu)\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^6 + 15\sum_{k=2}^{6} 2^{k-1}\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{6-k}(\gamma\tau_6)^k \enspace .\end{aligned}$$

We now express this sum as a third-power of a sum of two terms: one contraction, and one additive term due to stochasticity. Let $k = 2\ell + 1 \in \{2, \dots, 6\}$ be an odd number, which implies $\ell = 1$ or $\ell = 2$. Since $k \ge 2$, then $\ell \ge 1$, and $k \ge 3$. Using the fact that for odd values of $k = 2\ell + 1$, then $k - 1 = 2\ell \ge 2$ is even, we have

$$\begin{aligned}\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{6-k}(\gamma\tau_6)^k &= \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{5-2\ell}(\gamma\tau_6)^{2\ell+1} \\ &= \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{4-2\ell}(\gamma\tau_6)^{2\ell}\left(\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|\gamma\tau_6\right) \\ &\le \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{4-2\ell}(\gamma\tau_6)^{2\ell}\left(2\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 + 2\gamma^2\tau_6^2\right) \enspace .\end{aligned}$$

Using this fact, as well as the above inequalities, Hölder's inequality, and following the lines of proof of Dieuleveut et al. [2020]'s Lemma 13, there exists a constant $\beta > 0$ such that

$$\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^6\right] \le \left((1-\gamma\mu/3)\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^6\right]^{1/3} + \beta\gamma^2\tau_6^2\right)^3 \enspace .$$

Consequently, we have

$$\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^6\right]^{1/3} \le (1-\gamma\mu/3)\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^6\right]^{1/3} + \beta\gamma^2\tau_6^2 \enspace .$$

Iterating this for H iterations, we obtain that

$$\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_c^{1:H}) - \theta^\star\|^6\right]^{1/3} \le (1-\gamma\mu/3)^H\mathbb{E}\left[\|\theta - \theta^\star\|^6\right]^{1/3} + \beta H\gamma^2\tau_6^2 \enspace .$$

Then, combining this inequality with Minkowski's inequality, we obtain, for any $\theta \in \mathbb{R}^d$,

$$\begin{aligned}\mathbb{E}\left[\|\widetilde{\mathsf{T}}^{(\gamma,H)}(\theta; Z_{1:N,t}^{1:H}) - \theta^\star\|^6\right]^{1/3} &\le \frac{1}{N}\sum_{c=1}^{N}\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_{1:N,t}^{1:H}) - \theta^\star\|^6\right]^{1/3} \\ &\le (1-\gamma\mu/3)^H\mathbb{E}\left[\|\theta - \theta^\star\|^6\right]^{1/3} + \beta H\gamma^2\tau_6^2 \enspace ,\end{aligned}$$

and the first part of the result follows from iterating this inequality T times, starting from θ_T.

The second part of the result for $p = 3$ directly follows from the previous inequality. To obtain the result for $p = 2$, we use Hölder inequality and remark that

$$\begin{aligned}\int \|\theta - \theta^\star\|^4 \pi^{(\gamma,H)}(\mathrm{d}\theta) &= \int \|\theta - \theta^\star\| \times \|\theta - \theta^\star\|^3 \pi^{(\gamma,H)}(\mathrm{d}\theta) \\ &\le \left(\int \|\theta - \theta^\star\|^2 \pi^{(\gamma,H)}(\mathrm{d}\theta)\right)^{1/2}\left(\int \|\theta - \theta^\star\|^6 \pi^{(\gamma,H)}(\mathrm{d}\theta)\right)^{1/2} \\ &= O\left((\gamma \times \gamma^3)^{1/2}\right) \enspace ,\end{aligned}$$

where the last equality comes from Lemma 5 and from the first part of this Lemma. □

B.2.2 Heterogeneous Functions

Lemma 7. *Assume A 1, A 2, A 3, let $\gamma \le 1/L$, and $\gamma\mu H \le 1$. Then we have*

$$\mathbb{E}\left[\|\theta_t - \theta^\star\|^2\right] \le \left(1 - \frac{\gamma\mu}{2}\right)^{Ht} \|\theta - \theta^\star\|^2 + \frac{H(H-1)}{\mu}\left(4\gamma^3 L^2 + \frac{2\gamma^2 L^2}{\mu}\right)\zeta_{\star,1}^2 + \frac{8\gamma}{\mu} M_\epsilon^{1/2} \ .$$

This implies that, for $\theta \sim \pi^{(\gamma,H)}$, where $\pi^{(\gamma,H)}$ is the stationary distribution of FedAvg *with step size γ and H local updates, it holds that*

$$\int \|\theta - \theta^\star\|^2 \pi^{(\gamma,H)}(\mathrm{d}\theta) = O(\gamma + \gamma^2 H^2) \ , \quad \text{and} \int \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 \pi^{(\gamma,H)}(\mathrm{d}\theta) = O(\gamma + \gamma^2 H^2) \ . \tag{69}$$

Proof. We start from $\theta_{t+1} = \theta_t - \gamma \mathsf{G}^{(\gamma,H)}(\theta; Z_{1:N}^{1:H})$, with $\mathsf{G}^{(\gamma,H)}(\theta; Z_{1:N}^{1:H})$ as defined in Section 5, and use $\frac{1}{N}\sum_{c=1}^N \nabla f_c(\theta^\star) = 0$, to obtain

$$\theta_{t+1} = \theta_t - \frac{\gamma}{N}\sum_{c=1}^N \sum_{h=0}^{H-1} \left\{ \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star) \right\} \ .$$

Using Jensen's inequality, we have

$$\|\theta_{t+1} - \theta^\star\|^2 \le \frac{1}{N}\sum_{c=1}^N \left\| \theta_t - \frac{\gamma}{N}\sum_{c=1}^N \sum_{h=0}^{H-1} \left\{ \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star) \right\} \right\|^2 \ .$$

To derive an upper bound on this value, we study the following sequence of iterates, that correspond to the local parameters with recentered gradients, defined for $h \in \{0, \dots, H-1\}$,

$$\widetilde{\mathsf{V}}_c^{(\gamma,H)}(\theta; Z_c^{1:H}) \triangleq \widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_c^{1:H}) - \gamma h \nabla f_c(\theta^\star) \ , \tag{70}$$

which allows to rewrite the above inequality as $\|\theta_{t+1} - \theta^\star\|^2 \le \frac{1}{N}\sum_{c=1}^N \|\widetilde{\mathsf{V}}_c^{(\gamma,H)}(\theta; Z_c^{1:H}) - \theta^\star\|^2$. Next, we bound each term of this sum independently. We start by expanding the norm

$$\begin{aligned}
\|\widetilde{\mathsf{V}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2 &= \|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star - \gamma(\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star))\|^2 \\
&= \|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 + \gamma^2 \|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\|^2 \\
&\quad - 2\gamma \langle \widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star) \rangle \ .
\end{aligned}$$

We now take the expectation using the filtration $\mathcal{F}_c^h = \sigma(Z_c^\ell : \ell \le h)$, for $h \in \{0, \dots, H-1\}$,

$$\begin{aligned}
\mathbb{E}\left[\|\widetilde{\mathsf{V}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2 \,\middle|\, \mathcal{F}_c^h \right] &= \|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 \\
&\quad + \gamma^2 \mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\|^2 \,\middle|\, \mathcal{F}_c^h \right] \\
&\quad - 2\gamma \langle \widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star) \rangle \ .
\end{aligned}$$

