Tight Analysis of Decentralized SGD: A Markov Chain Perspective

Lucas Versini **Paul Mangold** **Aymeric Dieuleveut**
CMAP, CNRS, École polytechnique, Institut Polytechnique de Paris

Abstract

We propose a novel analysis of the Decentralized Stochastic Gradient Descent (DSGD) algorithm with constant step size, interpreting the iterates of the algorithm as a Markov chain. We show that DSGD converges to a stationary distribution, with its bias, to first order, decomposable into two components: one due to decentralization (growing with the graph's spectral gap and clients' heterogeneity) and one due to stochasticity. Remarkably, the variance of *local* parameters is, at the first-order, inversely proportional to the number of clients, regardless of the network topology and even when clients' iterates are not averaged at the end. As a consequence of our analysis, we obtain non-asymptotic convergence bounds for clients' local iterates, confirming that DSGD has linear speed-up in the number of clients, and that the network topology only impacts higher-order terms.

1 INTRODUCTION

Decentralized optimization is a key paradigm for large-scale machine learning. For such problems, it quickly becomes necessary to keep data distributed across multiple clients, sharing the computation among all participants. While many methods rely on a central server to coordinate communication [Kairouz et al., 2021], decentralized approaches offer a strong alternative for handling even larger and more heterogeneous clients. A fundamental method in this setting is Decentralized Stochastic Gradient Descent (DSGD), where clients perform local stochastic updates and exchange information with their neighbors [Nedic and Ozdaglar, 2009b, Duchi et al., 2011, Lian et al., 2017]. This algorithm alleviates the strong requirements of centralized and federated learning, which either necessitate collecting all data in a single location or relying on a central server to manage all communications.

Decentralized learning thus appears as a more flexible alternative to the widely studied federated learning [McMahan et al., 2017], and many variants of DSGD have been proposed an analyzed in a wide range of settings [Colla and Hendrickx, 2021, Koloskova et al., 2020, Yuan et al., 2016, Richards and Rebeschini, 2020, Bars et al., 2023, Sayed, 2014, Nedic and Ozdaglar, 2009a, Sundhar Ram et al., 2010]. Nonetheless, some of DSGD's theoretical properties still remain misunderstood. Specifically, this algorithm and its variants have been studied extensively in the deterministic setting, with corresponding upper-bounds on the error and the convergence rate [Vogels et al., 2022, Nedic and Ozdaglar, 2009a, Sundhar Ram et al., 2010, Yuan et al., 2016, Koloskova et al., 2020], but its behavior under stochastic noise is less precisely understood.

In this work, we present a refined analysis of DSGD, with a particular focus on the stochastic nature of local updates. We propose a novel interpretation of DSGD as a Markov chain, which we show to be geometrically ergodic and therefore convergent to a stationary distribution. By analyzing this stationary distribution, we derive exact first-order expansions in the step size for the bias and variance of DSGD. Two key insights stem from these expressions: a bias arises when the communication graph is not fully connected and clients are heterogeneous, while the variance depends on the graph only through higher-order terms. Finally, we provide a non-asymptotic convergence rate, enabling a precise characterization of DSGD's convergence.

We stress that the goal of our work is to *precisely* study the properties of DSGD, one of the most fundamental algorithm of decentralized learning. Here, we aim to develop an analytical framework for the analysis of DSGD, that (i) gives very precise insights on the bias of the algorithm through first-order expansions, rather than upper bounds, (ii) could readily be extended to more complex decentralized algorithms, and (iii) could guide the design of future decentralized algorithms. Our contributions can be summarized as follows:

- In Section 3, we first study the *deterministic* counterpart of DSGD, simply denoted as deterministic

GD (DGD), and derive an explicit first-order expansion in the step size of its bias. This shows that the limit point reached by the algorithm moves further away from the optimal point with both network's spectral gap and clients' heterogeneity. In particular, it vanishes when either of the two is zero.

- In Section 4, interpreting DSGD as a Markov chain, we show that DSGD's iterates converge geometrically fast (in Wasserstein distance) towards a stationary distribution. We provide first-order expansions of DSGD's bias and variance at stationarity, disentangling two sources of bias: one due to decentralization and heterogeneity combined, and one due to stochasticity. Interestingly, DSGD's variance decreases with the number of clients, and depends on the network topology only in higher-order terms.
- We then derive sharper non-asymptotic rates for DSGD. Specifically, we (i) show that clients iterates converge with linear speed-up *without averaging the parameters across clients*, and (ii) tightly control the impact of decentralization and heterogeneity, as well as that of the topology on the variance.
- In Section 5, we propose a novel decentralized algorithm, based on Richardson-Romberg extrapolation, eliminating first order bias. We study its sample complexity, proving that it reduces communication without any knowledge of the network topology.
- Finally, we provide experiments in Section 6.

Notations. For $f : \mathbb{R}^d \to \mathbb{R}$ a function i-times differentiable ($i \geq 1$), we denote $\nabla^i f$ its i-th tensor derivative. For a vector $\|\cdot\|$ is its Euclidean norm, and for a matrix, it is its ℓ_2 operator norm. Id is the identity matrix, and $\mathbf{1}_m$ the vector filled with 1 in dimension m. For a symmetric matrix A, we denote A's second largest eigenvalue $\lambda_2(A)$, and $\lambda_{\min}(A)$ its smallest eigenvalue. We write $\mathbf{W}_2(\rho_1, \rho_2)$ the second-order Wasserstein distance between two probability measures ρ_1 and ρ_2. For two matrices A, B, $A \otimes B$ is the Kronecker product of A and B, *i.e.*, the operator $A \otimes B : M \mapsto BMA^\top$. We also define $A^{\otimes k}$ as the k-th power of the tensor A.

2 PROBLEM SETTING

Decentralized Learning. We consider the distributed optimization problem

$$\theta^\star \in \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d} f(\theta) = \tfrac{1}{m} \textstyle\sum_{k=1}^m f_k(\theta) \ , \tag{1}$$

where for $k \in \{1, \ldots, m\}$, $f_k : \mathbb{R}^d \to \mathbb{R}$ is the local objective of agent k. Each agent starts from an initial vector $\theta_0^{(k)} \in \mathbb{R}^d$, and updates it with its neighbors

$$\theta_{t+1}^{(k)} = \sum_{\ell=1}^m W_{k\ell} \big(\theta_t^{(\ell)} - \gamma\big(\nabla f_\ell(\theta_t^{(\ell)}) + \varepsilon_{t+1}^{(\ell)}(\theta_t^{(\ell)})\big)\big) \ ,$$

where $W \in \mathbb{R}^{m \times m}$ is a fixed communication matrix, assumed to be symmetric and doubly stochastic, and $\varepsilon_{t+1}^{(\ell)}(\theta_t^{(\ell)})$ is a random noise term.

Note the clients first update their parameters, before averaging the parameters of their neighbors, which is usually referred to as "Adapt-then-Combine".

Algorithm 1 DSGD

Input: $\forall k$, $\theta_0^{(k)} = \theta_0 \in \mathbb{R}^d$, number of iterations T, step size γ, matrix W.
for $t = 0, \ldots, T-1$ **do**
  for each client $k = 1, \ldots, m$ **do**
    $\theta_{t+1}^{(k)} \leftarrow \sum_{\ell=1}^m W_{k\ell}\big(\theta_t^{(\ell)} - \gamma\big(\nabla f_\ell(\theta_t^{(\ell)}) + \varepsilon_{t+1}^{(\ell)}(\theta_t^{(\ell)})\big)\big)$
return $\theta_T^{(1)}, \ldots, \theta_T^{(m)}$.

It is convenient to rewrite the updates in global form. We define the stacked parameter vector Θ_t, the global gradient $\nabla F(\Theta_t)$ and the global noise $\varepsilon_{t+1}(\Theta_t)$ in $\mathbb{R}^{md}$ as

$$\begin{aligned}
\Theta_t &:= \big((\theta_t^{(1)})^\top \quad \cdots \quad (\theta_t^{(m)})^\top\big)^\top \ , \\
\nabla F(\Theta_t) &:= \big((\nabla f_1(\theta_t^{(1)}))^\top \quad \cdots \quad (\nabla f_m(\theta_t^{(m)}))^\top\big)^\top \ , \\
\varepsilon_{t+1}(\Theta_t) &:= \big((\varepsilon_{t+1}^{(1)}(\theta_t^{(1)}))^\top \quad \cdots \quad (\varepsilon_{t+1}^{(m)}(\theta_t^{(m)}))^\top\big)^\top \ .
\end{aligned}$$

With these notations, the algorithm recursion writes

$$\Theta_{t+1} = \boldsymbol{W}\left(\Theta_t - \gamma\big(\nabla F(\Theta_t) + \varepsilon_{t+1}(\Theta_t)\big)\right) \ , \tag{DSGD}$$

where $\boldsymbol{W} := W \otimes \mathrm{Id}$. Lastly, we also set $\Theta^\star = \mathbf{1}_m \otimes \theta^\star \in \mathbb{R}^{md}$, defined by repeating m times $\theta^\star$.

Assumptions. Throughout this paper, we work under the following regularity assumptions.

A 1 (Regularity)**.** *For all $k \in \{1, \ldots, m\}$:*

(a) f_k is μ-strongly convex.

(b) f_k is four times differentiable and L-smooth.

(c) f_k's third derivative is bounded: there exists $K_3 > 0$ such that for $\theta, u \in \mathbb{R}^d$, $\|\nabla^3 f_k(\theta) u^{\otimes 2}\| \leq K_3 \|u\|^2$.

This assumption is standard in the analysis of stochastic and decentralized algorithms [Dieuleveut et al., 2020, Mangold et al., 2025b,a, Vogels et al., 2022]. The third derivative's boundedness is useful to control higher-order terms, as can be seen in the proofs of the results presented in Section 4.

Finally, to measure heterogeneity, we also define

$$\zeta_\star^2 = \|\nabla F(\Theta^\star)\|^2 \ . \tag{2}$$

A 2. *The matrix W is symmetric and stochastic, with $\lambda_2(W) < 1$; we let $\Lambda = 2\|(\mathrm{Id} - W)^\dagger W\|_2$. We define $\rho = \max(|\lambda_2(W)|, |\lambda_{\min}(W)|)$.*

This assumption is classical in decentralized learning, and ensures that the underlying undirected communication graph is connected. In particular, information from each client eventually propagates to all other clients. For instance, given a weighted, connected, and undirected graph with Laplacian matrix L, the matrix $W = \mathrm{Id} - tL$ satisfies A 2 for a small enough $t > 0$.

3 DETERMINISTIC ALGORITHM

In this section, we start with the analysis of deterministic Decentralized GD (DGD), ie.

$$\Theta_{t+1} = \boldsymbol{W}\big(\Theta_t - \gamma \nabla F(\Theta_t)\big) \ . \tag{DGD}$$

One of the first notable analysis of DGD was provided by Nedic and Ozdaglar [2009a], and many rates were later established on the sub-optimality gap $f(\bar{\theta}_T) - \min f$, where $\bar{\theta}_T$ is the average of clients iterates at time T [Colla and Hendrickx, 2021, Yuan et al., 2016, Le Bars et al., 2023]. In particular, it is known that the clients' models do not all converge to the optimal point, motivating the introduction of *gradient tracking* [Shi et al., 2015, Nedic et al., 2017]. Most works focus on upper bounding the error between $\bar{\theta}_T$ and θ_*. Yet, computing the parameter $\bar{\theta}_T$ requires a global averaging step over the entire graph, which may be costly.

Furthermore, our goal is to understand the limit behavior of the limit as function of γ. We thus need to establish the convergence to a limit and characterize it precisely. We show that DGD converges, and derive an explicit expansion of its bias (*i.e.*, the difference between its limit point and the problem's global solution). We first recall the convergence of DGD to a fixed point. While this first result was established by [Vogels et al., 2022, Larsson and Michelusi, 2025], our following results are more precise, with an *exact first-order expansion of the limit point*, rather than upper bounds, allowing to obtain extrapolation results.

Lemma 3.1. *Assume A 1, A 2, and $\gamma \le 1/L$. Then the sequence $(\Theta_t)_t$ generated by (DGD) converges to a vector $\Theta_{\det}$, independent of Θ_0, which satisfies*

$$(\boldsymbol{I} - \boldsymbol{W})\Theta_{\det} = -\gamma \boldsymbol{W} \nabla F(\Theta_{\det}) \ . \tag{3}$$

Moreover, for all $t \ge 0$,

$$\|\Theta_t - \Theta_{\det}\| \le (1 - \gamma\mu)^t \|\Theta_0 - \Theta_{\det}\| \ . \tag{4}$$

We give a proof based on a contraction argument of this lemma in Appendix A. This establishes that the sequence $(\Theta_t)_t$ converges to a point $\Theta_{\det}$, which depends on both the step size γ and the communication matrix W. We next use (3) to derive an explicit expansion of $\Theta_{\det}$. Note that the matrix $\mathrm{Id} - W$ has different behavior on $\mathrm{Span}(\mathbf{1})$, where it is zero, and on its orthogonal space $\mathrm{Span}(\mathbf{1})^\perp$, where its largest eigenvalue is strictly smaller than 1. We thus decompose the vectors into two orthogonal components: a consensus part, where all the clients have the same value, and a disagreement part, characterizing individual specificity. We define the consensus and disagreement projection operators, which project a vector into the relevant subspace

$$\boldsymbol{\mathcal{P}} := \tfrac{1}{m}\mathbf{1}\mathbf{1}^\top \otimes \mathrm{Id} \ , \qquad \boldsymbol{\mathcal{Q}} := \boldsymbol{I} - \boldsymbol{\mathcal{P}} \ . \tag{5}$$

The operator $\boldsymbol{\mathcal{P}}$ projects a vector on the consensus space $\mathrm{Span}(\mathbf{1})$ by averaging its block-components, while $\boldsymbol{\mathcal{Q}}$ computes the difference between a vector and its consensus. We can now define the consensus and disagreement parts of the clients' iterates

$$\overline{\Theta}_{\det} := \boldsymbol{\mathcal{P}}\Theta_{\det} \ , \qquad \widetilde{\Theta}_{\det} := \boldsymbol{\mathcal{Q}}\Theta_{\det} \ , \tag{6}$$

so that $\Theta_{\det} = \overline{\Theta}_{\det} + \widetilde{\Theta}_{\det}$. With these notations, the identity (3) can be reformulated as follows.

Lemma 3.2. *Under A 1, A 2, and for $0 < \gamma \le 1/L$:*

$$\tfrac{1}{m}\textstyle\sum_{k=1}^m \nabla f_k(\theta_k^{\det}) = 0 \ , \tag{7}$$

$$\widetilde{\Theta}_{\det} = -\gamma(\boldsymbol{I} - \boldsymbol{W})^\dagger \boldsymbol{W} \nabla F(\Theta_{\det}) \ . \tag{8}$$

The identity (7) is a kind of optimality condition; if $\Theta_{\det}$ is already at consensus, it means that DGD converged to the solution of the global problem. The second equation (8) defines the disagreement part. Based on the identity $(1-x)^{-1} = \sum_{k=0}^\infty x^k$, (8) can be interpreted as the part of the gradients that reaches each client after any $k \ge 0$ rounds of propagation through the network, making iterates depart from consensus.

Now, we derive expansions of the error based on Lemma 3.2: first for quadratics, where we give *exact* expressions, and then for general functions.

Quadratic functions. When the functions f_k are quadratic, we can derive an exact expression of $\Theta_{\det}$.

A 3. *For $k \in \{1, \ldots, m\}$, there exist $A_k \in \mathbb{R}^{d\times d}$ positive definite and $\theta_k^\star \in \mathbb{R}^d$ such that $f_k(\theta) = \frac{1}{2}\|A_k^{1/2}(\theta - \theta_k^\star)\|^2$. We let*

$$\bar{A} := \tfrac{1}{m}\textstyle\sum_{k=1}^m A_k \ , \qquad \boldsymbol{A} := \mathrm{diag}(A_1, \ldots, A_m) \ ,$$
$$\Theta^\star_{(loc)} := (\theta_1^{\star\top}, \ldots, \theta_m^{\star\top})^\top \ , \qquad \bar{\boldsymbol{A}} := \mathrm{diag}(\bar{A}, \ldots, \bar{A}) \ .$$

In this setting, the fixed-point equation (3) reduces to a linear system $(\mathrm{Id} - W)\Theta_{\det} = -\gamma W \boldsymbol{A}(\Theta_{\det} - \Theta^\star_{(\mathrm{loc})})$.

Under A 3, we define the matrices

$$\boldsymbol{\mathcal{H}} := \boldsymbol{\mathcal{P}}\bar{\boldsymbol{A}}^{-1}\boldsymbol{A} \ , \qquad \boldsymbol{\mathcal{G}} := (\boldsymbol{I}-\boldsymbol{W})^{\dagger}\boldsymbol{W} \ , \tag{9}$$

$$\boldsymbol{\mathcal{B}} := (\boldsymbol{I}+\gamma\boldsymbol{\mathcal{G}}\boldsymbol{A})^{-1}\boldsymbol{\mathcal{G}}\boldsymbol{A} \tag{10}$$

The matrix $\boldsymbol{\mathcal{H}}$ measures heterogeneity among clients, and relates local and global solutions. Indeed, we have $\theta^\star = \boldsymbol{\mathcal{H}}(\theta_1^{\star\top}, \ldots, \theta_m^{\star\top})^\top$. If all matrices A_k are equal, then $\theta^\star$ is simply the average of the vectors θ_k. The matrix $\boldsymbol{\mathcal{G}}$, which is zero on the consensus subspace $\mathrm{Span}(\mathbf{1})$, acts as $\sum_{t=1}^{+\infty}\boldsymbol{W}^t$ on the disagreement subspace $\mathrm{Span}(\mathbf{1})^\perp$. When applied to a vector, $\boldsymbol{\mathcal{G}}$ sums all the disagreements propagated in the graph from this vector. The matrix $\boldsymbol{\mathcal{B}}$ mixes these two matrices, accounting both for heterogeneity and network topology. Using these matrices, we can give an exact expression, and first-order expansions, of the error.

Proposition 3.3. *Assume A 2-3 and $\gamma < 2/((1+L/\mu)L\Lambda)$. Then we have the following expressions*

$$\begin{aligned}
\overline{\Theta}_{\det} &= \Theta^\star - \gamma\boldsymbol{\mathcal{H}}(\boldsymbol{I}-\gamma\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{H}})^{-1}\boldsymbol{\mathcal{B}}(\Theta^\star_{(loc)}-\Theta^\star) \ , \\
\widetilde{\Theta}_{\det} &= \gamma(\boldsymbol{I}-\gamma\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{H}})^{-1}\boldsymbol{\mathcal{B}}(\Theta^\star_{(loc)}-\Theta^\star) \ , \\
\Theta_{\det} &= \Theta^\star + \gamma(\boldsymbol{I}-\boldsymbol{\mathcal{H}})(\boldsymbol{I}-\gamma\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{H}})^{-1}\boldsymbol{\mathcal{B}}(\Theta^\star_{(loc)}-\Theta^\star) \ ,
\end{aligned}$$

Moreover, the following expansions hold:

$$\begin{aligned}
\overline{\Theta}_{\det} &= \Theta^\star - \gamma\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{G}}\boldsymbol{A}(\Theta^\star_{(loc)}-\Theta^\star) + \mathcal{O}(\gamma^2) \ , \\
\widetilde{\Theta}_{\det} &= \gamma\boldsymbol{\mathcal{G}}\boldsymbol{A}(\Theta^\star_{(loc)}-\Theta^\star) + \mathcal{O}(\gamma^2) \ , \\
\Theta_{\det} &= \Theta^\star + \gamma(\boldsymbol{I}-\boldsymbol{\mathcal{H}})\boldsymbol{\mathcal{G}}\boldsymbol{A}(\Theta^\star_{(loc)}-\Theta^\star) + \mathcal{O}(\gamma^2) \ .
\end{aligned}$$

We prove this lemma in Appendix A. Remark that the condition on γ is only needed to guarantee the existence of all inverse matrices that appear in the expressions. For larger values of γ for which all inverse matrices exist, the above formulas still hold. The expressions from Proposition 3.3 allow to make multiple observations on the bias of DGD in this setting.