Now, we remark that

$$\begin{aligned}
\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star) &= \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star) \\
&\quad + \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \nabla F_c^{Z_c^{h+1}}(\theta^\star) - \nabla f_c(\theta^\star) \ ,
\end{aligned}$$

which allows to decompose the term $\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\|^2 \,\middle|\, \mathcal{F}_c^h \right]$ using Young's inequality twice, followed by A 1 and A 3,

$$
\begin{aligned}
&\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\|^2 \;\middle|\; \mathcal{F}_c^h\right] \\
&\quad \le 2\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^2 \;\middle|\; \mathcal{F}_c^h\right] \\
&\quad + 4\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|^2 \;\middle|\; \mathcal{F}_c^h\right] + 4\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\theta^\star) - \nabla f_c(\theta^\star)\|^2 \;\middle|\; \mathcal{F}_c^h\right] \\
&\quad \le 2\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^2 \;\middle|\; \mathcal{F}_c^h\right] + 4L^2\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})\|^2 + 4M_\epsilon^{1/2} \\
&\quad = 2\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^2 \;\middle|\; \mathcal{F}_c^h\right] + 4L^2\gamma^2h^2\|\nabla f_c(\theta^\star)\|^2 + 4M_\epsilon^{1/2} \ ,
\end{aligned}
$$

where the last equality comes from the definition of $\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})$. Furthermore, we have

$$
\begin{aligned}
\gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\rangle &= \gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla f_c(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\rangle \\
&\quad + \gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\rangle
\end{aligned}
$$

Then, we bound the second term using Young's inequality A 1 and the definition of $\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})$,

$$
\begin{aligned}
&-2\gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\rangle \\
&\quad \le \frac{\gamma\mu}{2}\|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 + \frac{2\gamma}{\mu}\|\nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|^2 \\
&\quad \le \frac{\gamma\mu}{2}\|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 + \frac{2\gamma^3h^2L^2}{\mu}\|\nabla f_c(\theta^\star)\|^2 \ .
\end{aligned}
$$

Finally, we remark that

$$
\begin{aligned}
&\|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 + 2\mathbb{E}\left[\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^2 \;\middle|\; \mathcal{F}_c^h\right] \\
&\quad - 2\gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla f_c(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\rangle \\
&\quad \le (1-\gamma\mu)\|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 \ .
\end{aligned}
$$

We now plug the bounds we obtained in the expansion above to obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|\widetilde{\mathsf{V}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2 \;\middle|\; \mathcal{F}_c^h\right] \\
&\quad \le \left(1 - \frac{\gamma\mu}{2}\right)\|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 + \left(4\gamma^4h^2L^2 + \frac{2\gamma^3h^2L^2}{\mu}\right)\|\nabla f_c(\theta^\star)\|^2 + 4\gamma^2M_\epsilon^{1/2} \ .
\end{aligned}
$$

Taking the expectation and unrolling the inequality, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|\widetilde{\mathsf{V}}_c^{(\gamma,H)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2\right] \\
&\quad \le \left(1 - \frac{\gamma\mu}{2}\right)^H\|\theta - \theta^\star\|^2 + \frac{H^2(H-1)}{2}\left(4\gamma^4L^2 + \frac{2\gamma^3L^2}{\mu}\right)\|\nabla f_c(\theta^\star)\|^2 + 4\gamma^2HM_\epsilon^{1/2} \ .
\end{aligned}
$$

Which gives the following inequality, that links two consecutive updates,

$$
\begin{aligned}
&\mathbb{E}\left[\|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_{1:N}^{1:H}) - \theta^\star\|^2\right] \\
&\quad \le \left(1 - \frac{\gamma\mu}{2}\right)^H\|\theta - \theta^\star\|^2 + \frac{H^2(H-1)}{2}\left(4\gamma^4L^2 + \frac{2\gamma^3L^2}{\mu}\right)\zeta_{\star,1}^2 + 4\gamma^2HM_\epsilon^{1/2} \ .
\end{aligned}
$$

Unrolling this inequality starting from a point $\theta_0 \in \mathbb{R}^d$, we obtain

$$
\mathbb{E}\left[\|\theta_t - \theta^\star\|^2\right] \le \left(1 - \frac{\gamma\mu}{2}\right)^{Ht}\|\theta - \theta^\star\|^2 + \frac{H(H-1)}{\mu}\left(4\gamma^3L^2 + \frac{2\gamma^2L^2}{\mu}\right)\zeta_{\star,1}^2 + \frac{8\gamma}{\mu}M_\epsilon^{1/2} \ ,
$$

which gives the first part of the Lemma. The second part follows the same lines as the second part of Lemma 5. □

Lemma 8. *Assume A 1, A 2 and A 3. Furthermore, assume that* $\mathbb{E}^{1/3}\left[\|\varepsilon_c^{Z_c}(\theta^\star)\|^6\right] \leq \tau_6^2$*. Let* $\gamma \leq 1/8L$*, and* $\gamma\mu H \leq 1$ *then there exist a universal constant* $\beta > 0$ *suc that*

$$\mathbb{E}^{1/3}\left[\|\theta_t - \theta^\star\|^6\right] \leq (1-\gamma\mu/3)^{Ht}\mathbb{E}^{1/3}[\|\theta - \theta^\star\|^6] + \frac{6\beta\gamma^2L^2H(H-1)}{\mu^2}\zeta_{\star,1}^2 + \frac{8\gamma}{\mu}M_\epsilon^{1/2} + \frac{3\beta\gamma}{\mu}\tau_6^2 \ .$$

This implies that, for $\theta \sim \pi^{(\gamma,H)}$*, where* $\pi^{(\gamma,H)}$ *is the stationary distribution of* FedAvg *with step size* γ *and* H *local updates, it holds that, for* $p \in \{2,3\}$*,*

$$\int \|\theta - \theta^\star\|^{2p}\pi^{(\gamma,H)}(\mathrm{d}\theta) = O\left(\gamma^p + \gamma^{2p}H^{2p}\right) \ , \quad \text{and} \int \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^{2p}\pi^{(\gamma,H)}(\mathrm{d}\theta) = O\left(\gamma^p + \gamma^{2p}H^{2p}\right) \ .$$

Proof. The proof follows the same lines as the proof of Lemma 6, with an additional heterogeneity term that is $O(\gamma^2H^2)$ that plays a role similar to the one of τ_6. We start with the expansion of the local updates, recentered by $\gamma h\nabla f_c(\theta^\star)$, as defined in Equation (70), in the proof of Lemma 7,

$$\begin{aligned}
\|\widetilde{\mathsf{V}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2 &= \|\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star - \gamma(\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star))\|^2 \\
&= \|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 + \gamma^2\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\|^2 \\
&\quad - 2\gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\rangle \\
&= \|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 + \gamma^2\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\|^2 \\
&\quad - 2\gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\rangle \\
&\quad - 2\gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\rangle \ .
\end{aligned}$$

We first bound the following quantity, following the derivations from Lemma 7,

$$\begin{aligned}
&\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\|^2 \\
&\quad \leq 2\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\theta^\star)\|^2 \\
&\qquad + 4\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|^2 + 4\|\nabla F_c^{Z_c^{h+1}}(\theta^\star) - \nabla f_c(\theta^\star)\|^2 \ ,
\end{aligned}$$

Then, we bound the last term, without the expectation, and with slightly different constants,

$$\begin{aligned}
&-2\gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\rangle \\
&\quad \leq \frac{\gamma\mu}{6}\|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 + \frac{6\gamma}{\mu}\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|^2 \ .
\end{aligned}$$