Impact of the step size. The expansion shows that $\Theta_{\det} - \Theta^\star = \mathcal{O}(\gamma)$, and thus $\lim_{\gamma\to 0}\Theta_{\det} = \Theta^\star$. Hence, decreasing the step size reduces the bias of DGD, which eventually vanishes. Note that the bias, as well a the disagreement, scales linearly with γ for both the consensus and disagreement components. Of course, as seen in Lemma 3.1, reducing the bias by decreasing γ also slows down the convergence of the algorithm.

Impact of data heterogeneity. In the homogeneous case where all local minimizers $\theta_k^\star$ coincide, we have $\Theta^\star_{(\mathrm{loc})} = \Theta^\star$, and therefore $\Theta_{\det} = \Theta^\star$. This shows that heterogeneity across clients is the only source of bias: the greater the differences between the vectors $\theta_k^\star$, the larger the deviation of $\Theta_{\det}$ from $\Theta^\star$.

Role of the communication matrix. The network topology appears through $\boldsymbol{\mathcal{G}} = (\boldsymbol{I}-\boldsymbol{W})^\dagger\boldsymbol{W}$. Recall that $\boldsymbol{\mathcal{G}}$ captures how disagreement between clients is smoothed out by iterated communication by summing all the disagreements obtained on the gradient $\nabla F(\Theta^\star)$ at each time step $t \geq 1$. As an example, when the network is fully connected ($W = \frac{1}{m}\mathbf{1}\mathbf{1}^\top$), $\boldsymbol{\mathcal{G}}$ is equal to 0, and $\Theta_{\det} = \Theta^\star$; for sparse graphs, $\boldsymbol{\mathcal{G}}$ is nonzero, leading to a larger bias. Generally, under A 2, the spectral norm of $\boldsymbol{\mathcal{G}}$ is equal to $\Lambda = 2\|(\mathrm{Id}-W)^\dagger W\|_2$, which decreases when the eigenvalues of W, other than the one at 1, are close to 0, i.e., when the graph is more connected.

General functions. We now study the case of strongly convex and smooth functions. To derive our refined analysis, which finely quantifies the prominent terms in the error, we need a crude bound on DGD's convergence, which we will use to bound higher-order terms. First, we control the bias $\Theta_{\det} - \Theta^\star$, using a bound recently established by Larsson and Michelusi [2025], showing that it is of order γ.

Lemma 3.4 ([Larsson and Michelusi, 2025, Formula 47])**.** *Assume A 1 and A 2. Then for $\gamma < \min\left(\frac{1}{\Lambda L}, \frac{1}{L}\right)$,*

$$\|\Theta_{\det} - \Theta^\star\| \leq \tfrac{\gamma L\Lambda}{\mu}\zeta_\star \ .$$

In particular, as $\gamma \to 0$, $\|\Theta_{\det} - \Theta^\star\| = \mathcal{O}(\gamma)$.

This directly gives a convergence rate for DGD.

Corollary 3.5. *Assume A 1-2. Then, for any $t \geq 0$. we have the convergence rate*

$$\|\Theta_t - \Theta^\star\| \leq (1-\gamma\mu)^t\|\Theta_0 - \Theta^\star\| + \tfrac{2\gamma L\Lambda}{\mu}\zeta_\star \ .$$

For $\epsilon > 0$, setting $\gamma = \min(\frac{1}{L}, \frac{1}{\Lambda L}, \frac{\mu\epsilon}{L\Lambda\zeta_\star})$, we have $\|\Theta_t - \Theta^\star\| \leq \epsilon$ for $T \gtrsim \max(\frac{L}{\mu}, \frac{\Lambda L}{\mu}, \frac{L\Lambda\zeta_\star}{\mu^2\epsilon})\log(\frac{\|\Theta_0-\Theta^\star\|}{\epsilon})$.

Notice that to achieve good precision, one needs to require γ to be smaller when the graph is less connected. Using this bound, we adapt Proposition 3.3 to the general strongly convex and smooth setting. To this end, and under A 1-2, we define the average Hessian $\nabla^2 f(\theta^\star) := \frac{1}{m}\sum_{k=1}^m \nabla^2 f_k(\theta^\star)$, as well as the the block-diagonal matrices

$$\begin{aligned}
\boldsymbol{A} &:= \mathrm{diag}\left(\nabla^2 f_1(\theta^\star), \ldots, \nabla^2 f_m(\theta^\star)\right) \ , \\
\bar{\boldsymbol{A}} &:= \mathrm{diag}(\nabla^2 f(\theta^\star), \ldots, \nabla^2 f(\theta^\star)) \ ,
\end{aligned}$$

and formally define $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mathcal{G}}$ as in (9)-(10) with theses matrices $\boldsymbol{A}$ and $\bar{\boldsymbol{A}}$.

Proposition 3.6 (First-order bias expansion)**.** *Assume A 1-2. Then we have*

$$\begin{aligned}
\overline{\Theta}_{\det} &= \Theta^\star + \gamma\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{G}}\nabla F(\Theta^\star) + \mathcal{O}(\gamma^2) \ , \\
\widetilde{\Theta}_{\det} &= -\gamma\boldsymbol{\mathcal{G}}\nabla F(\Theta^\star) + \mathcal{O}(\gamma^2) \ , \\
\Theta_{\det} &= \Theta^\star - \gamma(\boldsymbol{I}-\boldsymbol{\mathcal{H}})\boldsymbol{\mathcal{G}}\nabla F(\Theta^\star) + \mathcal{O}(\gamma^2) \ .
\end{aligned}$$

For $\gamma \leq \min\left(\frac{1}{\Lambda L}, \frac{1}{L}\right)$, the residual term is bounded by

$$\gamma^2 \frac{L^2}{2\mu^2}\Lambda^2\left(\frac{K_3}{\mu\sqrt{m}}\zeta_\star^2 + L\cdot\zeta_\star\right) .$$

The proof, given in Appendix A, linearizes the local gradients using Taylor expansion around $\theta^\star$, and uses Lemma 3.4 to control higher-order terms. Proposition 3.6 is the exact analogue of Proposition 3.3 for non-quadratic functions. The bias, to the first order, is the same as in the quadratic setting, the same discussion thus applies. In short, the same three factors as in the quadratic setting (step size, heterogeneity and network connectivity) determine the bias of deterministic decentralized gradient descent. The main difference is that, contrarily to the quadratic setting, no closed-form expressions are available, but only first-order expansions.

4 STOCHASTIC ALGORITHM

We now consider DSGD with stochastic gradients. Prior work [Sirb and Ye, 2016, Lan et al., 2020, Lian et al., 2017, Bars et al., 2023, Neglia et al., 2020] provided convergence rates on $\mathbb{E}\|\nabla f(\Theta_t)\|^2$ or $\mathbb{E}\left[f(\Theta_t)\right] - \min f$, where Θ_t is the vector of all local models at iteration t. Other papers focus directly on the sequence of iterates $(\Theta_t)_{t\geq 0}$ [Vogels et al., 2022, Theorem 9] and prove that the distance between Θ_t and $\Theta^\star$ can be upper-bounded by a sum of three terms: one which depends on the initialization and vanishes as the time step t tends to $+\infty$, one which depends on the variance, and one which depends on heterogeneity.

Our analysis departs from these results by analyzing the iterates of DSGD under a constant step size *as a Markov chain.* In the single-client setting, Pflug [1986a], Moulines et al. [1998], Chee and Toulis [2018] showed that, with a constant step size, SGD does not converge to the minimizer, but rather oscillates around it. Viewing SGD with a constant step size as a Markov chain was used to characterizes SGD's stationary distribution, with convergence rates and asymptotic bias expansions [Dieuleveut et al., 2020]. This approach has recently been extended to federated learning algorithms such as Scaffold and FedAvg [Mangold et al., 2025a,b], giving expansions of the bias and variance, with associated convergence rates.

In this paper, we thus show the convergence of the Markov chain generated by DSGD, and derive an explicit expansion of its bias and variance.

- We highlight the fact that DSGD has linear speed-up *regardless of the topology*, and even *without averaging the local iterates*, up to higher-order terms.
- We show that stochasticity introduces an additional bias in the limit, corresponding to SGD's bias.
- We provide non-asymptotic convergence bounds for clients' local iterates.

DSGD as a Markov Chain. We introduce the following three assumptions on the noise model.

A 4 (Zero-mean noise)**.** *There exists a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ such that for all $t \in \mathbb{N}$ and $\Theta \in \mathbb{R}^{md}$, $\varepsilon_{t+1}(\Theta)$ is $\mathcal{F}_t$-measurable, and $\mathbb{E}\left[\varepsilon_{t+1}(\Theta) \mid \mathcal{F}_t\right] = 0$. Moreover, the sequence of random functions $(\varepsilon_t)_t$ is i.i.d., and the noises $\varepsilon_{t+1}^{(k)}$ of the different clients are independent.*

A 5 (p)**.** *(Co-coercivity) For every $t \geq 1$, define the stochastic gradient mapping $\nabla F(\cdot;\varepsilon_t) := \nabla F(\cdot) + \varepsilon_t(\cdot)$. Then, almost surely, $\nabla F(\cdot;\varepsilon_t)$ is L-co-coercive, i.e.*

$$\forall \Theta, \Theta' \in \mathbb{R}^{md}, \quad \|\nabla F(\Theta;\varepsilon_t) - \nabla F(\Theta';\varepsilon_t)\|^2 \\ \leq L\left\langle \nabla F(\Theta;\varepsilon_t) - \nabla F(\Theta';\varepsilon_t), \Theta - \Theta'\right\rangle .$$

In addition, the noise has a finite p-th moment at $\Theta^\star$: $\forall t \geq 1, \quad \mathbb{E}\left[\|\varepsilon_t(\Theta^\star)\|^p\right]^{1/p} \leq \tau_p$.

A 6 (Smoothness of the noise covariance)**.** *Define the covariance operator $\boldsymbol{\mathcal{C}}(\Theta) := \mathbb{E}\left[\varepsilon_1(\Theta)^{\otimes 2}\right]$. Assume that $\boldsymbol{\mathcal{C}}$ is twice continuously differentiable. Furthermore, we define $\mathcal{C}(\theta) := \frac{1}{m}\sum_{k=1}^m \mathbb{E}[(\varepsilon_1^{(k)}(\theta))^{\otimes 2}]$.*

These assumptions are standard when analyzing SGD iterates as a Markov chain [Dieuleveut et al., 2020, Mangold et al., 2025a,b]. They assume regularity of the noisy gradients, as well as bounded moments for the noise, which both hold in natural contexts such as least-squares regression or logistic regression under mild assumptions. Under A 4, the sequence of iterates $(\Theta_t)_{t\geq 0}$ generated by (DSGD) defines a time-homogeneous Markov chain on $\mathbb{R}^{md}$ with a Markov kernel defined as $R_\gamma(\Theta, A) := \mathbb{P}\left[\Theta_{t+1} \in A \mid \Theta_t = \Theta\right]$ for any Borel set $A \subseteq \mathbb{R}^{md}$.

Convergence to a Stationary Distribution. In the stochastic setting, Lemma 3.1 obviously no longer holds. However, we show below that the Markov chain $(\Theta_t)_{t\geq 0}$ converges in Wasserstein distance.

Proposition 4.1 (Convergence of DSGD)**.** *Assume A 1, A 2, A 4, and A 5(2). If $\gamma < 2/L$, then the Markov chain $(\Theta_t)_{t\geq 0}$ admits a unique stationary distribution $\pi_\gamma \in \mathcal{P}_2(\mathbb{R}^{md})$, i.e., with finite second moment. Moreover, for any initial distribution ρ_0 and any $t \geq 0$,*

$$\mathbf{W}_2^2\left(\rho_0 R_\gamma^t, \pi_\gamma\right) \leq \left(1 - 2\gamma\mu\left(1 - L\gamma/2\right)\right)^t \mathbf{W}_2^2\left(\rho_0, \pi_\gamma\right).$$

We prove this result in Appendix B.1. Proposition 4.1 shows that DSGD converges to a stationary regime, independently of the initialization. The convergence holds in the Wasserstein metric, with a geometric rate $1 - 2\gamma\mu\left(1 - L\gamma/2\right) \leq 1 - \gamma\mu < 1$. As expected, smaller step sizes yield slower convergence, but also reduce the deterministic bias of the algorithm (see Proposition 3.6).

Variance of DSGD. We now analyze the stationary distribution π_γ, by characterizing its mean and variance (with a slight abuse of language),

$$\Theta_{\text{sto}} := \int_{\mathbb{R}^{md}} \Theta \, d\pi_\gamma(\Theta) \ ,$$
$$\boldsymbol{\Sigma}_\theta := \int_{\mathbb{R}^{md}} (\Theta - \Theta_{\text{det}})^{\otimes 2} \, d\pi_\gamma(\Theta) \ .$$

The matrix $\boldsymbol{\Sigma}_\theta$ is of size $(md) \times (md)$. For $k, \ell \in \{1, \ldots, m\}$ we denote $[\boldsymbol{\Sigma}_\theta]_{k,\ell}$ its $d \times d$-size block of indices k and ℓ. We first give an expansion of the variance.

Proposition 4.2. *Assume A 1, A 2, A 4, A 5(4), A 6. Then for any $k, \ell \in \{1, \ldots, m\}$, it holds that*

$$[\boldsymbol{\Sigma}_\theta]_{k,\ell} = \frac{\gamma}{m} \boldsymbol{J} \mathcal{C}(\theta^\star) + \mathcal{O}(\gamma^{3/2}) \ ,$$

where we defined $\boldsymbol{J} = (\boldsymbol{I} \otimes \bar{A} + \bar{A} \otimes \boldsymbol{I})^{-1}$.

To establish this proposition, we linearize the gradients in DSGD's updates, expand the square, and bound the remainder terms; we give the detailed proof in Appendix B.2. There are two key take-aways from this proposition: (i) at the first-order in the step size, DSGD's variance decreases with the number of clients, resulting in a linear speed-up, and (ii) this is the case *regardless of the topology.* In other words, for two different graph topologies, the variance of the stationary distribution is essentially the same, provided that the step size γ is small enough. Interestingly, this first-order expression exactly matches with classical federated learning (see Theorem 4 of Mangold et al. [2025b]). The dominant contribution is determined by the noise and the local objectives, while the influence of the network structure only appears in higher-order terms. To highlight this phenomenon, we derive the following bound on the limit variance.

Lemma 4.3. *Assume A 1, A 2, A 4, A 5(2). Let $\gamma \leq \min(1/L, 1/(\Lambda L))$, then the variance of DSGD at stationarity is bounded by*

$$\frac{1}{m} \sum_{k=1}^{m} \|[\boldsymbol{\Sigma}_\theta]_{k,k}\| \lesssim \frac{\gamma \tau_2^2 + \gamma^{3/2} K_3 B^{3/2}}{\mu m} + \frac{\gamma^2 \rho^2}{1 - \rho^2} \tau_2^2 C \ ,$$

where $C = (LB + K_3 B^{3/2} \gamma^{1/2} + \frac{1}{2} \gamma^2 K_3^2 B^2 + \tau_2^2)/\tau_2^2$ and $B = \big(\tau_4^2 + \gamma^2 \frac{L^4}{\mu^2} \Lambda^2 \zeta_\star^2\big)/\mu$. We use $\lesssim$ to indicate an absolute numerical constant dependence.

To prove this bound, we leverage the decomposition we used to prove Proposition 4.2, and decompose it between a *consensus*, which dominates and *disagreement* part, which gives higher-order terms. We give a detailed proof in Appendix C.1. The key takeaway of this lemma is that as long as

$$m = \mathcal{O}(1/(\mu \gamma \Lambda^2)) \ ,$$

the leading term is the term in γ, which does not depend on the topology. More precisely, as long as the number of clients is of order $O(1/\Lambda^2)$, the impact of the graph's topology on the variance is negligible.

Bias of DSGD. First, we remark that stochasticity does not affect the bias for *quadratic* functions.

Proposition 4.4. *Assume A 2, A 3, A 4, A 5(4), A 6. Then stochasticity does not introduce any additional bias, and it holds that $\Theta_{\text{sto}} = \Theta_{\text{det}}$.*

The proof follows from the linearity of the gradient, see Appendix B.2. For general smooth and strongly convex functions, however, stochasticity induces an additional first-order bias in the stationary mean.

Proposition 4.5. *Suppose A 1, A 2, A 4, A 5(4), A 6. Then for any $i \in \{1, \ldots, m\}$,*

$$\theta_i^{\text{sto}} - \theta_i^{\text{det}} = -\frac{\gamma}{2m} \nabla^2 f(\theta^\star)^{-1} \nabla^3 f(\theta^\star) \boldsymbol{J} \mathcal{C}(\theta^\star) + \mathcal{O}(\gamma^{3/2}) \, ,$$

where $\boldsymbol{J}$ is defined in Proposition 4.2.

Proposition 4.5 shows that for smooth and strongly convex functions, the mean of the stationary distribution does not coincide with Θ_{det}: an additional bias of order $\mathcal{O}(\gamma)$ appears. We emphasize two key points: (i) the stationary mean Θ_{sto} still depends on the network topology, since Θ_{det} does; (ii) yet, the *additional stochastic bias* $\Theta_{\text{sto}} - \Theta_{\text{det}}$ and the stationary variance are *independent of the topology at first order in γ.*

Non-Asymptotic Bounds We now state one of our main results, which gives a non-asymptotic convergence rate for DSGD based on bounds on the asymptotic variance and geometric ergodicity of DSGD.

Theorem 4.6. *Assume A 1, A 2, A 4, A 5(4), A 6. Taking $\gamma \leq \min(\frac{1}{\Lambda L}, \frac{1}{10L})$, it holds that*

$$\begin{aligned} \mathbb{E}[\|\Theta_t - \Theta^\star\|^2] \lesssim\ & (1 - \gamma\mu)^t \psi_0 + \gamma \tfrac{\tau_2^2}{\mu m} + \gamma^{3/2} \tfrac{K_3 B^{3/2}}{\mu m} \\ & + \gamma^2 \Big(\tfrac{L^2 \Lambda^2}{\mu^2} \zeta_\star^2 + \tfrac{L B \rho^2}{1 - \rho^2} + \tfrac{\rho^2}{1 - \rho^2} \tau_2^2 \Big) \\ & + \gamma^{5/2} \tfrac{\rho^2}{1 - \rho^2} \big(K_3 B^{3/2} + \gamma^{3/2} K_3^2 B^2 \big) \ , \end{aligned}$$

with $\psi_0 = \|\Theta_0 - \Theta^\star\|^2 + \frac{\gamma^2 L^2 \Lambda^2}{\mu^2} \zeta_\star^2 + \frac{\gamma}{\mu} (\tau_2^2 + \gamma^2 \frac{4L^4}{\mu^2} \Lambda^2 \zeta_\star^2)$ and $B = \frac{1}{\mu} \big(\tau_4^2 + \gamma^2 \frac{L^4}{\mu^2} \Lambda^2 \zeta_\star^2 \big)$.

We prove this Theorem in Appendix C. We stress that the proof scheme is *fundamentally different* from existing convergence proofs of DSGD, as it does not rely on establishing a bound on the convergence of consensus/disagreement part, but rather studies convergence of the algorithm, together with its properties in its stationary regime. We can now give a sample complexity for this algorithm.

Corollary 4.7. *Let $\epsilon > 0$ be small enough. Under the assumptions of Theorem 4.6, set* $\gamma = \min(\frac{1}{L}, \frac{\mu}{L^2\Lambda\zeta_\star}, \frac{\mu m\epsilon^2}{\tau_2^2}, \frac{\mu\epsilon}{L\Lambda\zeta_\star}, \epsilon(\frac{1-\rho^2}{LB\rho^2})^{1/2}, \epsilon(\frac{1-\rho^2}{\tau_2^2\rho^2})^{1/2})$*, the DSGD algorithm achieves* $\mathbb{E}[\|\Theta_t - \Theta^\star\|^2] \le \epsilon^2$ *after*

$$T \gtrsim \max(\tfrac{L}{\mu}, \tfrac{L^2\Lambda\zeta_\star}{\mu^2}, \tfrac{\tau_2^2}{\mu^2 m\epsilon^2}, \tfrac{L\Lambda\zeta_\star}{\mu^2\epsilon}, \tfrac{\tau_2}{\mu\epsilon}(\tfrac{\rho^2}{1-\rho^2})^{1/2}) \log(\tfrac{\psi_0}{\epsilon}) \,.$$

This corollary establishes the sample complexity of DSGD. It shows that, when high precision is desired, DSGD has linear speed-up until $m \lesssim \min(\frac{\tau^2}{L\Lambda\zeta_\star\epsilon}, \frac{((1-\rho^2)/\rho^2)^{1/2}}{\mu\epsilon})$. In other words, provided the network topology is not too disconnected and the problem not too heterogeneous, DSGD achieves linear speed-up in the number of clients.

5 RICHARDSON-ROMBERG FOR DECENTRALIZED LEARNING

Building on our analysis of DSGD's bias, we extend the Richardson-Romberg extrapolation technique to the decentralized world. The main idea of this technique, initially developed in numerical analysis, is to cancel leading-order error terms by combining iterates obtained with different step sizes. It has since found applications in diverse areas, like data science [Bach, 2021], and in the analysis of SGD algorithms, both in the single-client setting [Dieuleveut et al., 2020] and in federated learning [Mangold et al., 2025b].

To construct the decentralized Richardson-Romberg estimator, we set a step size $\gamma \le 1/L$, and run the algorithm two times, giving two sequences of iterates: one with step size γ and one with step size $\gamma/2$, denoted respectively by $(\Theta_t^\gamma)_{t\ge0}$ and $(\Theta_t^{\gamma/2})_{t\ge0}$. We then define the *Richardson-Romberg extrapolated iterate* as

$$\Theta_t^{\mathrm{RR},\gamma} := 2\Theta_t^{\gamma/2} - \Theta_t^\gamma \ . \tag{11}$$

We stress that building this estimator only requires running the algorithm twice. Contrary to gradient tracking methods [Di Lorenzo and Scutari, 2016, Koloskova et al., 2021], it does not require clients to maintain a slack variable in memory. This makes this algorithm fit for running on devices with low-resources, which are typical in decentralized settings.

Leveraging our expression of DSGD's bias, we show that the estimator built in (11) has reduced bias.

Proposition 5.1. *Assume A 1-2. If* $\gamma \le \min\left(\frac{1}{\Lambda L}, \frac{1}{L}\right)$,

$$\|\Theta_{\mathrm{det}}^{\mathrm{RR},\gamma} - \Theta^\star\| \le \gamma^2 \tfrac{L^2}{\mu^2}\Lambda^2\left(\tfrac{K_3}{\mu\sqrt{m}}\zeta_\star^2 + L\zeta_\star\right) \ .$$

This shows that Richardson-Romberg extrapolation successfully eliminates the deterministic first-order bias term $\gamma\cdot(\boldsymbol{I}-\boldsymbol{\mathcal{H}})\boldsymbol{\mathcal{G}}\nabla F(\Theta^\star)$, which depends on both the topology of the communication graph and the heterogeneity. Remarkably, this does not require explicit knowledge of the graph structure or of the heterogeneity. We then obtain the following convergence rate.

Corollary 5.2. *Assume A 1-2. Then, for any $t \ge 0$, in the deterministic setting, we have the convergence rate*

$$\begin{aligned}\|\Theta_t^{\mathrm{RR},\gamma} - \Theta^\star\| &\le 3(1-\gamma\mu/2)^t\|\Theta_0 - \Theta^\star\| \\ &\quad + 2(1-\gamma\mu/2)^t\frac{\gamma L\Lambda}{\mu}\zeta_\star \\ &\quad + \gamma^2\frac{5L^2}{8\mu^2}\Lambda^2\Big(\frac{K_3}{\mu\sqrt{m}}\zeta_\star^2 + L\zeta_\star\Big) \ .\end{aligned}$$

For $\epsilon > 0$, let $\gamma = \min(\frac{1}{L}, \frac{1}{\Lambda L}, \frac{\mu\epsilon^{1/2}}{L\Lambda\zeta_\star})$ *and assume $K_3/m^{1/2}$ is small, then we have* $\|\Theta_t^{\mathrm{RR},\gamma} - \Theta^\star\| \le \epsilon$ *after* $T \gtrsim \max(\frac{L}{\mu}, \frac{\Lambda L}{\mu}, \frac{L\Lambda\zeta_\star}{\mu^2\epsilon^{1/2}})\log(\frac{\|\Theta_0-\Theta^\star\|}{\epsilon})$ *iterations.*

This shows that the speed of convergence is not affected, but that the limit bias is quadratically reduced, allowing for a quadratic improvement in the sample complexity when heterogeneity is large or the network not very connected. Finally, we stress that similar results would hold in the stochastic setting, allowing to speed-up convergence when the variance is dominated by the decentralization error. Thus, this allows to handle decentralized learning with much less connected graphs, de facto reducing communications.

In the stochastic setting, combining Corollary 5.2 with Lemma 4.3 gives the following result.

Corollary 5.3. *Let $\epsilon > 0$ be small enough. Under the assumptions of Theorem 4.6, set* $\gamma = \min(\frac{1}{L}, \frac{\mu}{L^2\Lambda\zeta_\star}, \frac{\mu m\epsilon^2}{\tau_2^2}, \frac{\mu\epsilon^{1/2}}{L\Lambda\zeta_\star}, \epsilon(\frac{1-\rho^2}{LB\rho^2})^{1/2}, \epsilon(\frac{1-\rho^2}{\tau_2^2\rho^2})^{1/2})$*, the RR-DSGD algorithm achieves* $\mathbb{E}[\|\Theta_t^{\mathrm{RR},\gamma} - \Theta^\star\|^2] \le \epsilon^2$ *for*

$$T \gtrsim \max(\tfrac{L}{\mu}, \tfrac{L^2\Lambda\zeta_\star}{\mu^2}, \tfrac{\tau_2^2}{\mu^2 m\epsilon^2}, \tfrac{L\Lambda\zeta_\star}{\mu^2\epsilon^{1/2}}, \tfrac{\tau_2}{\mu\epsilon}(\tfrac{\rho^2}{1-\rho^2})^{1/2}) \log(\tfrac{\psi_0}{\epsilon}) \,.$$

This shows that the Richardson-Romberg extrapolation procedure has the same advantages in the stochastic regime and in the deterministic one. Notably, it can be particularly beneficial in settings where noise is dominated by the bias due to decentralization.

6 EXPERIMENTS

This section illustrates our theoretical results through numerical experiments. We consider the optimization problem (1), where for client $k \in \{1, \dots, m\}$, the local objective is given by $f_k(\theta) = \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(\langle\theta, \theta_{k,i}\rangle)) + \frac{\lambda}{2}\|\theta\|^2$. For each client k, the vectors $\theta_{k,i}$ are sampled independently from a distribution specific to that client, which enables us to model

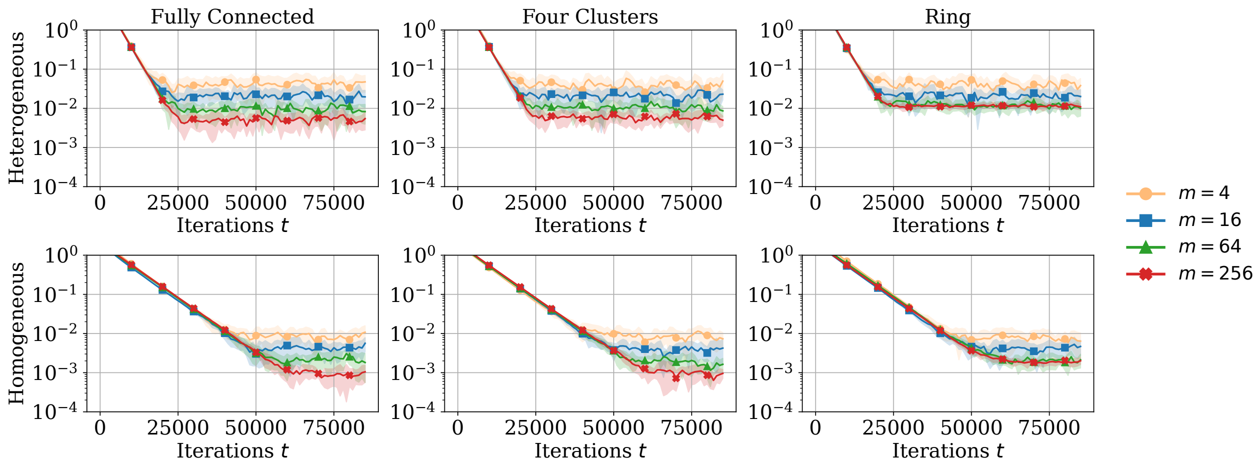

Figure 1: DSGD for heterogeneous (top row) and homogeneous (bottom row) clients, for various numbers of clients m and communication graphs. Graphs are fully connected (left), four clusters sparsely connected (middle), and ring (right). Colored areas indicate variations ($\pm$ standard deviation) obtained from 20 independent runs.

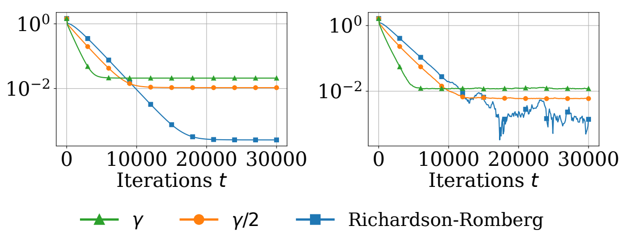

Figure 2: $\frac{1}{m}\sum_{i=1}^{m}\|\theta_i^t - \theta^*\|$ for deterministic (left) and stochastic (right) DGD with step size γ, step size $\gamma/2$ and RR extrapolation.

heterogeneity across the network. All experiments are conducted in dimension $d = 2$, and the communication graph is chosen to be either (i) fully connected, with $W = \frac{1}{m}\mathbf{1}\mathbf{1}^\top$), (ii) a ring topology, or (iii) a clustered topology with four well-connected clusters that are only sparsely connected to each other.

Deterministic DGD. In the deterministic setting, Lemma 3.1 and Proposition 3.6 show that using a smaller step size γ reduces the bias, but slows down convergence. Moreover, Proposition 5.1 shows that that Richardson-Romberg extrapolation improves the bias order from $\mathcal{O}(\gamma)$ for classical DGD to $\mathcal{O}(\gamma^2)$. This is indeed what we observe in Figure 2, where $m = 12, \gamma = 10^{-3}$, with the four-clusters graph.

Stochastic DGD. In the stochastic setting, Proposition 4.2 and Proposition 4.5 establish that both the variance and the bias scale as $1/m$ at the first-order. We illustrated this numerically by plotting the evolution of the distance to the deterministic limit Θ_{det} as a function of the iteraiton number, for different m and graphs. As observed in Figure 1 ($\gamma = 10^{-3}$), increasing the number of clients yields smaller variance and bias. However, for poorly connected graphs like the ring topology, the improvement stops when the bias begins: since the second-largest eigenvalue of W is very close to 1, higher-order error that depend on the graph prevent further speed-up.

Decentralized Richardson-Romberg. We now illustrate the performance of the decentralized Richardson-Romberg extrapolation method. In both deterministic and stochastic settings, it effectively reduces DSGD's bias by one order of magnitude. In Figure 2 (left), we run the deterministic variant of the algorithm, confirming a reduction of the bias from 10^{-2} to 10^{-4}. The stochastic counterpart of the algorithm, reported in Figure 2 (right), also reduces the bias: there, bias strongly hinders DSGD's convergence, while DSGD with Richardson-Romberg extrapolation reaches a stationary regime, where variance dominates.

7 CONCLUSION

In this paper, we analyzed the deterministic and stochastic Decentralized Gradient Descent algorithm, focusing on the characterization of their limiting points and the impact of heterogeneity, network topology, and stochastic noise. In the deterministic setting, we showed that DGD's bias depends on the step size, network topology, and heterogeneity, with a bias of order $\mathcal{O}(\gamma)$ that can be reduced to $\mathcal{O}(\gamma^2)$ with Richardson–Romberg extrapolation. In the stochastic setting, the iterates admit a unique stationary distribution with an additional $\mathcal{O}(\gamma)$ bias, and have a linear speed-up in the number of clients even without formally averaging the iterates.

Overall, our results reveal a clear contrast: deterministic bias is shaped by network effects, while stochastic bias is dominated by noise. This deepens our understanding of the impact of decentralization and stochasticity on optimization, and hints towards novel methods like, e.g., Richardson–Romberg extrapolation. We proved that this method reduces DSGD's sample complexity, notably in small noise regimes. We hope that our analysis of DSGD will serve as a basis for the development of decentralized algorithms tailored to handle stochasticity.

ACKNOWLEDGEMENTS

The work of Aymeric Dieuleveut and Lucas Versini is supported by French State aid managed by the Agence Nationale de la Recherche (ANR) under France 2030 program with the reference ANR-23-PEIA-005 (REDEEM project). This work is supported by Hi! PARIS and ANR/France 2030 program (ANR-23-IACL-0005), in particular Hi!Paris FLAG chair.

References

Alekh Agarwal and John C Duchi. Distributed delayed stochastic optimization. *Advances in neural information processing systems*, 24, 2011.

Alekh Agarwal, Olivier Chapelle, Miroslav Dudík, and John Langford. A reliable effective terascale linear learning system. *The Journal of Machine Learning Research*, 15(1):1111–1133, 2014.

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

Mahmoud Assran, Nicolas Loizou, Nicolas Ballas, and Mike Rabbat. Stochastic gradient push for distributed deep learning. In *International Conference on Machine Learning*, pages 344–353. PMLR, 2019.

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

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Kevin Scaman, and Giovanni Neglia. Improved stability and generalization guarantees of the decentralized sgd algorithm. *arXiv preprint arXiv:2306.02939*, 2023.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

Zachary Charles and Jakub Konečnỳ. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2575–2583. PMLR, 2021.

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

Jerry Chee and Panos Toulis. Convergence diagnostics for stochastic gradient descent with constant learning rate. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1476–1485. PMLR, 09–11 Apr 2018.

Woocheol Choi and Jimyeong Kim. On the convergence analysis of the decentralized projected gradient descent method. *SIAM Journal on Optimization*, 35(3):1673–1702, 2025. doi: 10.1137/23M1562032. URL https://doi.org/10.1137/23M1562032.

Sebastien Colla and Julien M Hendrickx. Automated worst-case performance analysis of decentralized gradient descent. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 2627–2633. IEEE, 2021.

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.

Christopher M De Sa, Ce Zhang, Kunle Olukotun, and Christopher Ré. Taming the wild: A unified analysis of hogwild-style algorithms. *Advances in neural information processing systems*, 28, 2015.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. *Advances in neural information processing systems*, 25, 2012.

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.

Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

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.

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.

John Duchi, Michael I Jordan, and Brendan McMahan. Estimation, optimization, and parallelism when data is sparse. *Advances in Neural Information Processing Systems*, 26, 2013.

John C Duchi, Alekh Agarwal, and Martin J Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3):592–606, 2011.

Mathieu Even, Anastasia Koloskova, and Laurent Massoulié. Asynchronous sgd on graphs: a unified framework for asynchronous decentralized and federated optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 64–72. PMLR, 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.

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

Hadrien Hendrikx, Francis Bach, and Laurent Massoulie. An optimal algorithm for decentralized finite-sum optimization. *SIAM Journal on Optimization*, 31(4):2753–2783, 2021.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alexander J Smola. On variance reduction in stochastic gradient descent and its asynchronous variants. *Advances in neural information processing systems*, 28, 2015.

Dušan Jakovetić. A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):31–46, 2018.

Yiming Jiang, Helei Kang, Jinlan Liu, and Dongpo Xu. On the convergence of decentralized stochastic gradient descent with biased gradients. *IEEE Transactions on Signal Processing*, 73:549–558, 2025. doi: 10.1109/TSP.2025.3531356.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

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

Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International conference on machine learning*, pages 3478–3487. PMLR, 2019.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized SGD with changing topology and local updates. In *International conference on machine learning*, pages 5381–5393. PMLR, 2020.

Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34:11422–11435, 2021.

Dmitry Kovalev, Adil Salim, and Peter Richtárik. Optimal and practical algorithms for smooth and strongly convex decentralized optimization. *Advances in Neural Information Processing Systems*, 33:18342–18352, 2020.

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

Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, 180(1):237–284, 2020.

Erik G Larsson and Nicolo Michelusi. Unified analysis of decentralized gradient descent: a contraction mapping framework. *IEEE Open Journal of Signal Processing*, 2025.

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Erick Lavoie, and Anne-Marie Kermarrec. Refined convergence and topology learning for decentralized sgd with heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 1672–1702. PMLR, 2023.

Batiste Le Bars, Aurélien Bellet, Marc Tommasi, Kevin Scaman, and Giovanni Neglia. Improved stability and generalization guarantees of the decentralized sgd algorithm. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Rémi Leblond, Fabian Pedregosa, and Simon Lacoste-Julien. Asaga: Asynchronous parallel saga. In *Artificial Intelligence and Statistics*, pages 46–54. PMLR, 2017.

Huan Li and Zhouchen Lin. Revisiting extra for smooth distributed optimization. *SIAM Journal on Optimization*, 30(3):1795–1821, 2020.

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

Zhi Li, Wei Shi, and Ming Yan. A decentralized proximal-gradient method with network independent step-sizes and separated convergence rates. *IEEE Transactions on Signal Processing*, 67(17): 4494–4506, 2019b.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Liu, Ji Li, Cho-Jui Hsieh, and Ji Zhang. Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30:5330–5340, 2017.

Grigory Malinovskiy, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtarik. From local sgd to local fixed-point methods for federated learning. In *International Conference on Machine Learning*, pages 6692–6701. PMLR, 2020.

Paul Mangold, Alain Durmus, Aymeric Dieuleveut, and Eric Moulines. Scaffold with stochastic gradients: New analysis with linear speed-up. *arXiv preprint arXiv:2503.07594*, 2025a.

Paul Mangold, Alain Durmus, Aymeric Dieuleveut, Sergey Samsonov, and Eric Moulines. Refined analysis of constant step size federated averaging and federated richardson-romberg extrapolation. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258, pages 5023–5031, 2025b.

Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, 2017.

Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon Mann. Efficient large-scale distributed training of conditional maximum entropy models. *Advances in neural information processing systems*, 22, 2009.

Ryan McDonald, Keith Hall, and Gideon Mann. Distributed training strategies for the structured perceptron. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 456–464, 2010.

H Brendan McMahan, Eider Moore, Daniel Ramage, and Seth Hampson. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282. PMLR, 2017.

E. Moulines, P. Priouret, and R. Aguech. On a perturbation approach for the analysis of stochastic tracking algorithms. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 3, pages 1681–1684 vol.3, 1998. doi: 10.1109/ICASSP.1998.681779.

Eric Moulines, Pierre Priouret, and François Roueff. On recursive estimation for time varying autoregressive processes. *The Annals of Statistics*, 2005.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009a. doi: 10.1109/TAC.2008.2009515.

Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on automatic control*, 54(1):48–61, 2009b.

Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

Giovanni Neglia, Chuan Xu, Don Towsley, and Gianmarco Calbi. Decentralized gradient methods: does topology matter? In *International Conference on Artificial Intelligence and Statistics*, pages 2348–2358. PMLR, 2020.

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.

Reese Pathak and Martin J Wainwright. Fedsplit: An algorithmic framework for fast federated optimization. *Advances in neural information processing systems*, 33:7057–7066, 2020.

Georg Ch. Pflug. Stochastic minimization with constant step-size: Asymptotic laws. *SIAM Journal on Control and Optimization*, 24(4):655–666, 1986a. doi: 10.1137/0324039.

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

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457, 2021.

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.

Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *Advances in neural information processing systems*, 24, 2011.