Using the same derivations as in the proof of Lemma 7, we obtain

$$\begin{aligned}
&\|\widetilde{\mathsf{V}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star\|^2 \\
&\leq (1+\gamma\mu/6)\|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 - 2\gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\rangle \\
&\quad + 2\gamma^2\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\|^2 \\
&\quad + \frac{10\gamma}{\mu}\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|^2 + 4\gamma^2\|\nabla F_c^{Z_c^{h+1}}(\theta^\star) - \nabla f_c(\theta^\star)\|^2 \ ,
\end{aligned}$$

where we also used $4\gamma^2 \leq \frac{4\gamma}{L} \leq \frac{4\gamma}{\mu}$. Then, we expand the third moment of this equation, using the same derivations as in Lemma 6, with

$$\begin{aligned}
a^2 &= (1+\gamma\mu/6)\|\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\|^2 \ , \\
-2\gamma b &= -2\gamma\langle\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star, \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\rangle \\
\gamma^2c^2 &= 2\gamma^2\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla f_c(\theta^\star)\|^2 \\
&\quad + \frac{10\gamma}{\mu}\|\nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \nabla F_c^{Z_c^{h+1}}(\widetilde{\mathsf{V}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\|^2 + 4\gamma^2\|\nabla F_c^{Z_c^{h+1}}(\theta^\star) - \nabla f_c(\theta^\star)\|^2 \ ,
\end{aligned}$$

and use the fact that $(1+\gamma\mu/6)^3 \leq (1+\gamma\mu/2)$. We notice that we indeed have $-\gamma b \leq -\frac{\gamma\mu}{2}a^2$, and that $b \leq ac$. Then, using the same derivation as in Lemma 6, we obtain the result of the lemma. □

B.3 Quadratic Setting

B.3.1 Study of the Bias

In this section, we study the particular case where the functions f_c's are quadratic. Specifically, we assume that there exist symmetric matrices $\bar{A}_c$'s and vectors $\theta_c^\star$'s such that

$$f_c(\theta) = \frac{1}{2}\left\|(\bar{A}_c)^{1/2}(\theta - \theta_c^\star)\right\|^2 \ . \tag{71}$$

This implies that f_c's gradients are linear, and satisfy $\nabla f_c(\theta) = \bar{A}_c(\theta - \theta_c^\star)$. Consequently, for all $h \le H$,

$$\mathbb{E}[\widetilde{\mathsf{T}}_c^{(\gamma,H)}\left(\theta; Z_c^{1:H}\right)] - \theta_c^\star = (\mathrm{Id} - \gamma\bar{A}_c)^h(\theta - \theta_c^\star) \ . \tag{72}$$

For further analysis, we introduce the notations

$$\Gamma_c^\star = (\mathrm{Id} - \gamma\bar{A}_c)^H \ , \quad \Gamma^\star = \frac{1}{N}\sum_{c=1}^N \Gamma_c^\star \ . \tag{73}$$

Proposition 6 (Bias of FedAvg for Quadratics)**.** *Assume A 1, A 2, A 3, and A 4 and $\gamma \le 1/L$, then the bias of* FedAvg *with quadratic functions is*

$$\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} = \theta^\star + (\mathrm{Id} - \Gamma^\star)^{-1} \cdot \frac{1}{N}\sum_{c=1}^N (\mathrm{Id} - \Gamma_c^\star)(\theta^\star - \theta_c^\star) \ . \tag{74}$$

Furthermore, when $\gamma\mu H \le 1$, it holds that

$$\left\|\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} - \theta^\star\right\| \le \frac{\gamma(H-1)\zeta_{\star,2}\zeta_{\star,1}}{2\mu} \ . \tag{75}$$

Proof. For any point $\theta \in \mathbb{R}^d$, it holds that

$$\mathbb{E}[\widetilde{\mathsf{T}}^{(\gamma,H)}\left(\theta; Z_{1:N}^{1:H}\right) - \theta^\star] = \frac{1}{N}\sum_{c=1}^N \mathbb{E}\left[\widetilde{\mathsf{T}}_c^{(\gamma,H)}\left(\theta; Z_c^{1:H}\right)\right] - \theta_c^\star - (\theta^\star - \theta_c^\star) \tag{76}$$

$$= \frac{1}{N}\sum_{c=1}^N \Gamma_c^\star(\theta - \theta_c^\star) - (\theta^\star - \theta_c^\star) \tag{77}$$

$$= \frac{1}{N}\sum_{c=1}^N \Gamma_c^\star(\theta - \theta^\star) + (\Gamma_c^\star - \mathrm{Id})(\theta^\star - \theta_c^\star) \ . \tag{78}$$

When $\theta \sim \pi^{(\gamma)}$ is sampled from the stationary distribution of FedAvg's iterates, we have $\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} = \mathbb{E}[\theta] = \mathbb{E}[\widetilde{\mathsf{T}}_H^{(Z)}\theta]$. This gives the equation

$$\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} - \theta^\star = \Gamma^\star(\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} - \theta^\star) + \frac{1}{N}\sum_{c=1}^N (\Gamma_c^\star - \mathrm{Id})(\theta^\star - \theta_c^\star) \ . \tag{79}$$

Subtracting $\Gamma^\star(\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} - \theta_c^\star)$ on both side, and multiplying by $(\mathrm{Id} - \Gamma^\star)^{-1}$, we obtain the following expression for $\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)}$ as a function of $\theta^\star$,

$$\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} = \theta^\star + (\mathrm{Id} - \Gamma^\star)^{-1} \cdot \frac{1}{N}\sum_{c=1}^N (\mathrm{Id} - \Gamma_c^\star)(\theta_c^\star - \theta^\star) \ , \tag{80}$$

which gives the first part of the result. Then, using the Neumann series together with Lemma 9, we obtain

$$\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} = \theta^\star + \sum_{t=0}^\infty (\Gamma^\star)^t \cdot \frac{1}{N}\sum_{c=1}^N \sum_{h=0}^H \gamma\Gamma_c^{\star,h+1:H}\bar{A}_c(\theta^\star - \theta_c^\star) \tag{81}$$

$$= \theta^\star + \sum_{t=0}^\infty (\Gamma^\star)^t \cdot \frac{1}{N}\sum_{c=1}^N \sum_{h=0}^H \gamma\left(\Gamma_c^{\star,h+1:H} - \Gamma_{\mathrm{avg}}^{\star,h+1:H}\right)\bar{A}_c(\theta^\star - \theta_c^\star) \ , \tag{82}$$

where we defined the notation $\Gamma_{\mathrm{avg}}^{\star,h+1:H} = \prod_{h+1}^{H}(\mathrm{Id} - \gamma \bar{A})$, and the second inequality comes from the fact that $\Gamma_{\mathrm{avg}}^{\star,h+1:H} \sum_{c=1}^{N} \bar{A}_c(\theta^\star - \theta_c^\star) = 0$. Now, we note that

$$\Gamma_c^{\star,h+1:H} - \Gamma_{\mathrm{avg}}^{\star,h+1:H} = \sum_{\ell=h+1}^{H} \Gamma_c^{\star,h+1:\ell-1}(\gamma \bar{A}_c - \gamma \bar{A})\Gamma_{\mathrm{avg}}^{\star,\ell+1:H} \enspace . \tag{83}$$

Therefore, we have

$$\begin{aligned}
\left\| \bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} - \theta^\star \right\| &\leq \sum_{t=0}^{\infty} (1-\gamma\mu)^{Ht} \cdot \sum_{h=0}^{H} \left\| \frac{1}{N} \sum_{c=1}^{N} \gamma \left(\Gamma_c^{\star,h+1:H} - \Gamma_{\mathrm{avg}}^{\star,h+1:H} \right) \bar{A}_c(\theta^\star - \theta_c^\star) \right\| && (84)\\
&= \sum_{t=0}^{\infty} (1-\gamma\mu)^{Ht} \cdot \sum_{h=0}^{H} \left\| \frac{1}{N} \sum_{c=1}^{N} \gamma \sum_{\ell=h+1}^{H} \Gamma_c^{\star,h+1:\ell-1}(\gamma \bar{A}_c - \gamma \bar{A}) \Gamma_{\mathrm{avg}}^{\star,\ell+1:H} \bar{A}_c(\theta^\star - \theta_c^\star) \right\| && (85)\\
&\leq \sum_{t=0}^{\infty} (1-\gamma\mu)^{Ht} \cdot \gamma^2 \sum_{h=0}^{H} \sum_{\ell=h+1}^{H} \left\| \frac{1}{N} \sum_{c=1}^{N} \Gamma_c^{\star,h+1:\ell-1}(\bar{A}_c - \bar{A}) \Gamma_{\mathrm{avg}}^{\star,\ell+1:H} \bar{A}_c(\theta^\star - \theta_c^\star) \right\| \enspace . && (86)
\end{aligned}$$

And we obtain

$$\begin{aligned}
&\left\| \bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} - \theta^\star \right\| && (87)\\
&\leq \sum_{t=0}^{\infty} (1-\gamma\mu)^{Ht} \cdot \gamma^2 \sum_{h=0}^{H} \sum_{\ell=h+1}^{H} \left(\frac{1}{N} \sum_{c=1}^{N} \left\| \Gamma_c^{\star,h+1:\ell-1}(\bar{A}_c - \bar{A}) \Gamma_{\mathrm{avg}}^{\star,\ell+1:H} \right\|^2 \right)^{1/2} \left(\frac{1}{N} \sum_{c=1}^{N} \| \bar{A}_c(\theta^\star - \theta_c^\star) \| \right)^{1/2} && (88)\\
&\leq \sum_{t=0}^{\infty} (1-\gamma\mu)^{Ht} \gamma^2 \frac{H(H-1)}{2} \zeta_{\star,2}\zeta_{\star,1} = \frac{\gamma(H-1)\zeta_{\star,2}\zeta_{\star,1}}{2\mu} \enspace , && (89)
\end{aligned}$$

which is the second part of the result. □

Corollary 4. *Assume A 1, A 2, A 3, A 4, $\gamma \leq 1/L$ and $\gamma H \leq 1$, then we can express $\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)}$ as*

$$\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} - \theta^\star = -\frac{\gamma(H-1)}{2N} \nabla^2 f(\theta^\star)^{-1} \sum_{c=1}^{N} (\nabla^2 f_c(\theta^\star) - \nabla^2 f(\theta^\star)) \nabla f_c(\theta^\star) + O(\gamma^2 H^2) \enspace , \tag{90}$$

Proof. The proof follows the same lines as Proposition 4, using the fact that $\nabla^2 f_c(\theta) = \bar{A}_c$ and $\nabla^2 f(\theta) = \frac{1}{N}\sum_{c=1}^{N} \bar{A}_c$ for all $\theta \in \mathbb{R}^d$ to replace $\mathit{\Gamma}^\star$ with $\Gamma^\star$, $\mathit{\Gamma}_c^{\star,h+1:H}$ with $\Gamma_c^{\star,h+1:H}$ and $\mathit{\Gamma}_{\mathrm{avg}}^{\star,h+1:H}$ with $\Gamma_{\mathrm{avg}}^{\star,h+1:H}$ for $h \in \{0, \ldots, H-1\}$. □

B.4 General Functions, with Homogeneous Agents

When functions are not quadratic, local iterates are inherently biased. We start in the simpler case where agents are homogeneous, which will serve as a skeleton for the general heterogeneous case. In this setting, the functions f_c are all identical, therefore we simply denote them f.

To study this case, we define the following matrices, for $h = 0$ to H,

$$\Gamma^{\star,h} = \left(\mathrm{Id} - \gamma \nabla^2 f(\theta^\star) \right)^h \enspace . \tag{91}$$

Crucially, in the homogeneous setting, all agents have the same local matrices. This will not be the case anymore in the next section, where agents will be heterogeneous.

Expansion of Local Updates (Homogeneous Case). We start by studying the local iterates of the algorithm, when starting from a point θ drawn from the local distribution of FedAvg. Using a second-order Taylor expansion of the gradient of ∇f at $\theta^\star$, we have

$$\nabla f(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \tag{92}$$

$$= \nabla f(\theta^\star) + \nabla^2 f(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star) + \frac{1}{2}\nabla^3 f(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star)^{\otimes 2} + \mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \tag{93}$$

$$= \nabla^2 f(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star) + \frac{1}{2}\nabla^3 f(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star)^{\otimes 2} + \mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \ , \tag{94}$$

where we used $\nabla f_c(\theta^\star) = 0$ due to homogeneity, and $\mathcal{R}_{3,h}^c$ is a function that satisfies

$$\sup_{\theta \in \mathbb{R}^d} \|\mathcal{R}_{3,h}^c(\theta)\| / \|\theta - \theta^\star\|^3 < +\infty \ . \tag{95}$$

We stress here that, although the local functions are all the same, the noise variables drawn by each agent are different from each other. Consequently, local iterates are different from each other.

We can use the above expression to expand FedAvg's recursion as

$$\begin{aligned}
&\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star \\
&\quad = \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star - \gamma \nabla f(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \gamma \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \\
&\quad = \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star \\
&\qquad - \gamma \left(\nabla^2 f(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star) + \frac{1}{2}\nabla^3 f(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star)^{\otimes 2} + \mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right) \\
&\qquad - \gamma \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \\
&\quad = \left(\mathrm{Id} - \gamma \nabla^2 f(\theta^\star) \right) (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star) \\
&\qquad - \frac{\gamma}{2}\nabla^3 f(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star)^{\otimes 2} - \gamma \mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \gamma \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \ .
\end{aligned}$$

Unrolling this recursion, we obtain

$$\begin{aligned}
&\widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_c^{1:H}) - \theta^\star = \Gamma_c^{\star,H}(\theta - \theta^\star) \\
&\quad - \gamma \sum_{h=0}^{H-1} \Gamma^{\star,H-h-1} \left(\frac{1}{2}\nabla^3 f(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star)^{\otimes 2} + \mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right) \ .
\end{aligned} \tag{96}$$

Expansion of $\mathbb{E}\left[(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star)^{\otimes 2}\right]$ (Homogeneous Case). We start with the expression

$$\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star = \theta - \theta^\star - \gamma \sum_{\ell=0}^{h-1} \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell})) + \varepsilon_c^{Z_c^{\ell+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell})) \ . \tag{97}$$

We use second-order Taylor expansion of the gradient to obtain

$$\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star = \theta - \theta^\star - \gamma \sum_{\ell=0}^{h-1} \nabla^2 f_c(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell}) - \theta^\star) + \mathcal{R}_{2,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell})) + \varepsilon_c^{Z_c^{\ell+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell})) \ ,$$

where $\mathcal{R}_{2,h}^c$ is such that $\sup_{\vartheta \in \mathbb{R}^d} \|\mathcal{R}_{2,h}^c(\vartheta)\| / \|\vartheta - \theta^\star\|^2 < +\infty$. Expanding the square of this equation, and taking the expectation, we get

$$\begin{aligned}
&\int \mathbb{E}\left(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star \right)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\theta) = \int (\theta - \theta^\star)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\theta) \\
&\quad - \gamma \int (\theta - \theta^\star) \otimes \left(\sum_{\ell=0}^{h-1} \nabla^2 f_c(\theta^\star)(\mathbb{E}\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell}) - \theta^\star) + \mathbb{E}\mathcal{R}_{2,\ell}^c(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell})) \right) \pi^{(\gamma,H)}(\mathrm{d}\theta) \\
&\quad - \gamma \int \left(\sum_{\ell=0}^{h-1} \nabla^2 f_c(\theta^\star)(\mathbb{E}\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell}) - \theta^\star) + \mathbb{E}\mathcal{R}_{2,\ell}^c(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell})) \right) \otimes (\theta - \theta^\star) \pi^{(\gamma,H)}(\mathrm{d}\theta) \\
&\quad + \gamma^2 \int \mathbb{E} \left(\sum_{\ell=0}^{h-1} \nabla^2 f_c(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell}) - \theta^\star) + \mathcal{R}_{2,\ell}^c(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell})) + \varepsilon_c^{Z_c^{\ell+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta; Z_c^{1:\ell})) \right)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\theta) \ .
\end{aligned} \tag{98}$$