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

Dominic Richards and Patrick Rebeschini. Graph-dependent implicit regularisation for distributed stochastic subgradient descent. *Journal of Machine Learning Research*, 21(34):1–44, 2020.

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.

Ali H Sayed. Diffusion adaptation over networks. In *Academic Press Library in Signal Processing*, volume 3, pages 323–453. Elsevier, 2014.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *international conference on machine learning*, pages 3027–3036. PMLR, 2017.

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.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Benjamin Sirb and Xiaojing Ye. Consensus optimization with delayed and stochastic gradients on decentralized networks. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 76–85, 2016. doi: 10.1109/BigData.2016.7840591.

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

Josef Stoer, Roland Bulirsch, R Bartels, Walter Gautschi, and Christoph Witzgall. *Introduction to numerical analysis*, volume 1993. Springer, 1980.

S Sundhar Ram, Angelia Nedić, and Venugopal V Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications*, 147(3):516–545, 2010.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856. PMLR, 2018.

John Tsitsiklis, Dimitri Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE transactions on automatic control*, 31(9):803–812, 2003.

César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. A dual approach for optimal algorithms in distributed optimization over networks. In *2020 Information theory and applications workshop (ITA)*, pages 1–37. IEEE, 2020.

Cédric Villani. *Optimal transport : old and new / Cédric Villani*. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009. ISBN 978-3-540-71049-3.

Thijs Vogels, Hadrien Hendrikx, and Martin Jaggi. Beyond spectral gap: The role of the topology in decentralized learning. *Advances in Neural Information Processing Systems*, 35:15039–15050, 2022.

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, Zheng Xu, Zachary Garrett, Zachary Charles, Luyang Liu, and Gauri Joshi. Local adaptivity in federated learning: Convergence and consistency. *arXiv preprint arXiv:2106.02305*, 2021.

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.

Jinming Xu, Ye Tian, Ying Sun, and Gesualdo Scutari. A unified algorithmic framework for distributed composite optimization. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 2309–2316. IEEE, 2020.

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.

Haoxiang Ye, Tao Sun, and Qing Ling. Generalization guarantee of decentralized learning with heterogeneous data. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

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.

Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

Jiaqi Zhang and Keyou You. Decentralized stochastic gradient tracking for non-convex empirical risk minimization. *arXiv preprint arXiv:1909.02712*, 2019.

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.

Martin Zinkevich, John Langford, and Alex Smola. Slow learners are fast. *Advances in neural information processing systems*, 22, 2009.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.

Supplementary Materials

In this appendix, we first provide additional related work.

- Appendix A contains the proofs of the deterministic DGD results stated in Section 3.
- In Appendix B, we give the proofs of the results on DSGD stated in Section 4. We also state and prove additional results:
  - Lemma B.1 and Lemma B.2 give contraction results between consecutive iterates of DSGD, as well as upper-bounds of moments of order 2 and 4 under the stationary distribution.
  - Based on these first two lemmas, Lemma B.4 establishes a first expansion of the stationary distribution's variance.
  - From this expression of the variance, Lemma B.5 deduces an expression of the stationary distribution's bias.
  - Lemma B.6 and Lemma B.7 give estimates of the covariance of the noise under the stationary distribution, based on Lemma B.1, Lemma B.2 and Lemma B.5.
  - Based on these previous lemmas, Corollary B.8 and Lemma B.9 then simplify the expression of the variance, which yields both Proposition 4.2 and Proposition 4.5.
- Appendix C contains the proofs of the non-asymptotic results stated in Section 4.
- In Appendix D, we give the proofs of the results on Richardson-Romberg extrapolation.
- Appendix E contains two general lemmas on matrices, used in the proofs of the previous results.

Additional Related Work

Decentralized Optimization. The term decentralized SGD was introduced by Lian et al. [2017]. Nonetheless, similar algorithms had been studied in *distributed* optimization, with the first analysis by Nedic and Ozdaglar [2009b], and algorithms like dual averaging [Duchi et al., 2011], Extra [Shi et al., 2015], and other algorithms mentioned below. Many analyses [Hendrikx et al., 2021] and variants of decentralized SGD have then been studied, with variance reduction [Tang et al., 2018], compression [Koloskova et al., 2019], for neural networks [Assran et al., 2019], with changing topology [Koloskova et al., 2020, Le Bars et al., 2023], biased gradients [Jiang et al., 2025], asynchronous updates [Even et al., 2024] or projections [Choi and Kim, 2025]. Other works have studied the generalization of decentralized SGD [Le Bars et al., 2024, Ye et al., 2025]. All these approaches still suffer from bias due to decentralization and heterogeneity, and some methods have been proposed to mitigate this bias [Zhang and You, 2019, Pu and Nedić, 2021, Koloskova et al., 2021]. Other algorithms have also been proposed, like DIGing [Nedic et al., 2017], NIDS [Li et al., 2019b], together with refined analyses of these methods [Jakovetić, 2018, Xu et al., 2020, Li and Lin, 2020] A long line of work have been dedicated to the study of dual algorithms [Scaman et al., 2017, Uribe et al., 2020, Kovalev et al., 2020]

Distributed and Parallel Stochastic Gradient Descent. The methods from decentralized optimization find their roots in distributed optimization [Tsitsiklis et al., 2003, Nedic and Ozdaglar, 2009b, Boyd et al., 2011]. In this context, the goal is essentially to accelerate the learning by leveraging the computational power of multiple machines. A key difficulty in this setting is to handle asynchronous updates, which cause delay [Zinkevich et al., 2009]. This was studied to learn conditional entropy models [Mcdonald et al., 2009], distributed neural networks [McDonald et al., 2010, Dean et al., 2012], and for parallel SGD in general [Zinkevich et al., 2010, Agarwal et al., 2014, Agarwal and Duchi, 2011]. and more generally in the context of asynchronous SGD without locks [Recht et al., 2011, Duchi et al., 2013, Leblond et al., 2017, De Sa et al., 2015, J Reddi et al., 2015, Mania et al., 2017]. Note that in the more specific setting of federated learning (with a central server) similar ideas of linear speed-up in the number of clients have been observed [Yang et al., 2021, Qu et al., 2021, Mangold et al., 2025b,a].

Fixed-Point Analyses of Decentralized Learning. Closely related to our work, Vogels et al. [2022] studied the convergence of decentralized GD to the solution of an equation involving the gradient and the network, providing convergence rates for DSGD to a neighborhood of this point in expectation. Similarly, Larsson and Michelusi [2025] proved the convergence to a fixed point in the deterministic setting. Additionally, Yuan et al. [2016] showed convergence of decentralized GD by interpreting it as gradient descent on a modified function.

In the federated learning setting, when a central server is available, FedAvg has also been studied under the fixed-point framework, first in the deterministic setting [Malinovskiy et al., 2020, Wang et al., 2021], establishing convergence to a point, which can be related to the global solution [Malinovskiy et al., 2020, Charles and Konečnỳ, 2021, Pathak and Wainwright, 2020], with an explicit characterization of the bias in the quadratic case. This fixed-point framework was later extended to the stochastic setting, providing new analyses of FedAvg and Scaffold [Mangold et al., 2025b,a].

Other Approaches in Federated Learning. When a central server can orchestrate the training, decentralized learning boils down to federated learning [McMahan et al., 2017]. The most prominent algorithm in this setting is the FedAvg [McMahan et al., 2017], which has been widely studied. A first line of work has studied this method in the homogeneous setting, where clients share the same objective function [Stich, 2019, Wang and Joshi, 2018, Haddadpour and Mahdavi, 2019, Yu et al., 2019b, Wang and Joshi, 2018, Li et al., 2019a]. Then, many works studied the properties of FedAvg when clients differ, inducing a client-drift phenomenon [Karimireddy et al., 2020]. Many heterogeneity measures have been proposed, based on first-order information [Yu et al., 2019a, Khaled et al., 2020, Karimireddy et al., 2020, Reddi et al., 2021, Zindari et al., 2023, Crawshaw et al., 2024], 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].

In analyses of FedAvg with stochastic gradients, a bias appears as higher order terms [Khaled et al., 2020, Glasgow et al., 2022, Wang et al., 2024]. This bias essentially stems from the bias of SGD, which has been widely studied in the single-client setting [Lan, 2012, Défossez and Bach, 2015, Dieuleveut and Bach, 2016, Chee and Toulis, 2017], and through Markov chain analyses of SGD [Dieuleveut et al., 2020, Pflug, 1986b].

Richardson-Romberg. Richardson-Romberg extrapolation, which we extend to the decentralized setting for the first time, can be traced back to Richardson [1911]. This is a classical method in numerical analysis, with many applications, including time-varying autoregressive processes [Moulines et al., 2005], data science [Bach, 2021], and many others [Stoer et al., 1980]. It was brought to stochastic approximation by Dieuleveut et al. [2020], and later studied by Sheshukova et al. [2024]. Most closely related to our work, Mangold et al. [2025b] proposed a variant of this method in the context of federated learning, when a central server is available.

A DETERMINISTIC DECENTRALIZED GRADIENT DESCENT

Hereafter, we recall the proof of the convergence in the deterministic case [Larsson and Michelusi, 2025].

Proof of Lemma 3.1. The proof consists in noticing that the operator $\Theta \mapsto \boldsymbol{W}\big(\Theta - \gamma \nabla F(\Theta)\big)$ is $(1-\gamma\mu)$-Lipschitz, which comes from

$$
\begin{aligned}
\|\boldsymbol{W}\big(\Theta - \gamma\nabla F(\Theta)\big) - \boldsymbol{W}\big(\Theta' - \gamma\nabla F(\Theta')\big)\| &= \|\boldsymbol{W}\big(\Theta - \Theta' - \gamma\nabla F(\Theta) + \gamma\nabla F(\Theta')\big)\| \\
&\leq \|\Theta - \Theta' - \gamma(\nabla F(\Theta) - \nabla F(\Theta'))\| \\
&= \|(\boldsymbol{I} - \gamma A)(\Theta - \Theta')\| \ ,
\end{aligned}
$$

where $A := \displaystyle\int_0^1 \nabla^2 F(\Theta' + t(\Theta - \Theta'))\, \mathrm{d}t$ satisfies $\mu \boldsymbol{I} \preceq A \preceq L\boldsymbol{I}$.

For $\gamma < 1/L$, $\|\boldsymbol{I} - \gamma A\|_2 \leq 1 - \gamma\mu$, hence

$$
\|\boldsymbol{W}\big(\Theta - \gamma\nabla F(\Theta)\big) - \boldsymbol{W}\big(\Theta' - \gamma\nabla F(\Theta')\big)\| \leq (1 - \gamma\mu)\|\Theta - \Theta'\| \ .
$$

We conclude using Banach fixed-point theorem. ☐

Proof of Lemma 3.2. Applying the projection $\boldsymbol{\mathcal{P}}$ on the consensus space to (3), and using the fact that $\boldsymbol{\mathcal{P}}\boldsymbol{W} = \boldsymbol{\mathcal{P}}$ since W is doubly stochastic, we have $\boldsymbol{\mathcal{P}}\nabla F(\Theta_{\mathrm{det}}) = 0$, which gives a first equation

$$\frac{1}{m}\sum_{k=1}^{m} \nabla f_k(\theta_k^{\mathrm{det}}) = 0 \ .$$

Next, since $\overline{\Theta}_{\mathrm{det}} = \boldsymbol{W}\overline{\Theta}_{\mathrm{det}}$, the fixed-point equation (3) implies

$$(\boldsymbol{I} - \boldsymbol{W})\widetilde{\Theta}_{\mathrm{det}} = -\gamma \boldsymbol{W}\nabla F(\Theta_{\mathrm{det}}) \ ,$$

which gives the expression for $\widetilde{\Theta}_{\mathrm{det}}$ by applying the Moore-Penrose pseudoinverse of $\boldsymbol{I} - \boldsymbol{W}$. □

Proof of Proposition 3.3. From Lemma 3.2, we know

$$\frac{1}{m}\sum_{k=1}^{m} \nabla f_k(\theta_k^{\mathrm{det}}) = \frac{1}{m}\sum_{k=1}^{m} A_k(\theta_k^{\mathrm{det}} - \theta_k^\star) = 0 \ .$$

Using the decomposition $\theta_k^{\mathrm{det}} = \bar{\theta}^{\mathrm{det}} + \widetilde{\theta}_k^{\mathrm{det}}$, and the expression of the gradient for quadratic functions, we obtain

$$\bar{A}\bar{\theta}^{\mathrm{det}} = \frac{1}{m}\sum_{k=1}^{m} A_k\theta_k^\star - \frac{1}{m}\sum_{k=1}^{m} A_k\widetilde{\theta}_k^{\mathrm{det}} \ ,$$

which can be rewritten in compact matrix form, using the matrix $\boldsymbol{\mathcal{H}}$ defined in Proposition 3.3 and the fact that $\bar{A}$ is invertible, as

$$\overline{\Theta}_{\mathrm{det}} = \boldsymbol{\mathcal{H}}\Theta^\star_{(\mathrm{loc})} - \boldsymbol{\mathcal{H}}\widetilde{\Theta}_{\mathrm{det}} \ . \tag{12}$$

Lemma 3.2 also provides the following equation:

$$\widetilde{\Theta}_{\mathrm{det}} = -\gamma(\boldsymbol{I} - \boldsymbol{W})^\dagger \boldsymbol{W}\nabla F(\Theta_{\mathrm{det}}) = -\gamma\boldsymbol{\mathcal{G}}\boldsymbol{A}(\Theta_{\mathrm{det}} - \Theta^\star_{(\mathrm{loc})}).$$

Plugging in $\Theta_{\mathrm{det}} = \overline{\Theta}_{\mathrm{det}} + \widetilde{\Theta}_{\mathrm{det}}$, we get:

$$(\boldsymbol{I} + \gamma\boldsymbol{\mathcal{G}}\boldsymbol{A})\widetilde{\Theta}_{\mathrm{det}} = -\gamma\boldsymbol{\mathcal{G}}\boldsymbol{A}(\overline{\Theta}_{\mathrm{det}} - \Theta^\star_{(\mathrm{loc})}).$$

Plugging this in (12), we obtain:

$$\widetilde{\Theta}_{\mathrm{det}} = -\gamma(\boldsymbol{I} + \gamma\boldsymbol{\mathcal{G}}\boldsymbol{A})^{-1}\boldsymbol{\mathcal{G}}\boldsymbol{A}\left((\boldsymbol{\mathcal{H}} - \boldsymbol{I})\Theta^\star_{(\mathrm{loc})} - \boldsymbol{\mathcal{H}}\widetilde{\Theta}_{\mathrm{det}}\right).$$

We then define

$$\boldsymbol{\mathcal{B}} := (\boldsymbol{I} + \gamma\boldsymbol{\mathcal{G}}\boldsymbol{A})^{-1}\boldsymbol{\mathcal{G}}\boldsymbol{A},$$

and we get the closed-form

$$\widetilde{\Theta}_{\mathrm{det}} = \gamma(\boldsymbol{I} - \gamma\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{H}})^{-1}\boldsymbol{\mathcal{B}}(\boldsymbol{I} - \boldsymbol{\mathcal{H}})\Theta^\star_{(\mathrm{loc})},$$

Substituting back into (12) gives the expression for $\overline{\Theta}_{\mathrm{det}}$, using the fact that $\boldsymbol{\mathcal{H}}\Theta^\star_{(\mathrm{loc})} = \Theta^\star$.

To find the upper-bound on γ, we need to make sure that all the above inverse matrices are well-defined. The matrix $(\boldsymbol{I} + \gamma\boldsymbol{\mathcal{G}}\boldsymbol{A})^{-1}$ is well-defined when $\gamma < 1/(L\|\boldsymbol{\mathcal{G}}\|_2)$. Moreover, $\|\boldsymbol{\mathcal{B}}\|_2 \le L\|\boldsymbol{\mathcal{G}}\|_2/(1 - \gamma L\|\boldsymbol{\mathcal{G}}\|_2)$, and $\|\boldsymbol{\mathcal{H}}\|_2 \le L/\mu$, so $(\boldsymbol{I} - \gamma\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{H}})^{-1}$ is well-defined as soon as $\gamma < \mu(1 - \gamma L\|\boldsymbol{\mathcal{G}}\|_2)/(L^2\|\boldsymbol{\mathcal{G}}\|_2)$, or equivalently $\gamma < \dfrac{1}{(1 + L/\mu)L\|\boldsymbol{\mathcal{G}}\|_2}$.

The first-order expansions in γ then follow from $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{G}}\boldsymbol{A} + \mathcal{O}(\gamma)$. □

Proof of Corollary 3.5. The convergence rate is a direct consequence of Lemma 3.1 and Lemma 3.4:

$$\begin{aligned}\|\Theta_t - \Theta^\star\| &\leq \|\Theta_t - \Theta_{\text{det}}\| + \|\Theta_{\text{det}} - \Theta^\star\| \\ &\leq (1-\gamma\mu)^t \|\Theta_0 - \Theta_{\text{det}}\| + \|\Theta_{\text{det}} - \Theta^\star\| \\ &\leq (1-\gamma\mu)^t \|\Theta_0 - \Theta^\star\| + \frac{2\gamma L \Lambda}{\mu}\zeta_\star \ .\end{aligned}$$

This rate holds if $\gamma \leq \min(\frac{1}{L}, \frac{1}{\Lambda L})$. If moreover $\gamma \leq \frac{\mu\epsilon}{L\Lambda\zeta_\star}$, then the above bound becomes

$$\|\Theta_t - \Theta^\star\| \leq (1-\gamma\mu)^T \|\Theta_0 - \Theta^\star\| + \varepsilon \ .$$

We then need to have $(1-\gamma\mu)^T \lesssim \dfrac{\varepsilon}{\|\Theta_0 - \Theta^\star\|}$, or equivalently $T \gtrsim \dfrac{\log(\varepsilon/\|\Theta_0 - \Theta^\star\|)}{\log(1-\gamma\mu)} \gtrsim \dfrac{\log(\|\Theta_0 - \Theta^\star\|/\varepsilon)}{\gamma\mu}$. We conclude using the value on γ. □

Proof of Proposition 3.6. We follow the approach used in the quadratic case. Using Lemma 3.2, a Taylor expansion, Lemma 3.4, and $\theta_k^{\text{det}} = \bar{\theta}^{\text{det}} + \widetilde{\theta}_k^{\text{det}}$, it holds that

$$\begin{aligned} 0 &= \frac{1}{m}\sum_{k=1}^m \nabla f_k(\theta_k^{\text{det}}) \\ &= \frac{1}{m}\sum_{k=1}^m \left(\nabla f_k(\theta^\star) + \nabla^2 f_k(\theta^\star)(\theta_k^{\text{det}} - \theta^\star) + \mathcal{O}(\|\theta_k^{\text{det}} - \theta^\star\|^2)\right) \\ &= \underbrace{\frac{1}{m}\sum_{k=1}^m \nabla f_k(\theta^\star)}_{=0} + \bar{A}\bar{\theta}^{\text{det}} + \frac{1}{m}\sum_{k=1}^m \nabla^2 f_k(\theta^\star)\widetilde{\theta}_k^{\text{det}} - \bar{A}\theta^\star + R_1(\gamma) \ , \end{aligned} \tag{13}$$

where $R_1(\gamma) = \mathcal{O}(\gamma^2)$, and more precisely, using A 1, $\|R_1(\gamma)\| \leq \dfrac{1}{2m}\sum_{k=1}^m K_3\|\theta_c^{\text{det}} - \theta^\star\|^2 = \dfrac{K_3}{2m}\|\Theta_{\text{det}} - \Theta^\star\|^2$. Rearranging (13) yields

$$\bar{\theta}^{\text{det}} = \theta^\star - \bar{A}^{-1}\left(\frac{1}{m}\sum_{k=1}^m \nabla^2 f_k(\theta^\star)\widetilde{\theta}_k^{\text{det}}\right) + \bar{A}^{-1}R_1(\gamma) \ ,$$

which in tensorized form reads:

$$\overline{\Theta}_{\text{det}} = \Theta^\star - \boldsymbol{\mathcal{H}}\widetilde{\Theta}_{\text{det}} + \mathbf{1}\otimes\bar{A}^{-1}R_1(\gamma) \ . \tag{14}$$

Now, using the second part of Lemma 3.2, we have

$$\begin{aligned}\widetilde{\Theta}_{\text{det}} &= -\gamma(\boldsymbol{I} - \boldsymbol{W})^\dagger \boldsymbol{W}\nabla F(\Theta_{\text{det}}) \\ &= -\gamma\boldsymbol{\mathcal{G}}\left(\nabla F(\Theta^\star) + R_2(\gamma)\right) \\ &= -\gamma\boldsymbol{\mathcal{G}}\nabla F(\Theta^\star) - \gamma\boldsymbol{\mathcal{G}}R_2(\gamma) \ ,\end{aligned}$$

where $R_2(\gamma) = \mathcal{O}(\Theta_{\text{det}} - \Theta^\star) = \mathcal{O}(\gamma)$, and more precisely, using A 1, $\|R_2(\gamma)\| \leq L\|\Theta_{\text{det}} - \Theta^\star\|$.