From this expansion, Hölder inequality, the definition of $\mathcal{R}^c_{2,\ell}$, the bound $\gamma H \leq 1$, A 3, Lemma 6, and the fact that the $Z_c^{1:H}$ are independent from an agent to another, we obtain

$$\int \mathbb{E}\left(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\right)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\theta) = \int (\theta - \theta^\star)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\theta) + O(\gamma^2 h) \ . \tag{99}$$

Expression of the Global Update (Homogeneous Case). After averaging the expression obtained for the local updates, we get an expression of the global update,

$$\begin{aligned}
&\widetilde{\mathsf{T}}^{(\gamma,H)}(\theta; Z_{1:N}^{1:H}) - \theta^\star = \Gamma^{\star,H}(\theta - \theta^\star) \\
&\quad - \frac{\gamma}{N}\sum_{c=1}^{N}\sum_{h=0}^{H-1} \Gamma^{\star,H-h-1}\left(\frac{1}{2}\nabla^3 f(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h} - \theta^\star)^{\otimes 2} + \mathcal{R}^c_{3,h}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) + \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})\right) \ .
\end{aligned}$$

Integrating over $\pi^{(\gamma,H)}$ and taking the expectation, we obtain

$$\begin{aligned}
&\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star = \Gamma^{\star,H}(\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star) \\
&\quad - \frac{\gamma}{N}\sum_{c=1}^{N}\sum_{h=0}^{H-1} \Gamma^{\star,H-h-1}\int \left\{\frac{1}{2}\nabla^3 f(\theta^\star)\mathbb{E}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star)^{\otimes 2} + \mathbb{E}\mathcal{R}^c_{3,h}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})\right\} \pi^{(\gamma,H)}(\mathrm{d}\theta) \ .
\end{aligned}$$

Using the expression (99), Hölder inequality, Lemma 6, and the definition of $\mathcal{R}^c_{3,h}$, we can simplify this expression as

$$\left(\mathrm{Id} - \Gamma^{\star,H}\right)\left(\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star\right) = -\frac{\gamma}{2}\sum_{h=0}^{H-1}\Gamma^{\star,H-h-1}\nabla^3 f(\theta^\star)\int(\theta - \theta^\star)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta) + O(\gamma^2 h) + O(\gamma^{3/2}) \ , \tag{100}$$

To give a simpler expression, we remark that Lemma 9 gives the following equality

$$-\frac{\gamma}{2}\sum_{h=0}^{H-1}\Gamma^{\star,H-h-1} = \frac{1}{2}\left(\mathrm{Id} - \Gamma^{\star,H}\right)\nabla^2 f(\theta^\star)^{-1} \ . \tag{101}$$

Therefore, starting from the previous equation, reorganizing the terms and using this equality, we obtain

$$\left(\mathrm{Id} - \Gamma^{\star,H}\right)\left(\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star\right) = \frac{1}{2}(\mathrm{Id} - \Gamma^{\star,H})\left\{\nabla^2 f(\theta^\star)^{-1}\nabla^3 f(\theta^\star)\int (\theta - \theta^\star)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta) + O(\gamma^2 h) + O(\gamma^{3/2})\right\} \ .$$

Multiplying by $(\mathrm{Id} - \Gamma^{\star,H})^{-1}$, we obtain

$$\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star = \frac{1}{2}\nabla^2 f(\theta^\star)^{-1}\nabla^3 f(\theta^\star)\int (\theta - \theta^\star)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta) + O(\gamma^2 H) + O(\gamma^{3/2}) \ . \tag{102}$$

Bound the Variance (Homogeneous Case). To bound $\int (\theta - \theta^\star)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta)$, we proceed as above but with one less term in the expansion, and study the square. We get

$$\begin{aligned}
&\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star \\
&\quad = \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star - \gamma\left(\nabla^2 f(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star) + \mathcal{R}^c_2(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\right) - \gamma\varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \\
&\quad = \left(\mathrm{Id} - \gamma\nabla^2 f(\theta^\star)\right)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star) - \gamma\mathcal{R}^c_{2,h}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \gamma\varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \ .
\end{aligned}$$

Unrolling this recursion and averaging over all agents, we get

$$\widetilde{\mathsf{T}}^{(\gamma,H)}(\theta; Z_{1:N}^{1:H}) - \theta^\star = \Gamma^{\star,H}(\theta - \theta^\star) - \frac{\gamma}{N}\sum_{c=1}^{N}\sum_{h=0}^{H-1}\Gamma^{\star,H-h-1}\left\{\mathcal{R}^c_{2,h}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\right\} \ .$$

Taking the second order moment of this equation, and using the fact that $\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1})$ follows the same distribution as θ, we obtain

$$
\begin{aligned}
&\int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) \\
&\quad = \int \left(\Gamma^{\star,H}(\theta - \theta^\star) - \frac{\gamma}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \Gamma^{\star,H-h-1} \left\{ \mathcal{R}_{2,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right\} \right)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\theta) \\
&\quad = \int \left(\Gamma^{\star,H}(\theta - \theta^\star) \right)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\theta) \\
&\qquad - \frac{\gamma}{N} \sum_{c=1}^{N} \int \left(\Gamma^{\star,H}(\theta - \theta^\star) \right) \otimes \left(\sum_{h=0}^{H-1} \Gamma^{\star,H-h-1} \left\{ \mathcal{R}_{2,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right\} \right) \pi^{(\gamma,H)}(\mathrm{d}\theta) \\
&\qquad - \frac{\gamma}{N} \sum_{c=1}^{N} \int \left(\sum_{h=0}^{H-1} \Gamma^{\star,H-h-1} \left\{ \mathcal{R}_{2,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right\} \right) \otimes \left(\Gamma^{\star,H}(\theta - \theta^\star) \right) \pi^{(\gamma,H)}(\mathrm{d}\theta) \\
&\qquad + \frac{\gamma^2}{N^2} \int \left(\sum_{c=1}^{N} \sum_{h=0}^{H-1} \Gamma^{\star,H-h-1} \left\{ \mathcal{R}_{2,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right\} \right)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\theta) \ .
\end{aligned}
$$

Which gives, using Hölder inequality, Lemma 6, A 3, the definition of $\mathcal{R}_{2,h}^c$, the definition of $\mathcal{C}$, and after taking the expectation,

$$
\int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) = \Gamma^{\star,H} \int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) \Gamma^{\star,H} + \frac{\gamma^2}{N} \sum_{h=0}^{H-1} \mathbb{E}\mathcal{C}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + O(\gamma^{5/2} H) \ .
$$

Now, using A 3 and Lemma 6, we have $\mathbb{E}\mathcal{C}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) = \mathcal{C}(\theta^\star) + O(\gamma)$, which results in the identity

$$
\int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) = \Gamma^{\star,H} \int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) \Gamma^{\star,H} + \frac{\gamma^2 H}{N} \mathcal{C}(\theta^\star) + O(\gamma^{5/2} H) \ . \tag{103}
$$