Substituting back into (14), we obtain:

$$\overline{\Theta}_{\text{det}} = \left(\Theta^\star + \gamma\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{G}}\nabla F(\Theta^\star)\right) + \left(\mathbf{1}\otimes\bar{A}^{-1}R_1(\gamma) + \gamma\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{G}}R_2(\gamma)\right) \ ,$$

where the residual term is indeed $\mathcal{O}(\gamma^2)$. More precisely, using Lemma 3.4,

$$\begin{aligned}\|\mathbf{1}\otimes\bar{A}^{-1}R_1(\gamma) + \gamma\boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{G}}R_2(\gamma)\| &\leq \sqrt{m}\|\bar{A}^{-1}\|_2\cdot\|R_1(\gamma)\| + \gamma\|\boldsymbol{\mathcal{H}}\|_2\cdot\|\boldsymbol{\mathcal{G}}\|_2\cdot\|R_2(\gamma)\| \\ &\leq \frac{1}{\mu\sqrt{m}}\cdot\frac{K_3}{2}\|\Theta_{\text{det}} - \Theta^\star\|^2 + \gamma\frac{L}{\mu}\cdot\|\boldsymbol{\mathcal{G}}\|_2\cdot L\|\Theta_{\text{det}} - \Theta^\star\| \\ &\leq \gamma^2\frac{L^2}{\mu^2}\Lambda\Big(\frac{K_3}{2\mu\sqrt{m}}\Lambda\zeta_\star^2 + L\|\boldsymbol{\mathcal{G}}\|_2\cdot\zeta_\star\Big) \ .\end{aligned}$$

□

B DECENTRALIZED STOCHASTIC GRADIENT DESCENT

B.1 Convergence of DSGD

Proof of Proposition 4.1. The proof follows closely the structure of the argument in [Dieuleveut et al., 2020, Proposition 2]. The key difference lies in the contraction between consecutive time steps, which we adapt below to our setting:

$$\begin{aligned}
\mathbb{E}\left[\|\Theta_{t+1}^{(1)}-\Theta_{t+1}^{(2)}\|^2\right] &= \mathbb{E}\left[\|\boldsymbol{W}(\Theta_t^{(1)}-\Theta_t^{(2)})-\gamma\boldsymbol{W}\big(\nabla F(\Theta_t^{(1)};\varepsilon_t)-\nabla F(\Theta_t^{(2)};\varepsilon_t)\big)\|^2\right] \\
&\overset{(i)}{\leq} \mathbb{E}\left[\|(\Theta_t^{(1)}-\Theta_t^{(2)})-\gamma\big(\nabla F(\Theta_t^{(1)};\varepsilon_t)-\nabla F(\Theta_t^{(2)};\varepsilon_t)\big)\|^2\right] \\
&= \mathbb{E}\left[\|\Theta_t^{(1)}-\Theta_t^{(2)}\|^2+\gamma^2\|\nabla F(\Theta_t^{(1)};\varepsilon_t)-\nabla F(\Theta_t^{(2)};\varepsilon_t)\|^2\right] \\
&\quad -2\gamma\mathbb{E}\left[\left\langle\Theta_t^{(1)}-\Theta_t^{(2)},\nabla F(\Theta_t^{(1)};\varepsilon_t)-\nabla F(\Theta_t^{(2)};\varepsilon_t)\right\rangle\right] \qquad (15)\\
&\overset{(ii)}{\leq} \mathbb{E}\left[\|\Theta_t^{(1)}-\Theta_t^{(2)}\|^2-2\gamma(1-L\gamma/2)\left\langle\Theta_t^{(1)}-\Theta_t^{(2)},\nabla F(\Theta_t^{(1)};\varepsilon_t)-\nabla F(\Theta_t^{(2)};\varepsilon_t)\right\rangle\right] \\
&\overset{(iii)}{=} \mathbb{E}\left[\|\Theta_t^{(1)}-\Theta_t^{(2)}\|^2-2\gamma(1-L\gamma/2)\left\langle\Theta_t^{(1)}-\Theta_t^{(2)},\nabla F(\Theta_t^{(1)})-\nabla F(\Theta_t^{(2)})\right\rangle\right] \\
&\overset{(iv)}{\leq} \big(1-2\mu\gamma(1-L\gamma/2)\big)\mathbb{E}\left[\|\Theta_t^{(1)}-\Theta_t^{(2)}\|^2\right] \ , \qquad (16)
\end{aligned}$$

where:

- (i) uses $\|W\|_2=1$, since W is doubly stochastic.
- (ii) comes from the co-coercivity $A\,5(2)$.
- (iii) is obtained by taking the expected value conditional on $\mathcal{F}_t$, using the fact that $\Theta_t^{(1)}$ and $\Theta_t^{(2)}$ are $\mathcal{F}_t$-measurable, and that $\mathbb{E}\left[\nabla F(\Theta_t;\varepsilon_t)\mid\mathcal{F}_t\right]=\nabla F(\Theta_t)$, using $A\,4$.
- (iv) comes from strong convexity $A\,1$ and from $\gamma\leq 2/L$.

For any probability measures ρ_1,ρ_2 with finite second order moment, Villani [2009, Theorem 4.1] shows the existence of random variables $\Theta_0^{(1)}$ and $\Theta_0^{(2)}$ which are independent of $(\varepsilon_t)_{t\geq 1}$ and such that $\mathbf{W}_2^2(\rho_1,\rho_2)=\mathbb{E}\left[\|\Theta_0^{(1)}-\Theta_0^{(2)}\|^2\right]$.

By definition of $\mathbf{W}_2$, and using the fact that the distribution of $\Theta_t^{(1)}$ (resp. $\Theta_t^{(2)}$) is $\rho_1R_\gamma^t$ (resp. $\rho_2R_\gamma^t$), we have $\mathbf{W}_2^2(\rho_1R_\gamma^t,\rho_2R_\gamma^t)\leq\mathbb{E}\left[\|\Theta_t^{(1)}-\Theta_t^{(2)}\|^2\right]$. Then, (16) yields

$$\mathbf{W}_2^2(\rho_1R_\gamma^t,\rho_2R_\gamma^t)\leq\big(1-2\mu\gamma(1-L\gamma/2)\big)^t\,\mathbf{W}_2^2(\rho_1,\rho_2)\ , \qquad (17)$$

and taking $\rho_2=\rho_1R_\gamma$ shows that $\sum_{t=1}^{+\infty}\mathbf{W}_2^2(\rho_1R_\gamma^t,\rho_1R_\gamma^{t+1})\leq\sum_{t=1}^{+\infty}\rho^t\mathbf{W}_2^2(\rho_1,\rho_1R_\gamma)<+\infty$.

Then, since by Villani [2009, Theorem 6.16] the set of probability measures with finite second moment endowed with $\mathbf{W}_2$ is a Polish space, we get that $(\rho_1R_\gamma^t)_{t\geq1}$ is a Cauchy sequence, which converges to some probability measure with finite second moment $\pi_\gamma^{\rho_1}$: $\lim_{t\to+\infty}\mathbf{W}_2(\rho_1R_\gamma^t,\pi_\gamma^{\rho_1})=0$.

Morevoer, if we assume that $\lim_{t\to+\infty}\mathbf{W}_2(\rho_2R_\gamma^t,\pi_\gamma^{\rho_2})=0$ for some $\pi_\gamma^{\rho_2}$, then using (17) we get:

$$\mathbf{W}_2(\pi_\gamma^{\rho_1},\pi_\gamma^{\rho_2})\leq\mathbf{W}_2(\pi_\gamma^{\rho_1},\rho_1R_\gamma^t)+\mathbf{W}_2(\rho_1R_\gamma^t,\rho_2R_\gamma^t)+\mathbf{W}_2(\rho_2R_\gamma^t,\pi_\gamma^{\rho_2})\underset{t\to+\infty}{\longrightarrow}0\ ,$$

so $\pi_\gamma:=\pi_\gamma^{\rho_1}=\pi_\gamma^{\rho_2}$ does not depend on ρ_1. Finally, $\mathbf{W}_2(\pi_\gamma R_\gamma,\pi_\gamma)\leq\mathbf{W}_2(\pi_\gamma R_\gamma,\pi_\gamma R_\gamma^t)+\mathbf{W}_2(\pi_\gamma R_\gamma^t,\pi_\gamma)\underset{t\to+\infty}{\longrightarrow}0$, so $\pi_\gamma R=\pi_\gamma$, and π_γ is indeed a stationary distribution. The uniqueness comes from (17). □

B.2 Variance and Bias of DSGD

Proof of Proposition 4.4. Using

$$\Theta_1 = \boldsymbol{W}\left(\Theta_0 - \gamma\big(\nabla F(\Theta_0) + \varepsilon_1(\Theta_0)\big)\right) \ ,$$

integrating over $\Theta_0 \sim \pi_\gamma$ and taking the expected value on ε_1, it holds that

$$(\boldsymbol{I} - \boldsymbol{W})\Theta_{\text{sto}} = -\gamma \boldsymbol{W} \int_{\mathbb{R}^{md}} \nabla F(\Theta)\, \pi_\gamma(d\Theta) \ .$$

Since the gradient is linear for quadratic functions, we have

$$(\boldsymbol{I} - \boldsymbol{W})\Theta_{\text{sto}} = -\gamma \boldsymbol{W} \nabla F(\textstyle\int_{\mathbb{R}^{md}} \Theta\, \pi_\gamma(d\Theta)) = -\gamma \boldsymbol{W} \nabla F(\Theta_{\text{sto}}) \ .$$

This is precisely the fixed-point equation (3), which has a unique solution: Θ_{det}. Hence $\Theta_{\text{sto}} = \Theta_{\text{det}}$. □

To establish the expression of Θ_{sto} and $\int_{\mathbb{R}^{md}} (\Theta - \Theta_{\text{det}})^{\otimes 2}\, \pi_\gamma(d\Theta)$ in the general case, we need a few technical lemmas.

Note that in many of the lemmas below, we give explicit upper-bounds on residual terms. These upper-bounds are not necessary to prove Proposition 4.2 or Proposition 4.5, and could easily be replaced by $\mathcal{O}$. They are only given here for completeness.

Lemma B.1. *For $\gamma < 1/L$, and under A 1, A 4, and A 5(2):*

$$\mathbb{E}\left[\|\Theta_{t+1} - \Theta_{\text{det}}\|^2 \;\middle|\; \mathcal{F}_t\right] \leq (1 - 2\gamma\mu(1 - L\gamma))\|\Theta_t - \Theta_{\text{det}}\|^2 + 2\tau_2(\Theta_{\text{det}})^2\gamma^2 \ ,$$

where

$$\tau_2(\Theta_{\text{det}}) = \mathbb{E}\left[\|\varepsilon_1(\Theta_{\text{det}})\|^2\right]^{1/2} \ .$$

Moreover

$$\int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\text{det}}\|^2\, \pi_\gamma(d\Theta) \leq \frac{\gamma\tau_2(\Theta_{\text{det}})^2}{\mu(1 - \gamma L)} \ .$$

If $\gamma \leq 1/(\Lambda L)$, then

$$\int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\text{det}}\|^2\, \pi_\gamma(d\Theta) \leq \frac{2\gamma\left(\tau_2^2 + \gamma^2 \frac{4L^4}{\mu^2}\Lambda^2\zeta_\star^2\right)}{\mu(1 - \gamma L)} \ ,$$

where τ_2 is defined in A 5.

Proof. By definition of Θ_{t+1} and Θ_{det} and expanding the square, we have

$$\begin{aligned}
\mathbb{E}\left[\|\Theta_{t+1} - \Theta_{\text{det}}\|^2 \;\middle|\; \mathcal{F}_t\right] &= \mathbb{E}\left[\|\boldsymbol{W}(\Theta_t - \Theta_{\text{det}}) - \gamma\boldsymbol{W}(\nabla F(\Theta_t; \varepsilon_{t+1}) - \nabla F(\Theta_{\text{det}}))\|^2 \;\middle|\; \mathcal{F}_t\right] \\
&\leq \mathbb{E}\left[\|(\Theta_t - \Theta_{\text{det}}) - \gamma(\nabla F(\Theta_t; \varepsilon_{t+1}) - \nabla F(\Theta_{\text{det}}))\|^2 \;\middle|\; \mathcal{F}_t\right] \\
&= \|\Theta_t - \Theta_{\text{det}}\|^2 + \gamma^2\mathbb{E}\left[\|\nabla F(\Theta_t; \varepsilon_{t+1}) - \nabla F(\Theta_{\text{det}})\|^2 \;\middle|\; \mathcal{F}_t\right] \\
&\quad - 2\gamma\mathbb{E}\left[\langle \Theta_t - \Theta_{\text{det}}, \nabla F(\Theta_t; \varepsilon_{t+1}) - \nabla F(\Theta_{\text{det}})\rangle \mid \mathcal{F}_t\right] \ .
\end{aligned}$$

Then, using A 4 and A 5:

$$\begin{aligned}
\mathbb{E}\left[\|\nabla F(\Theta_t; \varepsilon_{t+1}) - \nabla F(\Theta_{\text{det}})\|^2 \;\middle|\; \mathcal{F}_t\right] &\leq 2\mathbb{E}\left[\|\nabla F(\Theta_t; \varepsilon_{t+1}) - \nabla F(\Theta_{\text{det}}; \varepsilon_{t+1})\|^2 \;\middle|\; \mathcal{F}_t\right] + 2\mathbb{E}\left[\|\varepsilon_{t+1}(\Theta_{\text{det}})\|^2 \;\middle|\; \mathcal{F}_t\right] \\
&\leq 2L\mathbb{E}\left[\langle \nabla F(\Theta_t; \varepsilon_{t+1}) - \nabla F(\Theta_{\text{det}}; \varepsilon_{t+1}), \Theta_t - \Theta_{\text{det}}\rangle \mid \mathcal{F}_t\right] + 2\tau_2(\Theta_{\text{det}})^2 \\
&= 2L\left\langle \nabla F(\Theta_t) - \nabla F(\Theta_{\text{det}}), \Theta_t - \Theta_{\text{det}}\right\rangle + 2\tau_2(\Theta_{\text{det}})^2 \ .
\end{aligned}$$

Using A 1, we then obtain

$$\begin{aligned}
\mathbb{E}\left[\|\Theta_{t+1} - \Theta_{\text{det}}\|^2 \;\middle|\; \mathcal{F}_t\right] &\leq \|\Theta_t - \Theta_{\text{det}}\|^2 + 2\gamma^2\tau_2(\Theta_{\text{det}})^2 - 2\gamma(1 - L\gamma)\left\langle \nabla F(\Theta_t) - \nabla F(\Theta_{\text{det}}), \Theta_t - \Theta_{\text{det}}\right\rangle \\
&\leq (1 - 2\gamma\mu(1 - L\gamma))\|\Theta_t - \Theta_{\text{det}}\|^2 + 2\tau_2(\Theta_{\text{det}})^2\gamma^2 \ ,
\end{aligned}$$

hence the first inequality. The second statement is obtained by considering $\mathbb{E}[\|\Theta_t - \Theta_{\mathrm{det}}\|^2] \wedge M$ for $M > 0$, taking the limit as $t \to +\infty$ in the previous inequalities, and letting $M \to +\infty$ using the monotone convergence theorem.

To prove the third inequality, we now upper-bound $\tau_2(\Theta_{\mathrm{det}})^2$. Using A 1 and A 5(2), the noise is almost surely $2L$-Lipschitz continuous, hence

$$\|\varepsilon_t(\Theta_{\mathrm{det}}) - \varepsilon_t(\Theta^\star)\| \leq 2L\|\Theta_{\mathrm{det}} - \Theta^\star\| \ ,$$

from which we deduce

$$\begin{aligned}
\left|\tau_2(\Theta_{\mathrm{det}})^2 - \tau_2(\Theta^\star)^2\right| &= \left|\mathbb{E}\Big[\|\varepsilon_t(\Theta_{\mathrm{det}})\|^2 - \|\varepsilon_t(\Theta^\star)\|^2\Big]\right| \\
&\leq \mathbb{E}\Big[\|\varepsilon_t(\Theta_{\mathrm{det}}) - \varepsilon_t(\Theta^\star)\| \cdot (\|\varepsilon_t(\Theta_{\mathrm{det}})\| + \|\varepsilon_t(\Theta^\star)\|)\Big] \\
&\leq \mathbb{E}\Big[\|\varepsilon_t(\Theta_{\mathrm{det}}) - \varepsilon_t(\Theta^\star)\| \cdot (\|\varepsilon_t(\Theta_{\mathrm{det}}) - \varepsilon_t(\Theta^\star)\| + 2\|\varepsilon_t(\Theta^\star)\|)\Big] \\
&\leq 4L\|\Theta_{\mathrm{det}} - \Theta^\star\|\big(L\|\Theta_{\mathrm{det}} - \Theta^\star\| + \tau_1\big),
\end{aligned}$$

and using $\tau_1 \leq \tau_2$ and Lemma 3.4,

$$\begin{aligned}
\tau_2(\Theta_{\mathrm{det}})^2 &\leq \tau_2^2 + 4L\|\Theta_{\mathrm{det}} - \Theta^\star\|\big(L\|\Theta_{\mathrm{det}} - \Theta^\star\| + \tau_2\big) \\
&= \big(\tau_2 + 2L\|\Theta_{\mathrm{det}} - \Theta^\star\|\big)^2 \\
&\leq \left(\tau_2 + \gamma\frac{2L^2}{\mu}\Lambda\zeta_\star\right)^2 \ .
\end{aligned}$$

□

Lemma B.2. *For $\gamma < 1/(10L)$, and under A 1, A 4, and A 5(4),*

$$\mathbb{E}\left[\|\Theta_{t+1} - \Theta_{\mathrm{det}}\|^4\right]^{1/2} \leq (1 - 2\gamma\mu(1 - 9L\gamma))^t\mathbb{E}\left[\|\Theta_0 - \Theta_{\mathrm{det}}\|^4\right]^{1/2} + \frac{100\gamma\tau_4(\Theta_{\mathrm{det}})^2}{\mu} \ ,$$

where

$$\tau_4(\Theta_{\mathrm{det}}) = \mathbb{E}\left[\|\varepsilon_t(\Theta_{\mathrm{det}})\|^4\right]^{1/4} \ .$$

Moreover

$$\int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\mathrm{det}}\|^4 \, \pi_\gamma(d\Theta) \leq \left(\frac{100\gamma\tau_4(\Theta_{\mathrm{det}})^2}{\mu}\right)^2 \ .$$

If $\gamma \leq 1/(\Lambda L)$ and $\gamma < 1/(10L)$, then

$$\int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\mathrm{det}}\|^4 \, \pi_\gamma(d\Theta) \leq \frac{250 \times 100^2\gamma^2\big(\tau_4^4 + \gamma^4\frac{L^8}{\mu^4}\Lambda^4\zeta_\star^4\big)}{\mu^2}.$$