We now use the fact that $\Gamma^{\star,H} = \mathrm{Id} - \gamma H \nabla^2 f_c(\theta^\star) + O(\gamma^2 H^2)$, which allows to rewrite

$$
\begin{aligned}
\int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) &= \left(\mathrm{Id} - \gamma H \nabla^2 f_c(\theta^\star) \right) \int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) \left(\mathrm{Id} - \gamma H \nabla^2 f_c(\theta^\star) \right) && (104) \\
&\quad + \frac{\gamma^2 H}{N} \mathcal{C}(\theta^\star) + O(\gamma^{5/2} H) + O(\gamma^3 H^2) \ . && (105)
\end{aligned}
$$

Simplifying this expression, we obtain

$$
\int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) = \frac{\gamma}{N} \mathbf{A} \mathcal{C}(\theta^\star) + O(\gamma^{3/2}) + O(\gamma^2 H) \ , \tag{106}
$$

where we recall that

$$
\mathbf{A} = \left(\mathrm{Id} \otimes \nabla^2 f(\theta^\star) + \nabla^2 f(\theta^\star) \otimes \mathrm{Id} \right)^{-1} \ , \tag{107}
$$

Plugging this expression in (102), we obtain the result

$$
\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} - \theta^\star = \frac{\gamma}{2N} \nabla^2 f(\theta^\star)^{-1} \nabla^3 f(\theta^\star) \mathbf{A} \mathcal{C}(\theta^\star) + O(\gamma^2 H) + O(\gamma^{3/2}) \ . \tag{108}
$$

B.5 General Functions, with Heterogeneous Agents

When functions are not quadratic nor homogeneous, local iterates are inherently biased. There are thus two sources of bias: heterogeneity, as in the quadratic case, and "iterate bias", that is due to stochasticity of gradients and the fact that derivatives of order greater than two are non zero.

To study this case, we define the following matrices, for $h = 0$ to H, that will be central in the analysis

$$\Gamma_c^{\star,h} = \left(\mathrm{Id} - \gamma \nabla^2 f_c(\theta^\star)\right)^h \quad , \tag{109}$$

as well as the aggregated variant of these matrices

$$\Gamma^{\star,h} = \frac{1}{N} \sum_{c=1}^N \Gamma_c^{\star,h} = \frac{1}{N} \sum_{c=1}^N \left(\mathrm{Id} - \gamma \nabla^2 f_c(\theta^\star)\right)^h \quad . \tag{110}$$

Note that, contrarily to the homogeneous setting, the $\Gamma_c^{\star,h}$'s differ from an agent to another. This will result in additional bias due to heterogeneity.

Expansion of Local Updates (Heterogeneous Case). We start by studying the local iterates of the algorithm. Using a second-order Taylor expansion of the gradient of ∇f_c at $\theta^\star$, we have

$$\begin{aligned}\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) = \nabla f_c(\theta^\star) + \nabla^2 f_c(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \theta^\star) \\ + \frac{1}{2}\nabla^3 f_c(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \theta^\star)^{\otimes 2} + \mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \quad ,\end{aligned}$$

where $\mathcal{R}_3^c$ is a function that satisfies $\sup_{\theta \in \mathbb{R}^d} \left\{ \frac{\|\mathcal{R}_{3,h}^c(\theta)\|}{\|\theta - \theta^\star\|^3} \right\} < +\infty$. We can use this expression to expand FedAvg's recursion as

$$\begin{aligned}\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star &= \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star - \gamma \nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \gamma \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \\ &= \widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star - \gamma \Big(\nabla f_c(\theta^\star) + \nabla^2 f_c(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \theta^\star) \\ &\qquad + \frac{1}{2}\nabla^3 f_c(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \theta^\star)^{\otimes 2} + \mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\Big) - \gamma \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) \\ &= \left(\mathrm{Id} - \gamma \nabla^2 f_c(\theta^\star)\right)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \theta^\star) - \gamma \nabla f_c(\theta^\star) \\ &\qquad - \frac{\gamma}{2}\nabla^3 f_c(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \theta^\star)^{\otimes 2} - \gamma \mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \gamma \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \quad .\end{aligned}$$

Unrolling this recursion, we obtain

$$\begin{aligned}\widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_c^{1:H}) - \theta^\star = \Gamma_c^{\star,H}(\theta - \theta^\star) - \gamma \sum_{h=0}^{H-1} \Gamma_c^{\star,H-h-1} \Big(\nabla f_c(\theta^\star) + \frac{1}{2}\nabla^3 f_c(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \theta^\star)^{\otimes 2} \\ + \mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})\Big) \quad .\end{aligned} \tag{111}$$

Expansion of Global Updates (Heterogeneous Case). We start by summing (111) over all agents

$$\begin{aligned}\frac{1}{N}\sum_{c=1}^N \theta_H^c - \theta^\star = \Gamma^{\star,H}(\theta - \theta^\star) - \frac{\gamma}{N}\sum_{c=1}^N \sum_{h=0}^{H-1} \Gamma_c^{\star,H-h-1} \Big(\nabla f_c(\theta^\star) + \frac{1}{2}\nabla^3 f_c(\theta^\star)\left(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\right)^{\otimes 2} \\ + \mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\Big) \quad .\end{aligned}$$

Similarly to the homogeneous setting, we integrate over $\pi^{(\gamma,H)}$, take the expectation and use the fact that $\frac{1}{N}\sum_{c=1}^N \theta_H^c$ follows the same distribution as θ, to obtain

$$\begin{aligned}(\mathrm{Id} - \Gamma^{\star,H})(\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} - \theta^\star) = -\frac{\gamma}{N}\sum_{c=1}^N \sum_{h=0}^{H-1} \Gamma_c^{\star,H-h-1} \nabla f_c(\theta^\star) \\ - \frac{\gamma}{2N}\sum_{c=1}^N \sum_{h=0}^{H-1} \Gamma_c^{\star,H-h-1} \nabla^3 f_c(\theta^\star) \int \left\{ \mathbb{E}\left(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) - \theta^\star\right)^{\otimes 2} + \mathbb{E}\mathcal{R}_{3,h}^c(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}))\right\} \pi^{(\gamma,H)}(\mathrm{d}\theta) \quad .\end{aligned} \tag{112}$$

Now we use Lemma 9 to write $-\gamma\sum_{h=0}^{H-1}\Gamma_c^{\star,H-h-1} = \left(\mathrm{Id}-\Gamma_c^{\star,H}\right)\nabla^2 f_c(\theta^\star)^{-1}$, and plug it in (112) to obtain

$$
\begin{aligned}
\left(\mathrm{Id}-\Gamma^{\star,H}\right)\left(\bar\theta^{(\gamma,H)}_{\mathrm{sto}}-\theta^\star\right) &= \frac{1}{N}\sum_{c=1}^{N}(\mathrm{Id}-\Gamma_c^{\star,H})\nabla^2 f_c(\theta^\star)^{-1}\nabla f_c(\theta^\star) \qquad (113)\\
&- \frac{\gamma}{2N}\sum_{c=1}^{N}\sum_{h=0}^{H-1}\Gamma_c^{\star,H-h-1}\nabla^3 f_c(\theta^\star)\int\!\left(\mathbb{E}\left(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h})-\theta^\star\right)^{\otimes 2}+\mathbb{E}\mathcal{R}^c_{3,h}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h}))\right)\pi^{(\gamma,H)}(\mathrm{d}\theta)\ .
\end{aligned}
$$

Interestingly, Equation (113) is composed of two terms. The first term is due to heterogeneity, and is the same as in the quadratic setting. From Proposition 6, we thus know that this term is of order $O(\gamma H)$. The second one reflects the bias of FedAvg that is due to stochasticity of the gradients.