Proof. Using Cauchy-Schwarz inequality:

$$\begin{aligned}
\mathbb{E}\left[\|\Theta_{t+1} - \Theta_{\mathrm{det}}\|^4 \;\middle|\; \mathcal{F}_t\right] &= \mathbb{E}\left[\|\boldsymbol{W}\big((\Theta_t - \Theta_{\mathrm{det}}) - \gamma(\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}}))\big)\|^4 \;\middle|\; \mathcal{F}_t\right] \\
&\leq \mathbb{E}\left[\|(\Theta_t - \Theta_{\mathrm{det}}) - \gamma(\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}}))\|^4 \;\middle|\; \mathcal{F}_t\right] \\
&= \|\Theta_t - \Theta_{\mathrm{det}}\|^4 + \gamma^4\mathbb{E}\left[\|\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}})\|^4 \;\middle|\; \mathcal{F}_t\right] \\
&+ 4\gamma^2\mathbb{E}\left[\langle\Theta_t - \Theta_{\mathrm{det}}, \nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}})\rangle^2 \;\middle|\; \mathcal{F}_t\right] \\
&+ 2\gamma^2\|\Theta_t - \Theta_{\mathrm{det}}\|^2\mathbb{E}\left[\|\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}})\|^2 \;\middle|\; \mathcal{F}_t\right] \\
&- 4\gamma\|\Theta_t - \Theta_{\mathrm{det}}\|^2\mathbb{E}\left[\langle\Theta_t - \Theta_{\mathrm{det}}, \nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}})\rangle \mid \mathcal{F}_t\right] \\
&- 4\gamma^3\mathbb{E}\left[\|\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}})\|^2 \langle\Theta_t - \Theta_{\mathrm{det}}, \nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}})\rangle \;\middle|\; \mathcal{F}_t\right] \\
\\
&\leq \|\Theta_t - \Theta_{\mathrm{det}}\|^4 - 4\gamma\|\Theta_t - \Theta_{\mathrm{det}}\|^2 \langle\Theta_t - \Theta_{\mathrm{det}}, \nabla F(\Theta_t) - \nabla F(\Theta_{\mathrm{det}})\rangle \\
&+ \gamma^4\mathbb{E}\left[\|\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}})\|^4 \;\middle|\; \mathcal{F}_t\right] \\
&+ 6\gamma^2\|\Theta_t - \Theta_{\mathrm{det}}\|^2\mathbb{E}\left[\|\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}})\|^2 \;\middle|\; \mathcal{F}_t\right] \\
&+ 4\gamma^3\|\Theta_t - \Theta_{\mathrm{det}}\|\mathbb{E}\left[\|\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}})\|^3 \;\middle|\; \mathcal{F}_t\right] \ .
\end{aligned}$$

Then, we use $(x+y)^n \leq 2^{n-1}(x^n+y^n)$ for $n \in \{2,3,4\}$, $\tau_n \leq \tau_4$, and A 5:

$$\begin{aligned}
&\mathbb{E}\left[\|\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}})\|^n \mid \mathcal{F}_t\right] \\
&\leq 2^{n-1}\mathbb{E}\left[\|\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}};\varepsilon_{t+1})\|^n \mid \mathcal{F}_t\right] + 2^{n-1}\mathbb{E}\left[\|\varepsilon_{t+1}(\Theta_{\mathrm{det}})\|^n \mid \mathcal{F}_t\right] \\
&\leq 2^{n-1}L^{n-2}\|\Theta_t - \Theta_{\mathrm{det}}\|^{n-2}\mathbb{E}\left[\|\nabla F(\Theta_t;\varepsilon_{t+1}) - \nabla F(\Theta_{\mathrm{det}};\varepsilon_{t+1})\|^2 \;\middle|\; \mathcal{F}_t\right] + 2^{n-1}\tau_4(\Theta_{\mathrm{det}})^n \\
&\leq 2^{n-1}L^{n-1}\|\Theta_t - \Theta_{\mathrm{det}}\|^{n-2}\left\langle \Theta_t - \Theta_{\mathrm{det}}, \nabla F(\Theta_t) - \nabla F(\Theta_{\mathrm{det}})\right\rangle + 2^{n-1}\tau_4(\Theta_{\mathrm{det}})^n \;\; .
\end{aligned}$$

Plugging this in the previous inequalities:

$$\begin{aligned}
\mathbb{E}\left[\|\Theta_{t+1} - \Theta_{\mathrm{det}}\|^4 \;\middle|\; \mathcal{F}_t\right] \leq{}& \|\Theta_t - \Theta_{\mathrm{det}}\|^4 \\
&- 4\gamma\big(1 - 3L\gamma - 4L^2\gamma^2 - 2L^3\gamma^3\big)\|\Theta_t - \Theta_{\mathrm{det}}\|^2\left\langle \Theta_t - \Theta_{\mathrm{det}}, \nabla F(\Theta_t) - \nabla F(\Theta_{\mathrm{det}})\right\rangle \\
&+ 8\tau_4(\Theta_{\mathrm{det}})^4\gamma^4 + 12\tau_4(\Theta_{\mathrm{det}})^2\gamma^2\|\Theta_t - \Theta_{\mathrm{det}}\|^2 + \underbrace{16\tau_4(\Theta_{\mathrm{det}})^3\gamma^3\|\Theta_t - \Theta_{\mathrm{det}}\|}_{\leq 8\tau_4(\Theta_{\mathrm{det}})^2\gamma^2(\tau_4(\Theta_{\mathrm{det}})^2\gamma^2 + \|\Theta_t - \Theta_{\mathrm{det}}\|^2)} \\
\leq{}& \big(1 - 4\mu\gamma\big(1 - 3L\gamma - 4L^2\gamma^2 - 2L^3\gamma^3\big)\big)\|\Theta_t - \Theta_{\mathrm{det}}\|^4 \\
&+ 16\tau_4(\Theta_{\mathrm{det}})^4\gamma^4 + 20\tau_4(\Theta_{\mathrm{det}})^2\gamma^2\|\Theta_t - \Theta_{\mathrm{det}}\|^2 \;\; .
\end{aligned}$$

Then, we define $c_t = \mathbb{E}\left[\|\Theta_t - \Theta_{\mathrm{det}}\|^4\right]^{1/2}$, and use $\mathbb{E}\left[\|\Theta_t - \Theta_{\mathrm{det}}\|^2\right] \leq c_t$, $1 - 2x \leq (1-x)^2$ and $1 - 2\mu\gamma\big(1 - 3L\gamma - 4L^2\gamma^2 - 2L^3\gamma^3\big) \geq 1 - 2L\gamma \geq 1 - 2/10 \geq 1/2$:

$$\begin{aligned}
c_{t+1}^2 &\leq \big(1 - 4\mu\gamma\big(1 - 3L\gamma - 4L^2\gamma^2 - 2L^3\gamma^3\big)\big)c_t^2 + 16\tau_4(\Theta_{\mathrm{det}})^4\gamma^4 + 20\tau_4(\Theta_{\mathrm{det}})^2\gamma^2 c_t \\
&\leq \big(1 - 2\mu\gamma\big(1 - 3L\gamma - 4L^2\gamma^2 - 2L^3\gamma^3\big)\big)^2 c_t^2 + 16\tau_4(\Theta_{\mathrm{det}})^4\gamma^4 \\
&\qquad + 40\tau_4(\Theta_{\mathrm{det}})^2\gamma^2\big(1 - 2\mu\gamma\big(1 - 3L\gamma - 4L^2\gamma^2 - 2L^3\gamma^3\big)c_t \\
&\leq \big(\big(1 - 2\mu\gamma\big(1 - 3L\gamma - 4L^2\gamma^2 - 2L^3\gamma^3\big)\big)c_t + 20\tau_4(\Theta_{\mathrm{det}})^2\gamma^2\big)^2 \\
&\leq \big(\big(\underbrace{1 - 2\mu\gamma\big(1 - 9L\gamma\big)}_{=r(\gamma)}\big)c_t + 20\tau_4(\Theta_{\mathrm{det}})^2\gamma^2\big)^2 \;\; ,
\end{aligned}$$

i.e., $c_{t+1} \leq r(\gamma)c_t + 20\tau_4(\Theta_{\mathrm{det}})^2\gamma^2$. Therefore $c_t \leq r(\gamma)^t c_0 + \dfrac{10\tau_4(\Theta_{\mathrm{det}})^2\gamma}{\mu(1 - 9L\gamma)} \leq r(\gamma)^t c_0 + \dfrac{100\tau_4(\Theta_{\mathrm{det}})^2\gamma}{\mu}$, which proves the first inequality.

The second inequality is obtained similarly to Lemma B.1.

For the third inequality, we upper-bound $\tau_4(\Theta_{\mathrm{det}})^4$. Using Cauchy-Schwarz inequality,

$$\begin{aligned}
\left|\tau_4(\Theta_{\mathrm{det}})^4 - \tau_4(\Theta^\star)^4\right| &= \left|\mathbb{E}\left[\|\varepsilon_1(\Theta_{\mathrm{det}})\|^4 - \|\varepsilon_1(\Theta^\star)\|^4\right]\right| \\
&\leq \mathbb{E}\Big[\left|\|\varepsilon_1(\Theta_{\mathrm{det}})\|^2 - \|\varepsilon_1(\Theta^\star)\|^2\right| \cdot \left(\|\varepsilon_1(\Theta_{\mathrm{det}})\|^2 + \|\varepsilon_1(\Theta^\star)\|^2\right)\Big] \\
&\leq \sqrt{\mathbb{E}\left[\left(\|\varepsilon_1(\Theta_{\mathrm{det}})\|^2 - \|\varepsilon_1(\Theta^\star)\|^2\right)^2\right]\mathbb{E}\left[\left(\|\varepsilon_1(\Theta_{\mathrm{det}})\|^2 + \|\varepsilon_1(\Theta^\star)\|^2\right)^2\right]} \;\; .
\end{aligned}$$

For $\mathbb{E}\left[\left(\|\varepsilon_1(\Theta_{\mathrm{det}})\|^2 - \|\varepsilon_1(\Theta^\star)\|^2\right)^2\right]$, using A 5, $\forall a, b \geq 0, (a+b)^2 \leq 2(a^2+b^2)$, and the upper-bound on $\tau_2(\Theta_{\mathrm{det}})^2$ established in the proof of Lemma B.1:

$$\begin{aligned}
\mathbb{E}\left[\left(\|\varepsilon_1(\Theta_{\mathrm{det}})\|^2 - \|\varepsilon_1(\Theta^\star)\|^2\right)^2\right] &= \mathbb{E}\left[\Big(\|\varepsilon_1(\Theta_{\mathrm{det}}) - \varepsilon_1(\Theta^\star)\| \cdot \|\varepsilon_1(\Theta_{\mathrm{det}}) + \varepsilon_1(\Theta^\star)\|\Big)^2\right] \\
&\leq 4L^2\|\Theta_{\mathrm{det}} - \Theta^\star\|^2\mathbb{E}\left[\|\varepsilon_1(\Theta_{\mathrm{det}}) + \varepsilon_1(\Theta^\star)\|^2\right] \\
&\leq 8L^2\|\Theta_{\mathrm{det}} - \Theta^\star\|^2\Big(\mathbb{E}\left[\|\varepsilon_1(\Theta_{\mathrm{det}})\|^2\right] + \mathbb{E}\left[\|\varepsilon_1(\Theta^\star)\|^2\right]\Big) \\
&\leq 8L^2\|\Theta_{\mathrm{det}} - \Theta^\star\|^2\Big(\tau_2(\Theta_{\mathrm{det}})^2 + \tau_2^2\Big) \\
&\leq 8L^2\|\Theta_{\mathrm{det}} - \Theta^\star\|^2\Big(\big(\tau_2 + 2L\|\Theta_{\mathrm{det}} - \Theta^\star\|\big)^2 + \tau_2^2\Big) \\
&\leq 2^6 \times L^2\|\Theta_{\mathrm{det}} - \Theta^\star\|^2\big(\tau_2^2 + L^2\|\Theta_{\mathrm{det}} - \Theta^\star\|^2\big) \;\; .
\end{aligned}$$

For $\mathbb{E}\left[\left(\|\varepsilon_1(\Theta_{\det})\|^2+\|\varepsilon_1(\Theta^\star)\|^2\right)^2\right]$, using A 5 and, for $a,b\geq 0$, $(a+b)^2\leq 2(a^2+b^2)$ and $(a+b)^4\leq 8(a^4+b^4)$:

$$\begin{aligned}
\mathbb{E}\left[\left(\|\varepsilon_1(\Theta_{\det})\|^2+\|\varepsilon_1(\Theta^\star)\|^2\right)^2\right] &\leq 2\mathbb{E}\left[\|\varepsilon_1(\Theta_{\det})\|^4+\|\varepsilon_1(\Theta^\star)\|^4\right]\\
&\leq 2\mathbb{E}\left[8\|\varepsilon_1(\Theta_{\det})-\varepsilon_1(\Theta^\star)\|^4+9\|\varepsilon_1(\Theta^\star)\|^4\right]\\
&\leq 2^8L^4\|\Theta_{\det}-\Theta^\star\|^4+18\tau_4^4\\
&\leq 2^8\left(L^4\|\Theta_{\det}-\Theta^\star\|^4+\tau_4^4\right)\ .
\end{aligned}$$

We thus obtain, using, for $a,b\geq 0$, $(a^2+b^2)(a^4+b^4)\leq 2(a^6+b^6)$, $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ and $\tau_2\leq\tau_4$,

$$\begin{aligned}
\left|\tau_4(\Theta_{\det})^4-\tau_4(\Theta^\star)^4\right| &\leq \sqrt{\mathbb{E}\left[\left(\|\varepsilon_1(\Theta_{\det})\|^2-\|\varepsilon_1(\Theta^\star)\|^2\right)^2\right]\mathbb{E}\left[\left(\|\varepsilon_1(\Theta_{\det})\|^2+\|\varepsilon_1(\Theta^\star)\|^2\right)^2\right]}\\
&\leq \sqrt{2^{14}\times L^2\|\Theta_{\det}-\Theta^\star\|^2\left(\tau_2^2+L^2\|\Theta_{\det}-\Theta^\star\|^2\right)\left(L^4\|\Theta_{\det}-\Theta^\star\|^4+\tau_4^4\right)}\\
&\leq \sqrt{2^{15}\times L^2\|\Theta_{\det}-\Theta^\star\|^2\left(\tau_4^6+L^6\|\Theta_{\det}-\Theta^\star\|^6\right)}\\
&\leq 200L\|\Theta_{\det}-\Theta^\star\|\left(\tau_4^3+L^3\|\Theta_{\det}-\Theta^\star\|^3\right)\ .
\end{aligned}$$

Using $\forall a,b\geq 0$, $a^4+200b(a^3+b^3)\leq 151a^4+250b^4\leq 250(a^4+b^4)$, and Lemma 3.4:

$$\begin{aligned}
\tau_4(\Theta_{\det})^4 &\leq \tau_4^4+200L\|\Theta_{\det}-\Theta^\star\|\left(\tau_4^3+L^3\|\Theta_{\det}-\Theta^\star\|^3\right)\\
&\leq 250(\tau_4^4+L^4\|\Theta_{\det}-\Theta^\star\|^4)\\
&\leq 250\left(\tau_4^4+\gamma^4\frac{L^8}{\mu^4}\Lambda^4\zeta_\star^4\right).
\end{aligned}$$

□

Remark B.3. *Using Jensen's inequality,*

$$\begin{aligned}
\int_{\mathbb{R}^{md}}\|\Theta-\Theta_{\det}\|^3\,\pi_\gamma(d\Theta) &\leq \left(\int_{\mathbb{R}^{md}}\|\Theta-\Theta_{\det}\|^4\,\pi_\gamma(d\Theta)\right)^{3/4}\\
&\leq \frac{2^6 100^{3/2}\gamma^{3/2}\left(\tau_4^3+\gamma^3\frac{L^6}{\mu^3}\Lambda^3\zeta_\star^3\right)}{\mu^{3/2}}\ .
\end{aligned}$$

We now establish a first expansion of the stationary distribution's variance, which depends on the topology through $\Theta_{\det}$. This dependence will be removed in later results.

Lemma B.4. *Suppose A 1, A 2, A 4, A 5(4) and A 6 hold. Then the following expansion holds as $\gamma\to 0$:*

$$\begin{aligned}
&\int_{\mathbb{R}^{md}}(\Theta-\Theta_{\det})^{\otimes 2}\,\pi_\gamma(d\Theta)\\
&=\gamma\Big((\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\big(\boldsymbol{I}\otimes\nabla^2F(\Theta_{\det})+\nabla^2F(\Theta_{\det})\otimes\boldsymbol{I}\big)(\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\Big)^\dagger\int_{\mathbb{R}^{md}}\mathbb{E}[\varepsilon_1(\Theta)^{\otimes 2}]\,\pi_\gamma(d\Theta)+\mathcal{O}(\gamma^{3/2})\ .
\end{aligned}$$

Moreover, if $\gamma\leq\min\left(\frac{1}{\Lambda L},\frac{1}{10L}\right)$, then

$$\begin{aligned}
&\left\|\int_{\mathbb{R}^{md}}(\Theta-\Theta_{\det})^{\otimes 2}\,\pi_\gamma(d\Theta)-\gamma\Big((\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\big(\boldsymbol{I}\otimes\nabla^2F(\Theta_{\det})+\nabla^2F(\Theta_{\det})\otimes\boldsymbol{I}\big)(\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\Big)^\dagger\int_{\mathbb{R}^{md}}\mathbb{E}[\varepsilon_1(\Theta)^{\otimes 2}]\,\pi_\gamma(d\Theta)\right\|\\
&\lesssim\gamma^{3/2}\left(\frac{100K_3C^2\tau_4^4}{\mu^3}+\frac{K_3}{8\mu}\right)+\gamma^2\tau_2^2\left(1+\frac{3L^2}{\mu^2}+\frac{1}{1-\lambda_2(W)}\left(1+\frac{L}{\mu}\right)^2\right)\ ,
\end{aligned}$$

where we only kept the lowest orders in γ.

Proof. Starting from the update equation with $t=0$, we use a Taylor expansion:

$$\begin{aligned}\Theta_1 - \Theta_{\text{det}} &= \boldsymbol{W}(\Theta_0 - \Theta_{\text{det}}) + (\boldsymbol{W} - \boldsymbol{I})\Theta_{\text{det}} - \gamma \boldsymbol{W}\big(\nabla F(\Theta_{\text{det}}) + \nabla^2 F(\Theta_{\text{det}})(\Theta_0 - \Theta_{\text{det}}) + \varepsilon_1(\Theta_0) + R(\Theta_0)\big) \\ &= \boldsymbol{W}(\Theta_0 - \Theta_{\text{det}}) - \gamma \boldsymbol{W}\big(\nabla^2 F(\Theta_{\text{det}})(\Theta_0 - \Theta_{\text{det}}) + \varepsilon_1(\Theta_0) + R(\Theta_0)\big) \\ &= \boldsymbol{W}(\boldsymbol{I} - \gamma \nabla^2 F(\Theta_{\text{det}}))(\Theta_0 - \Theta_{\text{det}}) - \gamma \boldsymbol{W}\big(\varepsilon_1(\Theta_0) + R(\Theta_0)\big) \ , \end{aligned} \tag{18}$$

where for each $\Theta \in \mathbb{R}^{md}$,

$$\|R(\Theta)\| \le \frac{K_3}{2}\|\Theta - \Theta_{\text{det}}\|^2 \ , \tag{19}$$

with K_3 introduced in A 1.

We integrate over $\Theta_0 \sim \pi_\gamma$, take the expected value on ε_1 and use Lemma B.1 and Lemma B.2:

$$A_\gamma \int_{\mathbb{R}^{md}} (\Theta - \Theta_{\text{det}})^{\otimes 2}\, \pi_\gamma(d\Theta) = \gamma^2 (\boldsymbol{W} \otimes \boldsymbol{W}) \int_{\mathbb{R}^{md}} \mathbb{E}[\varepsilon_1(\Theta)^{\otimes 2}]\, \pi_\gamma(d\Theta) + R_0(\gamma) \ , \tag{20}$$

where:

- $A_\gamma = \boldsymbol{I} \otimes \boldsymbol{I} - \boldsymbol{W}\big(\boldsymbol{I} - \gamma \nabla^2 F(\Theta_{\text{det}})\big) \otimes \boldsymbol{W}\big(\boldsymbol{I} - \gamma \nabla^2 F(\Theta_{\text{det}})\big)$
- $R_0(\gamma) = \mathcal{O}(\gamma^{5/2})$. More precisely :

$$\begin{aligned} R_0(\gamma) = &-\gamma \int_{\mathbb{R}^{md}} \Big(\boldsymbol{W}(\boldsymbol{I} - \gamma \nabla^2 F(\Theta_{\text{det}}))(\Theta - \Theta_{\text{det}}) \otimes \boldsymbol{W} R(\Theta)\Big)\, \pi_\gamma(d\Theta) \\ &- \gamma \int_{\mathbb{R}^{md}} \Big(\boldsymbol{W} R(\Theta) \otimes \boldsymbol{W}(\boldsymbol{I} - \gamma \nabla^2 F(\Theta_{\text{det}}))(\Theta - \Theta_{\text{det}})\Big)\, \pi_\gamma(d\Theta) \\ &+ \gamma^2 (\boldsymbol{W} \otimes \boldsymbol{W}) \int_{\mathbb{R}^{md}} \big(R(\Theta) \otimes R(\Theta)\big)\, \pi_\gamma(d\Theta) \ . \end{aligned}$$

Hence, using (19) and $\forall a, b \ge 0, ab \le \frac{1}{4}a^4 + \frac{3}{4}b^{4/3}$:

$$\begin{aligned} \|R_0(\gamma)\| &\le \int_{\mathbb{R}^{md}} \big(\gamma^2 \|R(\Theta)\|^2 + 2(1 - \gamma\mu)\gamma \|\Theta - \Theta_{\text{det}}\| \cdot \|R(\Theta)\|\big)\, \pi_\gamma(d\Theta) \\ &\le \int_{\mathbb{R}^{md}} \left(\frac{K_3^2}{4}\gamma^2 \|\Theta - \Theta_{\text{det}}\|^4 + K_3 \gamma \|\Theta - \Theta_{\text{det}}\|^3\right) \pi_\gamma(d\Theta) \\ &\le \int_{\mathbb{R}^{md}} \left(\frac{K_3^2}{4}\gamma^2 \|\Theta - \Theta_{\text{det}}\|^4 + \big(K_3^{1/4}\gamma^{5/8}\big)\big(K_3^{3/4}\gamma^{3/8}\|\Theta - \Theta_{\text{det}}\|^3\big)\right) \pi_\gamma(d\Theta) \\ &\le \int_{\mathbb{R}^{md}} \left(\frac{K_3^2}{4}\gamma^2 \|\Theta - \Theta_{\text{det}}\|^4 + \left(\frac{1}{4}K_3\gamma^{5/2} + \frac{3}{4}K_3\gamma^{1/2}\|\Theta - \Theta_{\text{det}}\|^4\right)\right) \pi_\gamma(d\Theta) \\ &\le \frac{1}{4}\left(K_3^2\gamma^2 + 3K_3\gamma^{1/2}\right) \int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\text{det}}\|^4\, \pi_\gamma(d\Theta) + \frac{1}{4}K_3\gamma^{5/2} \ . \end{aligned}$$

Using Lemma B.2:

$$\|R_0(\gamma)\| \le \gamma^{5/2}\left(\frac{65 K_3 C^2\big(\tau_4^4 + \gamma^4 \frac{L^8}{\mu^4}\Lambda^4 \zeta_\star^4\big)(3 + K_3\gamma^{3/2})}{\mu^2} + \frac{1}{4}K_3\right) \ . \tag{21}$$

Then, we use Woodbury's formula

$$(I - UV)^{-1} = I + U(V^{-1} - U)^{-1} \ ,$$

from which we deduce:

$$\begin{aligned} A_\gamma^{-1} &= \Big(\boldsymbol{I} \otimes \boldsymbol{I} - (\boldsymbol{W} \otimes \boldsymbol{W})\big((\boldsymbol{I} - \gamma \nabla^2 F(\Theta_{\text{det}})) \otimes (\boldsymbol{I} - \gamma \nabla^2 F(\Theta_{\text{det}}))\big)\Big)^{-1} \\ &= \boldsymbol{I} \otimes \boldsymbol{I} + (\boldsymbol{W} \otimes \boldsymbol{W})\Big(\big((\boldsymbol{I} - \gamma \nabla^2 F(\Theta_{\text{det}})) \otimes (\boldsymbol{I} - \gamma \nabla^2 F(\Theta_{\text{det}}))\big)^{-1} - \big(\boldsymbol{W} \otimes \boldsymbol{W}\big)\Big)^{-1} \\ &= \boldsymbol{I} \otimes \boldsymbol{I} + (\boldsymbol{W} \otimes \boldsymbol{W})\Big(\boldsymbol{I} \otimes \boldsymbol{I} - \boldsymbol{W} \otimes \boldsymbol{W} + \gamma\big(\boldsymbol{I} \otimes \nabla^2 F(\Theta_{\text{det}}) + \nabla^2 F(\Theta_{\text{det}}) \otimes \boldsymbol{I}\big) + R_1(\gamma)\Big)^{-1} \\ &= \boldsymbol{I} \otimes \boldsymbol{I} + (\boldsymbol{W} \otimes \boldsymbol{W})(M_\gamma + R_1(\gamma))^{-1} \ , \end{aligned}$$

where:

- $M_\gamma = \boldsymbol{I}\otimes\boldsymbol{I} - \boldsymbol{W}\otimes\boldsymbol{W} + \gamma\big(\boldsymbol{I}\otimes\nabla^2 F(\Theta_{\text{det}}) + \nabla^2 F(\Theta_{\text{det}})\otimes\boldsymbol{I}\big)$,
- $R_1(\gamma) = \mathcal{O}(\gamma^2)$, and more precisely $R_1(\gamma) = \sum_{n=2}^{\infty}\gamma^n\left(\sum_{k=0}^{n}\nabla^2 F(\Theta_{\text{det}})^k\otimes\nabla^2 F(\Theta_{\text{det}})^{n-k}\right)$, so

$$\|R_1(\gamma)\| \le \sum_{n=2}^{\infty}(n+1)(\gamma\|\nabla^2 F(\Theta_{\text{det}})\|)^n \le \sum_{n=2}^{\infty}(n+1)(\gamma L)^n = \frac{(\gamma L)^2(3-2\gamma L)}{(1-\gamma L)^2} \ .$$

If we take $\gamma < \frac{1}{10L}$, then

$$\|R_1(\gamma)\| \le 6(\gamma L)^2 \ . \tag{22}$$

According to Lemma E.1,

$$M_\gamma^{-1} = \frac{1}{\gamma}\Big((\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\big(\boldsymbol{I}\otimes\nabla^2 F(\Theta_{\text{det}}) + \nabla^2 F(\Theta_{\text{det}})\otimes\boldsymbol{I}\big)(\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\Big)^\dagger + R_2(\gamma) \ , \tag{23}$$

where:

- $R_2(\gamma) = \mathcal{O}(1)$, and more precisely

$$\|R_2(\gamma)\| \le \frac{1}{1-\lambda_2(W)}\Big(1+\frac{L}{\mu}\Big)^2 \ . \tag{24}$$

- $\boldsymbol{\mathcal{P}} = \frac{1}{m}\underbrace{\boldsymbol{1}\boldsymbol{1}^\top}_{\in\mathbb{R}^{m\times m}}\otimes\underbrace{\mathrm{Id}}_{\in\mathbb{R}^{d\times d}}$ is the matrix of the orthogonal projection onto the kernel of $\boldsymbol{I}-\boldsymbol{W}$.
- $\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}}$ is the matrix of the orthogonal projection onto the kernel of $\boldsymbol{I}\otimes\boldsymbol{I}-\boldsymbol{W}\otimes\boldsymbol{W}$.

Therefore:

$$\begin{aligned}
A_\gamma^{-1} &= \boldsymbol{I}\otimes\boldsymbol{I} + (\boldsymbol{W}\otimes\boldsymbol{W})(M_\gamma + R_1(\gamma))^{-1}\\
&= (\boldsymbol{W}\otimes\boldsymbol{W})M_\gamma^{-1} + (\boldsymbol{I}\otimes\boldsymbol{I} + R_3(\gamma))\\
&= \frac{1}{\gamma}(\boldsymbol{W}\otimes\boldsymbol{W})\Big((\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\big(\boldsymbol{I}\otimes\nabla^2 F(\Theta_{\text{det}}) + \nabla^2 F(\Theta_{\text{det}})\otimes\boldsymbol{I}\big)(\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\Big)^\dagger\\
&+ \big(\boldsymbol{I}\otimes\boldsymbol{I} + R_3(\gamma) + (\boldsymbol{W}\otimes\boldsymbol{W})R_2(\gamma)\big)\\
&= \frac{1}{\gamma}\Big((\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\big(\boldsymbol{I}\otimes\nabla^2 F(\Theta_{\text{det}}) + \nabla^2 F(\Theta_{\text{det}})\otimes\boldsymbol{I}\big)(\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\Big)^\dagger\\
&+ \big(\boldsymbol{I}\otimes\boldsymbol{I} + R_3(\gamma) + (\boldsymbol{W}\otimes\boldsymbol{W})R_2(\gamma)\big) \ ,
\end{aligned}$$

where, according to Lemma E.2, $\|R_3(\gamma)\| \le \|\boldsymbol{W}\otimes\boldsymbol{W}\|_2 \times \|M_\gamma^{-1}\|_2^2\cdot\|R_1(\gamma)\| = \|M_\gamma^{-1}\|_2^2\cdot\|R_1(\gamma)\|$.

Using (23) and (24), $\|M_\gamma^{-1}\|_2 \le \frac{1}{2\gamma\mu} + \frac{1}{1-\lambda_2(W)}\Big(1+\frac{L}{\mu}\Big)^2$, so $\|M_\gamma^{-1}\|_2^2 \le \frac{1}{2\gamma^2\mu^2} + \frac{2}{(1-\lambda_2(W))^2}\Big(1+\frac{L}{\mu}\Big)^4$, from which we deduce, with (22):

$$\|R_3(\gamma)\| \le \frac{3L^2}{\mu^2} + \frac{12\gamma^2L^2}{(1-\lambda_2(W))^2}\Big(1+\frac{L}{\mu}\Big)^4 \ . \tag{25}$$

We then have

$$A_\gamma^{-1} = \frac{1}{\gamma}\Big((\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\big(\boldsymbol{I}\otimes\nabla^2 F(\Theta_{\text{det}}) + \nabla^2 F(\Theta_{\text{det}})\otimes\boldsymbol{I}\big)(\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\Big)^\dagger + R_4(\gamma) \ , \tag{26}$$

with

$$\|R_4(\gamma)\|_2 \leq 1 + \frac{3L^2}{\mu^2} + \frac{12\gamma^2 L^2}{(1-\lambda_2(W))^2}\Big(1+\frac{L}{\mu}\Big)^4 + \frac{1}{1-\lambda_2(W)}\Big(1+\frac{L}{\mu}\Big)^2 \ . \tag{27}$$

Plugging this back in (20), we get

$$\int_{\mathbb{R}^{md}} (\Theta-\Theta_{\mathrm{det}})^{\otimes 2}\, \pi_\gamma(d\Theta) = \gamma\Big((\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\big(\boldsymbol{I}\otimes\nabla^2 F(\Theta_{\mathrm{det}}) + \nabla^2 F(\Theta_{\mathrm{det}})\otimes\boldsymbol{I}\big)(\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\Big)^\dagger \int_{\mathbb{R}^{md}} \mathbb{E}[\varepsilon_1(\Theta)^{\otimes 2}]\,\pi_\gamma(d\Theta)$$
$$+ \gamma^2 R_4(\boldsymbol{W}\otimes\boldsymbol{W})\int_{\mathbb{R}^{md}} \mathbb{E}[\varepsilon_1(\Theta)^{\otimes 2}]\,\pi_\gamma(d\Theta) + A_\gamma^{-1}R_0 \ .$$

Using (27), A 4, (26) and (21):

$$\begin{aligned}
&\Big\|\gamma^2 R_4(\boldsymbol{W}\otimes\boldsymbol{W})\int_{\mathbb{R}^{md}} \mathbb{E}[\varepsilon_1(\Theta)^{\otimes 2}]\,\pi_\gamma(d\Theta) + A_\gamma^{-1}R_0\Big\| \\
&\leq \gamma^2\tau_2^2\left(1 + \frac{3L^2}{\mu^2} + \frac{12\gamma^2 L^2}{(1-\lambda_2(W))^2}\Big(1+\frac{L}{\mu}\Big)^4 + \frac{1}{1-\lambda_2(W)}\Big(1+\frac{L}{\mu}\Big)^2\right) \\
&+ \gamma^{5/2}\left(\frac{1}{2\gamma\mu} + 1 + \frac{3L^2}{\mu^2} + \frac{12\gamma^2 L^2}{(1-\lambda_2(W))^2}\Big(1+\frac{L}{\mu}\Big)^4 + \frac{1}{1-\lambda_2(W)}\Big(1+\frac{L}{\mu}\Big)^2\right) \\
&\times\left(\frac{65K_3C^2\big(\tau_4^4 + \gamma^4\frac{L^8}{\mu^4}\Lambda^4\zeta_\star^4\big)(3+K_3\gamma^{3/2})}{\mu^2} + \frac{1}{4}K_3\right) \ .
\end{aligned}$$

□

Using the previous lemma, we can now give a first expansion of the stationary distribution's bias.

Lemma B.5. *Suppose A 1, A 2, A 4, A 5(4) and A 6 hold. Also assume that for each k, f_k's fourth derivative is bounded: there exists $K_4 > 0$ such that for $\theta, u \in \mathbb{R}^d$, $\|\nabla^4 f_k(\theta)u^{\otimes 3}\| \leq K_4\|u\|^3$. Then the following expansion holds as $\gamma \to 0$:*

$$\Theta_{\mathrm{sto}} - \Theta_{\mathrm{det}} = -\frac{\gamma}{2}\big(\boldsymbol{\mathcal{P}}\nabla^2 F(\Theta_{\mathrm{det}})\boldsymbol{\mathcal{P}}\big)^\dagger \nabla^3 F(\Theta_{\mathrm{det}})\int_{\mathbb{R}^{md}} (\Theta-\Theta_{\mathrm{det}})^{\otimes 2}\,\pi_\gamma(d\Theta) + \mathcal{O}(\gamma^{3/2}) \ .$$

Moreover, if $\gamma \leq \min\left(\frac{1}{\Lambda L}, \frac{1}{10L}\right)$, then

$$\begin{aligned}
&\Big\|\Theta_{\mathrm{sto}} - \Theta_{\mathrm{det}} + \frac{\gamma}{2}\big(\boldsymbol{\mathcal{P}}\nabla^2 F(\Theta_{\mathrm{det}})\boldsymbol{\mathcal{P}}\big)^\dagger \nabla^3 F(\Theta_{\mathrm{det}})\int_{\mathbb{R}^{md}} (\Theta-\Theta_{\mathrm{det}})^{\otimes 2}\,\pi_\gamma(d\Theta)\Big\| \\
&\lesssim \gamma^{3/2}\frac{20C^{3/2}\tau_4^3K_4}{\mu^{5/2}} + \gamma^2\left(1+\frac{4L^2}{\mu^2}\right)\frac{2K_3\tau_2^2}{\mu} \ ,
\end{aligned}$$

where we only kept the lowest orders in γ.

Proof of Lemma B.5. Using the update equation with $t = 0$, integrating over $\Theta_0 \sim \pi_\gamma$ and taking the expected value on ε_1, it holds that

$$(\boldsymbol{I}-\boldsymbol{W})\Theta_{\mathrm{sto}} = -\gamma\boldsymbol{W}\int_{\mathbb{R}^{md}} \nabla F(\Theta)\,\pi_\gamma(d\Theta) \ .$$

Then, with a Taylor expansion of $\nabla F(\Theta)$ around Θ_{det}:

$$\nabla F(\Theta) = \nabla F(\Theta_{\mathrm{det}}) + \nabla^2 F(\Theta_{\mathrm{det}})(\Theta-\Theta_{\mathrm{det}}) + \frac{1}{2}\nabla^3 F(\Theta_{\mathrm{det}})(\Theta-\Theta_{\mathrm{det}})^{\otimes 2} + R_1(\Theta) \ ,$$

where $\forall \Theta \in \mathbb{R}^{md}$, $\|R_1(\Theta)\| \leq \frac{1}{6}K_4\|\Theta - \Theta_{\text{det}}\|^3$. Thus:

$$\begin{aligned}
&(\boldsymbol{I} - \boldsymbol{W})\Theta_{\text{sto}} \\
&= -\gamma \boldsymbol{W}\nabla F(\Theta_{\text{det}}) - \gamma \boldsymbol{W}\nabla^2 F(\Theta_{\text{det}})(\Theta_{\text{sto}} - \Theta_{\text{det}}) - \frac{\gamma}{2}\boldsymbol{W}\nabla^3 F(\Theta_{\text{det}}) \int_{\mathbb{R}^{md}} (\Theta - \Theta_{\text{det}})^{\otimes 2}\, \pi_\gamma(d\Theta) \\
&- \gamma \boldsymbol{W} \int_{\mathbb{R}^{md}} R_1(\Theta)\, \pi_\gamma(d\Theta) \ \ .
\end{aligned}$$

Since $(\boldsymbol{I} - \boldsymbol{W})\Theta_{\text{det}} = -\gamma \boldsymbol{W}\nabla F(\Theta_{\text{det}})$, we obtain

$$\begin{aligned}
&(\boldsymbol{I} - \boldsymbol{W} + \gamma \boldsymbol{W}\nabla^2 F(\Theta_{\text{det}}))(\Theta_{\text{sto}} - \Theta_{\text{det}}) \\
&= -\frac{\gamma}{2}\boldsymbol{W}\nabla^3 F(\Theta_{\text{det}}) \int_{\mathbb{R}^{md}} (\Theta - \Theta_{\text{det}})^{\otimes 2}\, \pi_\gamma(d\Theta) - \gamma \boldsymbol{W} \int_{\mathbb{R}^{md}} R_1(\Theta)\, \pi_\gamma(d\Theta) \ \ .
\end{aligned}$$

Using Woodbury's formula:

$$(I - UV)^{-1} = I + U(V^{-1} - U)^{-1} \ \ ,$$

and Lemma E.1:

$$\begin{aligned}
(\boldsymbol{I} - \boldsymbol{W} + \gamma \boldsymbol{W}\nabla^2 F(\Theta_{\text{det}}))^{-1} &= (\boldsymbol{I} - \boldsymbol{W}(\boldsymbol{I} - \gamma\nabla^2 F(\Theta_{\text{det}})))^{-1} \\
&= \boldsymbol{I} + \boldsymbol{W}\big((\boldsymbol{I} - \gamma\nabla^2 F(\Theta_{\text{det}}))^{-1} - \boldsymbol{W}\big)^{-1} \\
&= \boldsymbol{I} + \boldsymbol{W}\big(\boldsymbol{I} - \boldsymbol{W} + \gamma\nabla^2 F(\Theta_{\text{det}}) + R_2(\gamma)\big)^{-1} \\
&= \frac{1}{\gamma}\boldsymbol{W}\big(\boldsymbol{\mathcal{P}}\nabla^2 F(\Theta_{\text{det}})\boldsymbol{\mathcal{P}}\big)^{\dagger} + R_3(\gamma) \\
&= \frac{1}{\gamma}\big(\boldsymbol{\mathcal{P}}\nabla^2 F(\Theta_{\text{det}})\boldsymbol{\mathcal{P}}\big)^{\dagger} + R_3(\gamma) \ \ ,
\end{aligned}$$

where

- $R_2(\gamma) = \sum_{k=2}^{+\infty}(\gamma\nabla^2 F(\Theta_{\text{det}}))^k$ satisfies $\|R_2(\gamma)\|_2 \leq \frac{\gamma^2 L^2}{1 - \gamma L} \leq 2\gamma^2 L^2$ if $\gamma < \frac{1}{10L}$.
- According to Lemma E.2, $\|R_3(\gamma)\|_2 \leq 1 + \|(\boldsymbol{I} - \boldsymbol{W} + \gamma\nabla^2 F(\Theta_{\text{det}}))^{-1}\|_2^2 \cdot \|R_2(\gamma)\|_2$.
- According to Lemma E.1, $\|(\boldsymbol{I} - \boldsymbol{W} + \gamma\nabla^2 F(\Theta_{\text{det}}))^{-1}\|_2 \leq \frac{1}{\gamma\mu} + \frac{1}{1 - \lambda_2(W)}\Big(1 + \frac{L}{\mu}\Big)^2$.