Expansion of $\int\left(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h})-\theta^\star\right)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta)$ (Heterogeneous Case). We start with the following explicit expression of one round of the local updates

$$
\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h})-\theta^\star = \theta-\theta^\star-\gamma\sum_{\ell=0}^{h-1}\nabla f_c(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell}))+\varepsilon_c^{Z_c^{\ell+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell}))\ . \qquad (114)
$$

We use the first-order Taylor expansion of the gradient at $\theta^\star$ to obtain

$$
\begin{aligned}
&\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h})-\theta^\star \qquad (115)\\
&\quad=\theta-\theta^\star-\gamma\sum_{\ell=0}^{h-1}\nabla f_c(\theta^\star)+\nabla^2 f_c(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell})-\theta^\star)+\mathcal{R}^c_{2,\ell}(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell}))+\varepsilon_c^{Z_c^{\ell+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell}))\ ,
\end{aligned}
$$

where $\mathcal{R}^c_{2,\ell}:\mathbb{R}^d\to\mathbb{R}^d$ is a function such that $\sup_{\vartheta\in\mathbb{R}^d}\|\mathcal{R}^c_{2,\ell}(()\vartheta)\|/\|\vartheta-\theta^\star\|^2<+\infty$. Expanding the square of this equation, integrating over $\pi^{(\gamma,H)}$ and taking the expectation, we get

$$
\begin{aligned}
&\int\mathbb{E}\left(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h})-\theta^\star\right)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta)=\int(\theta-\theta^\star)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta)\\
&-\gamma\int(\theta-\theta^\star)\otimes\left(\sum_{\ell=0}^{h-1}\nabla f_c(\theta^\star)+\nabla^2 f_c(\theta^\star)(\mathbb{E}\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell})-\theta^\star)+\mathbb{E}\mathcal{R}^c_{2,\ell}(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell}))\right)\pi^{(\gamma,H)}(\mathrm{d}\theta)\\
&-\gamma\int\left(\sum_{\ell=0}^{h-1}\nabla f_c(\theta^\star)+\nabla^2 f_c(\theta^\star)(\mathbb{E}\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell})-\theta^\star)+\mathbb{E}\mathcal{R}^c_{2,\ell}(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell}))\right)\otimes(\theta-\theta^\star)\pi^{(\gamma,H)}(\mathrm{d}\theta)\\
&+\gamma^2\!\int\!\mathbb{E}\left(\sum_{\ell=0}^{h-1}\nabla f_c(\theta^\star)+\nabla^2 f_c(\theta^\star)(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell})-\theta^\star)+\mathcal{R}^c_{2,\ell}(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell}))+\varepsilon_c^{Z_c^{\ell+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,\ell)}(\theta;Z_c^{1:\ell}))\right)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta)\,.
\end{aligned}
$$

From this expansion, Hölder inequality, the definition of $\mathcal{R}^c_{2,\ell}$, A 3 and Lemma 8, we obtain

$$
\int\mathbb{E}\left(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h})-\theta^\star\right)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta)=\int(\theta-\theta^\star)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta)+O(\gamma^{3/2}H+\gamma^2H^2)\ . \qquad (116)
$$

Expression of the Global Update (Heterogeneous Case). Plugging (116) in (113), using Lemma 8 to bound $\int\mathcal{R}^c_{3,h}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h}))\pi^{(\gamma,H)}(\mathrm{d}\theta)=O(\gamma^{3/2}h^{3/2})$, and expanding the first term of (113) as in the quadratic setting (see Corollary 4), we now obtain

$$
\begin{aligned}
\bar\theta^{(\gamma,H)}_{\mathrm{sto}}-\theta^\star &= -\frac{\gamma(H-1)}{2N}\nabla^2 f(\theta^\star)^{-1}\sum_{c=1}^{N}(\nabla^2 f_c(\theta^\star)-\nabla^2 f(\theta^\star))\nabla f_c(\theta^\star)+O(\gamma^2H^2) \qquad (117)\\
&\quad-\frac{\gamma}{2N}(\mathrm{Id}-\Gamma^{\star,H})^{-1}\sum_{c=1}^{N}\sum_{h=0}^{H-1}\Gamma_c^{\star,H-h-1}\nabla^3 f_c(\theta^\star)\int(\theta-\theta^\star)^{\otimes 2}\pi^{(\gamma,H)}(\mathrm{d}\theta)+O(\gamma^{3/2}H+\gamma^2H^2)\ .
\end{aligned}
$$

Use Lemma 9, that is, $-\gamma \sum_{h=0}^{H-1} \Gamma_c^{\star, H-h-1} = \left(\mathrm{Id} - \Gamma_c^{\star, H}\right) \nabla^2 f_c(\theta^\star)^{-1}$, again, we obtain

$$\begin{aligned}\bar{\theta}_{\mathrm{sto}}^{(\gamma,H)} - \theta^\star = &-\frac{\gamma(H-1)}{2N} \nabla^2 f(\theta^\star)^{-1} \sum_{c=1}^{N} (\nabla^2 f_c(\theta^\star) - \nabla^2 f(\theta^\star)) \nabla f_c(\theta^\star) \\ &- \frac{1}{2N} \nabla^2 f(\theta^\star)^{-1} \nabla^3 f(\theta^\star) \int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) + O(\gamma^{3/2} H + \gamma^2 H^2) \ .\end{aligned} \tag{118}$$

Bound the Variance (Heterogeneous Case). To bound $\int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta)$, we proceed as above but with one less term in the expansion, and study the square. We get

$$\begin{aligned}&\widetilde{\mathsf{T}}_c^{(\gamma,h+1)}(\theta; Z_c^{1:h+1}) - \theta^\star \\ &= \left(\mathrm{Id} - \gamma \nabla^2 f_c(\theta^\star)\right) (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \theta^\star) - \gamma \nabla f_c(\theta^\star) - \gamma \mathcal{R}_{2,h}^c (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) - \gamma \varepsilon_c^{Z_c^{h+1}} (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h}) \ .\end{aligned}$$

Unrolling this recursion and averaging over all agents, we get

$$\begin{aligned}\widetilde{\mathsf{T}}_c^{(\gamma,H)}(\theta; Z_{1:N}^{1:H}) - \theta^\star = &\Gamma^{\star,H}(\theta - \theta^\star) \\ &- \frac{\gamma}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \Gamma_c^{\star, H-h-1} \left\{ \nabla f_c(\theta^\star) + \mathcal{R}_{2,h}^c (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}} (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right\} \ .\end{aligned}$$

Taking the second order moment of this equation, using the fact that $\frac{1}{N} \sum_{c=1}^{N} \theta_H^c$ follows the same distribution as θ, and integrating over $\pi^{(\gamma,H)}$, we obtain

$$\begin{aligned}&\int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) \\ &= \int \left(\Gamma^{\star,H}(\theta - \theta^\star) - \frac{\gamma}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \Gamma_c^{\star, H-h-1} \left\{ \nabla f_c(\theta^\star) + \mathcal{R}_{2,h}^c (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}} (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right\} \right)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\theta) \\ &= \Gamma^{\star,H} \int (\theta - \theta^\star)^{\otimes 2} \, \pi^{(\gamma,H)}(\mathrm{d}\theta) \Gamma^{\star,H} \\ &- \gamma \int \left(\Gamma^{\star,H}(\theta - \theta^\star) \right) \otimes \left(\frac{1}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \Gamma_c^{\star, H-h-1} \left\{ \nabla f_c(\theta^\star) + \mathcal{R}_{2,h}^c (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}} (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right\} \right) \pi^{(\gamma,H)}(\mathrm{d}\theta) \\ &- \gamma \int \left(\frac{1}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \Gamma_c^{\star, H-h-1} \left\{ \nabla f_c(\theta^\star) + \mathcal{R}_{2,h}^c (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}} (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right\} \right) \otimes \left(\Gamma^{\star,H}(\theta - \theta^\star) \right) \pi^{(\gamma,H)}(\mathrm{d}\theta) \\ &+ \gamma^2 \int \left(\frac{1}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \Gamma_c^{\star, H-h-1} \left\{ \nabla f_c(\theta^\star) + \mathcal{R}_{2,h}^c (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) + \varepsilon_c^{Z_c^{h+1}} (\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta; Z_c^{1:h})) \right\} \right)^{\otimes 2} \pi^{(\gamma,H)}(\mathrm{d}\theta) \ .\end{aligned}$$

Now, we expand $\Gamma_c^{\star, H-h-1}$ and use the fact that $\frac{1}{N} \sum_{c=1}^{N} \nabla f_c(\theta^\star) = 0$, which gives

$$\begin{aligned}\frac{1}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \Gamma_c^{\star, H-h-1} \nabla f_c(\theta^\star) &= \frac{1}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \nabla f_c(\theta^\star) - \gamma H \nabla^2 f_c(\theta^\star) \nabla f_c(\theta^\star) + O(\gamma^2 H^2) \\ &= \frac{1}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} -\gamma H \nabla^2 f_c(\theta^\star) \nabla f_c(\theta^\star) + O(\gamma^2 H^2) \ ,\end{aligned}$$

which, combined with $\gamma H \leq 1$, implies that

$$\frac{1}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \Gamma_c^{\star, H-h-1} \nabla f_c(\theta^\star) = O(\gamma H^2) \ , \quad \text{and} \quad \left(\frac{1}{N} \sum_{c=1}^{N} \sum_{h=0}^{H-1} \Gamma_c^{\star, H-h-1} \nabla f_c(\theta^\star) \right)^{\otimes 2} = O(\gamma^2 H^4) \ .$$

Combining the expansions above with Hölder inequality, the definition of $\mathcal{R}^c_{2,\ell}$, A 3 and Lemma 8, we obtain

$$
\begin{aligned}
\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta) &= \Gamma^{\star,H}\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta)\Gamma^{\star,H} \\
&\quad + \frac{\gamma^2}{N}\sum_{h=0}^{H-1}\int \mathbb{E}\left[\frac{1}{N}\sum_{c=1}^{N}\varepsilon_c^{Z_c^{h+1}}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h}))^{\otimes 2}\right]\pi^{(\gamma,H)}(\mathrm{d}\theta) + O(\gamma^3H^3) + O(\gamma^{5/2}H^2) \\
&= \Gamma^{\star,H}\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta)\Gamma^{\star,H} + \frac{\gamma^2}{N}\sum_{h=0}^{H-1}\int \mathcal{C}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h}))\pi^{(\gamma,H)}(\mathrm{d}\theta) + O(\gamma^3H^3) + O(\gamma^{5/2}H^2)\ .
\end{aligned}
$$

Now, using A 3 and Lemma 8 we have $\int \mathcal{C}(\widetilde{\mathsf{T}}_c^{(\gamma,h)}(\theta;Z_c^{1:h}))\pi^{(\gamma,H)}(\mathrm{d}\theta) = \mathcal{C}(\theta^\star) + O(\gamma H)$, which results in the identity

$$
\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta) = \Gamma^{\star,H}\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta)\Gamma^{\star,H} + \frac{\gamma^2H}{N}\mathcal{C}(\theta^\star) + O(\gamma^3H^3) + O(\gamma^{5/2}H^2)\ .
$$

We now use the fact that $\Gamma^{\star,H} = \mathrm{Id} - \gamma H\nabla^2 f(\theta^\star) + O(\gamma^2H^2)$, which allows to rewrite

$$
\begin{aligned}
\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta) &= \big(\mathrm{Id} - \gamma H\nabla^2 f(\theta^\star)\big)\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta)\,\big(\mathrm{Id} - \gamma H\nabla^2 f(\theta^\star)\big) && (119)\\
&\quad + \frac{\gamma^2H}{N}\mathcal{C}(\theta^\star) + O(\gamma^3H^3) + O(\gamma^{5/2}H^2)\ . && (120)
\end{aligned}
$$

Developing this expression and using Lemma 8, we get

$$
\begin{aligned}
\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta) &= \int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta) \\
&\quad - \gamma H\nabla^2 f(\theta^\star)\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta) - \gamma H\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta)\nabla^2 f(\theta^\star) \\
&\quad + \frac{\gamma^2H}{N}\mathcal{C}(\theta^\star) + O(\gamma^3H^3) + O(\gamma^{5/2}H^2)\ .
\end{aligned}
$$

Simplifying this expression, we obtain

$$
\int (\theta-\theta^\star)^{\otimes 2}\,\pi^{(\gamma,H)}(\mathrm{d}\theta) = \frac{\gamma}{N}\mathbf{A}\mathcal{C}(\theta^\star) + O(\gamma^2H^2) + O(\gamma^{3/2}H)\ , \tag{121}
$$

where we recall that

$$
\mathbf{A} = \big(\mathrm{Id}\otimes\nabla^2 f(\theta^\star) + \nabla^2 f(\theta^\star)\otimes\mathrm{Id}\big)^{-1}\ , \tag{122}
$$

Plugging this expression in (118), we obtain the result

$$
\begin{aligned}
\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}} - \theta^\star &= -\frac{\gamma(H-1)}{2N}\nabla^2 f(\theta^\star)^{-1}\sum_{c=1}^{N}(\nabla^2 f_c(\theta^\star) - \nabla^2 f(\theta^\star))\nabla f_c(\theta^\star) \\
&\quad - \frac{\gamma}{N}\nabla^2 f(\theta^\star)^{-1}\nabla^3 f(\theta^\star)\mathbf{A}\mathcal{C}(\theta^\star) + O(\gamma^2H^2) + O(\gamma^{3/2}H)\ .
\end{aligned}
$$

B.6 Proof of Theorem 5

The only statement to show is that for $\gamma \le 1/L \wedge 1/H$, then the iterates $\{\bar{\theta}^{(\gamma,H)}_T\}_{T\ge 1}$ defined as

$$
\bar{\theta}^{(\gamma,H)}_T = \frac{1}{T}\sum_{t=0}^{T-1}\theta^{(\gamma,H)}_t\ ,
$$

converge in L^2 to $\bar{\theta}^{(\gamma,H)}_{\mathrm{sto}}$. This is an easy consequence of [Durmus et al., 2024, Theorem 8] whose assumptions are satisfied by Lemma 6 and Proposition 3.

C Technical Lemma on Matrix Products

Lemma 9. *For any matrix-valued sequences $(M_k)_{k\in\mathbb{N}}$, $(M'_k)_{k\in\mathbb{N}}$ and for any $K \in \mathbb{N}$, it holds that:*

$$\prod_{k=1}^{K} M_k - \prod_{k=1}^{K} M'_k = \sum_{k=1}^{K} \left\{ \prod_{\ell=1}^{k-1} M_\ell \right\} \left(M_k - M'_k \right) \left\{ \prod_{\ell=k+1}^{M} M'_\ell \right\} \ . \tag{123}$$