Combining these inequalities yields

$$\|R_3(\gamma)\|_2 \leq 1 + \frac{4L^2}{\mu^2} + \gamma^2\frac{4L^2}{(1 - \lambda_2(W))^2}\Big(1 + \frac{L}{\mu}\Big)^4 \ \ .$$

so

$$\begin{aligned}
\Theta_{\text{sto}} - \Theta_{\text{det}} &= -\frac{\gamma}{2}(\boldsymbol{I} - \boldsymbol{W} + \gamma \boldsymbol{W}\nabla^2 F(\Theta_{\text{det}}))^{-1}\boldsymbol{W}\nabla^3 F(\Theta_{\text{det}}) \int_{\mathbb{R}^{md}} (\Theta - \Theta_{\text{det}})^{\otimes 2}\, \pi_\gamma(d\Theta) \\
&\quad - \gamma(\boldsymbol{I} - \boldsymbol{W} + \gamma \boldsymbol{W}\nabla^2 F(\Theta_{\text{det}}))^{-1}\boldsymbol{W} \int_{\mathbb{R}^{md}} R_1(\Theta)\, \pi_\gamma(d\Theta) \\
&= -\frac{1}{2}\big(\boldsymbol{\mathcal{P}}\nabla^2 F(\Theta_{\text{det}})\boldsymbol{\mathcal{P}}\big)^{\dagger}\nabla^3 F(\Theta_{\text{det}}) \int_{\mathbb{R}^{md}} (\Theta - \Theta_{\text{det}})^{\otimes 2}\, \pi_\gamma(d\Theta) + R_4(\gamma) \ \ ,
\end{aligned}$$

where

$$\begin{aligned}
R_4(\gamma) = &-\frac{\gamma}{2}R_3(\gamma)\boldsymbol{W}\nabla^3 F(\Theta_{\text{det}}) \int_{\mathbb{R}^{md}} (\Theta - \Theta_{\text{det}})^{\otimes 2}\, \pi_\gamma(d\Theta) \\
&- \gamma(\boldsymbol{I} - \boldsymbol{W} + \gamma \boldsymbol{W}\nabla^2 F(\Theta_{\text{det}}))^{-1}\boldsymbol{W} \int_{\mathbb{R}^{md}} R_1(\Theta)\, \pi_\gamma(d\Theta) \ \ .
\end{aligned}$$

Using the upper-bounds we established previously, as well as A 1, Lemma B.1 and Remark B.3:

$$\|R_4(\gamma)\|_2 \leq \gamma^2 K_3 \left(1 + \frac{4L^2}{\mu^2} + \gamma^2 \frac{4L^2}{(1-\lambda_2(W))^2}\left(1+\frac{L}{\mu}\right)^4\right) \frac{\left(\tau_2 + \gamma \frac{2L^2}{\mu}\Lambda\zeta_\star\right)^2}{\mu}$$
$$+ \gamma^{5/2}\frac{K_4}{6}\left(\frac{1}{\gamma\mu} + 1 + \frac{4L^2}{\mu^2} + \gamma^2\frac{4L^2}{(1-\lambda_2(W))^2}\left(1+\frac{L}{\mu}\right)^4\right)\frac{2^6 C^{3/2}\left(\tau_4^3 + \gamma^3 \frac{L^6}{\mu^3}\Lambda^3\zeta_\star^3\right)}{\mu^{3/2}} \ .$$

□

The two previous lemmas use the covariance of the noise under the stationary distribution, for which we give an expansion in terms of the optimum $\Theta^\star$ in the next two lemmas.

Lemma B.6. *Under the assumptions of Proposition 4.2, and if all local functions have a bounded fourth derivative as in Lemma B.5, then it holds that*

$$\int_{\mathbb{R}^{md}} \boldsymbol{\mathcal{C}}(\Theta)\, \pi_\gamma(d\Theta) = \boldsymbol{\mathcal{C}}(\Theta_{\det}) + \mathcal{O}(\gamma) \ .$$

Moreover, if we assume that there exist $\sigma_\star^2, \beta \geq 0$ such that for any Θ,

$$\|\nabla \boldsymbol{\mathcal{C}}(\Theta)\| \leq \sigma_\star^2 + \beta\|\Theta - \Theta^\star\|^2, \qquad \|\nabla^2 \boldsymbol{\mathcal{C}}(\Theta)\| \leq \sigma_\star^2 + \beta\|\Theta - \Theta^\star\|^2 \ ,$$

and if $\gamma \leq \min\left(\frac{1}{\Lambda L}, \frac{1}{10L}\right)$, then

$$\left\| \int_{\mathbb{R}^{md}} \boldsymbol{\mathcal{C}}(\Theta)\, \pi_\gamma(d\Theta) - \boldsymbol{\mathcal{C}}(\Theta_{\det})\right\| \lesssim \gamma \frac{4\sigma_\star^2 \tau_2}{\mu} \ .$$

Proof. We use a Taylor expansion

$$\boldsymbol{\mathcal{C}}(\Theta) = \boldsymbol{\mathcal{C}}(\Theta_{\det}) + \nabla\boldsymbol{\mathcal{C}}(\Theta_{\det})(\Theta - \Theta_{\det}) + R(\Theta) \ ,$$

where

$$\begin{aligned}
\|R(\Theta)\| &\leq \left(\sigma_\star^2 + \beta\left(\|\Theta_{\det} - \Theta^\star\|^2 + \|\Theta - \Theta^\star\|^2\right)\right)\|\Theta - \Theta_{\det}\|^2 \\
&= \left(\sigma_\star^2 + \beta\|\Theta_{\det} - \Theta^\star\|^2\right)\|\Theta - \Theta_{\det}\|^2 + \beta\|\Theta - \Theta^\star\|^2\|\Theta - \Theta_{\det}\|^2 \\
&\leq \left(\sigma_\star^2 + \beta\|\Theta_{\det} - \Theta^\star\|^2\right)\|\Theta - \Theta_{\det}\|^2 + 2\beta\|\Theta - \Theta_{\det}\|^4 + 2\beta\|\Theta_{\det} - \Theta^\star\|^2\|\Theta - \Theta_{\det}\|^2 \\
&\leq \left(\sigma_\star^2 + 3\beta\|\Theta_{\det} - \Theta^\star\|^2\right)\|\Theta - \Theta_{\det}\|^2 + 2\beta\|\Theta - \Theta_{\det}\|^4 \ .
\end{aligned}$$

We then have

$$\int_{\mathbb{R}^{md}} \boldsymbol{\mathcal{C}}(\Theta)\, \pi_\gamma(d\Theta) = \boldsymbol{\mathcal{C}}(\Theta_{\det}) + \nabla\boldsymbol{\mathcal{C}}(\Theta_{\det})(\Theta_{\text{sto}} - \Theta_{\det}) + \int_{\mathbb{R}^{md}} R(\Theta)\, \pi_\gamma(d\Theta) \ .$$

Using A 6, Lemma B.5, Lemma B.1 and Lemma B.2:

$$\int_{\mathbb{R}^{md}} \boldsymbol{\mathcal{C}}(\Theta)\, \pi_\gamma(d\Theta) = \boldsymbol{\mathcal{C}}(\Theta_{\det}) + \mathcal{O}(\gamma) \ .$$

More precisely,

$$\begin{aligned}
\left\|\nabla\boldsymbol{\mathcal{C}}(\Theta_{\det})(\Theta_{\text{sto}} - \Theta_{\det}) + \int_{\mathbb{R}^{md}} R(\Theta)\, \pi_\gamma(d\Theta)\right\| &\leq \|\nabla\boldsymbol{\mathcal{C}}(\Theta_{\det})\| \cdot \|\Theta_{\text{sto}} - \Theta_{\det}\| + \int_{\mathbb{R}^{md}} \|R(\Theta)\|\, \pi_\gamma(d\Theta) \\
&\leq \left(\sigma_\star^2 + \beta\|\Theta_{\det} - \Theta^\star\|^2\right) \cdot \|\Theta_{\text{sto}} - \Theta_{\det}\| \\
&\quad + \left(\sigma_\star^2 + 3\beta\|\Theta_{\det} - \Theta^\star\|^2\right)\int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\det}\|^2\, \pi_\gamma(d\Theta) \\
&\quad + 2\beta\int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\det}\|^4\, \pi_\gamma(d\Theta) \ .
\end{aligned}$$

Using Lemma B.5, Lemma B.1 and Lemma B.2, if $\gamma \leq \min\left(\frac{1}{\Lambda L}, \frac{1}{10L}\right)$:

$$
\begin{aligned}
&\left\|\nabla \mathcal{C}(\Theta_{\mathrm{det}})(\Theta_{\mathrm{sto}} - \Theta_{\mathrm{det}}) + \int_{\mathbb{R}^{md}} R(\Theta)\, \pi_\gamma(d\Theta)\right\| \\
&\leq \left(\sigma_\star^2 + \beta\|\Theta_{\mathrm{det}} - \Theta^\star\|^2\right) \cdot \|\Theta_{\mathrm{sto}} - \Theta_{\mathrm{det}}\| + \left(\sigma_\star^2 + 3\beta\|\Theta_{\mathrm{det}} - \Theta^\star\|^2\right) \int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\mathrm{det}}\|^2\, \pi_\gamma(d\Theta) \\
&\quad + 2\beta \int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\mathrm{det}}\|^4\, \pi_\gamma(d\Theta) \\
&\leq \left(\sigma_\star^2 + 3\beta\|\Theta_{\mathrm{det}} - \Theta^\star\|^2\right) \cdot \left(\|\Theta_{\mathrm{sto}} - \Theta_{\mathrm{det}}\| + \int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\mathrm{det}}\|^2\, \pi_\gamma(d\Theta)\right) + 2\beta \int_{\mathbb{R}^{md}} \|\Theta - \Theta_{\mathrm{det}}\|^4\, \pi_\gamma(d\Theta) \\
&\leq \gamma\left(\sigma_\star^2 + 3\beta\gamma^2 \frac{L^2}{\mu^2}\Lambda^2\zeta_\star^2\right) \cdot \left(\left(1 + \frac{\gamma K_3}{2\mu}\right) \frac{4\left(\tau_2^2 + \gamma^2 \frac{4L^4}{\mu^2}\Lambda^2\zeta_\star^2\right)}{\mu}\right. \\
&+ \qquad \gamma^2 K_3 \left(1 + \frac{4L^2}{\mu^2} + \gamma^2 \frac{4L^2}{(1-\lambda_2(W))^2}\left(1 + \frac{L}{\mu}\right)^4\right) \frac{\left(\tau_2 + \gamma \frac{2L^2}{\mu}\Lambda\zeta_\star\right)^2}{\mu} \\
&+ \qquad \left.\gamma^{5/2} \frac{K_4}{6}\left(\frac{1}{\gamma\mu} + 1 + \frac{4L^2}{\mu^2} + \gamma^2 \frac{4L^2}{(1-\lambda_2(W))^2}\left(1 + \frac{L}{\mu}\right)^4\right) \frac{2^6 C^{3/2}\left(\tau_4^3 + \gamma^3 \frac{L^6}{\mu^3}\Lambda^3\zeta_\star^3\right)}{\mu^{3/2}}\right) \\
&+ \qquad \gamma^2 \frac{500 C^2 \beta\left(\tau_4^4 + \gamma^4 \frac{L^8}{\mu^4}\Lambda^4\zeta_\star^4\right)}{\mu^2} \ .
\end{aligned}
$$

□

Lemma B.7. *Under the assumptions of Lemma B.6, it holds that*

$$
\int_{\mathbb{R}^{md}} \mathcal{C}(\Theta)\, \pi_\gamma(d\Theta) = \mathcal{C}(\Theta^\star) + \mathcal{O}(\gamma) \ .
$$

Moreover, if we assume, like in Lemma B.6, that the first two derivatives of $\mathcal{C}$ have polynomial growth, then for $\gamma \leq \max\left(\frac{1}{\Lambda L}, \frac{1}{10L}\right)$,

$$
\begin{aligned}
&\left\|\int_{\mathbb{R}^{md}} \mathcal{C}(\Theta)\, \pi_\gamma(d\Theta) - \mathcal{C}(\Theta^\star)\right\| \\
&\leq \gamma \frac{L}{\mu}\Lambda\zeta_\star \cdot \left(\sigma_\star^2 + \gamma^2 \frac{\beta L^2}{2\mu^2}\Lambda^2\zeta_\star^2\right) \\
&+ \quad \gamma\left(\sigma_\star^2 + 3\beta\gamma^2 \frac{L^2}{\mu^2}\Lambda^2\zeta_\star^2\right) \cdot \left(\left(1 + \frac{\gamma K_3}{2\mu}\right) \frac{4\left(\tau_2^2 + \gamma^2 \frac{4L^4}{\mu^2}\Lambda^2\zeta_\star^2\right)}{\mu}\right. \\
&+ \qquad \gamma^2 K_3 \left(1 + \frac{4L^2}{\mu^2} + \gamma^2 \frac{4L^2}{(1-\lambda_2(W))^2}\left(1 + \frac{L}{\mu}\right)^4\right) \frac{\left(\tau_2 + \gamma \frac{2L^2}{\mu}\Lambda\zeta_\star\right)^2}{\mu} \\
&+ \qquad \left.\gamma^{5/2}\left(\frac{1}{\gamma\mu} + 1 + \frac{4L^2}{\mu^2} + \gamma^2 \frac{4L^2}{(1-\lambda_2(W))^2}\left(1 + \frac{L}{\mu}\right)^4\right) \frac{2^6 \cdot 100^{3/2} K_4\left(\tau_4^3 + \gamma^3 \frac{L^6}{\mu^3}\Lambda^3\zeta_\star^3\right)}{6\mu^{3/2}}\right) \\
&+ \quad \gamma^2 \frac{500 \cdot 100^2 \beta\left(\tau_4^4 + \gamma^4 \frac{L^8}{\mu^4}\Lambda^4\zeta_\star^4\right)}{\mu^2} \ .
\end{aligned}
$$

Proof. Write

$$
\left\|\int_{\mathbb{R}^{md}} \mathcal{C}(\Theta)\, \pi_\gamma(d\Theta) - \mathcal{C}(\Theta^\star)\right\| \leq \left\|\int_{\mathbb{R}^{md}} \mathcal{C}(\Theta)\, \pi_\gamma(d\Theta) - \mathcal{C}(\Theta_{\mathrm{det}})\right\| + \left\|\mathcal{C}(\Theta_{\mathrm{det}}) - \mathcal{C}(\Theta^\star)\right\| \ ,
$$

upper-bound the first term using Lemma B.6, and the second one using Lemma 3.4:

$$\begin{aligned}\left\|\boldsymbol{\mathcal{C}}(\Theta_{\mathrm{det}})-\boldsymbol{\mathcal{C}}(\Theta^\star)\right\| &\leq \|\Theta_{\mathrm{det}}-\Theta^\star\|\cdot\left(\sigma_\star^2+\frac{\beta}{2}\|\Theta_{\mathrm{det}}-\Theta^\star\|^2\right)\\ &\leq \gamma\frac{L}{\mu}\Lambda\zeta_\star\cdot\left(\sigma_\star^2+\gamma^2\frac{\beta L^2}{2\mu^2}\Lambda^2\zeta_\star^2\right) \ .\end{aligned}$$

□

The following corollary gives an expansion of the stationary distribution's variance at the first order in γ, without any dependence in the network topology at the first order.

Corollary B.8. *Suppose A 1, A 2, A 4, A 5(4) and A 6 hold. Then the following expansion holds as $\gamma \to 0$:*

$$\int_{\mathbb{R}^{md}}(\Theta-\Theta_{\mathrm{det}})^{\otimes 2}\,\pi_\gamma(d\Theta)=\gamma\Big((\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\big(\boldsymbol{I}\otimes\nabla^2F(\Theta^\star)+\nabla^2F(\Theta^\star)\otimes\boldsymbol{I}\big)(\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\Big)^\dagger\boldsymbol{\mathcal{C}}(\Theta^\star)+\mathcal{O}(\gamma^{3/2}) \ .$$

Proof. This is easily obtained from Lemma B.4 and Lemma B.7 using A 1.

Note that $A\mapsto A^\dagger$ is not continuous, but that $A\mapsto((\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}})\,A\,(\boldsymbol{\mathcal{P}}\otimes\boldsymbol{\mathcal{P}}))^\dagger$ is continuous at every symmetric positive definite A. □

The previous result is rewritten in a more readable in the following lemma.

Lemma B.9. *Under the same assumptions as Lemma B.4, the first-order term of the quadratic error can be expressed, for any $k,\ell\in\{1,\ldots,m\}$, as*

$$\int_{\mathbb{R}^{md}}(\theta_k-\theta_k^{\mathrm{det}})\otimes(\theta_\ell-\theta_\ell^{\mathrm{det}})\,\pi_\gamma(d\Theta)=\frac{\gamma}{m}\big(\boldsymbol{I}\otimes\bar{A}+\bar{A}\otimes\boldsymbol{I}\big)^{-1}\mathcal{C}(\theta^\star)+\mathcal{O}(\gamma^{3/2}) \ .$$

Proof. We simply need to simplify $\Big(\boldsymbol{\mathcal{P}}\otimes(\boldsymbol{\mathcal{P}}\nabla^2F(\Theta^\star)\boldsymbol{\mathcal{P}})+(\boldsymbol{\mathcal{P}}\nabla^2F(\Theta^\star)\boldsymbol{\mathcal{P}})\otimes\boldsymbol{\mathcal{P}}\Big)^\dagger$.

First note that $\boldsymbol{\mathcal{P}}\nabla^2F(\Theta^\star)\boldsymbol{\mathcal{P}}=\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\bar{A}$, with $\bar{A}$ the average Hessian at $\theta^\star$.

Let Π be the operator such that $\Pi(x_1\otimes x_2\otimes x_3\otimes x_4)=x_1\otimes x_3\otimes x_2\otimes x_4$.

Then:

$$\begin{aligned}&\Pi\Big(\boldsymbol{\mathcal{P}}\otimes(\boldsymbol{\mathcal{P}}\nabla^2F(\Theta^\star)\boldsymbol{\mathcal{P}})+(\boldsymbol{\mathcal{P}}\nabla^2F(\Theta^\star)\boldsymbol{\mathcal{P}})\otimes\boldsymbol{\mathcal{P}}\Big)\Pi^\top\\ &=\Pi\Big(\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\boldsymbol{I}\otimes\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\bar{A}+\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\bar{A}\otimes\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\boldsymbol{I}\Big)\Pi^\top\\ &=\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\big(\boldsymbol{I}\otimes\bar{A}+\bar{A}\otimes\boldsymbol{I}\big)\\ &=\frac{1}{m}(\mathbf{1}\otimes\mathbf{1})\times\frac{1}{m}(\mathbf{1}\otimes\mathbf{1})^\top\otimes\big(\boldsymbol{I}\otimes\bar{A}+\bar{A}\otimes\boldsymbol{I}\big) \ .\end{aligned}$$

This operator is zero on $\mathrm{Span}(\mathbf{1}\otimes\mathbf{1})^\perp\otimes\mathbb{R}^{md\times md}$, and acts as $\boldsymbol{I}\otimes\big(\boldsymbol{I}\otimes\bar{A}+\bar{A}\otimes\boldsymbol{I}\big)$ on $\mathrm{Span}(\mathbf{1}\otimes\mathbf{1})\otimes\mathbb{R}^{md\times md}$.

Its pseudo-inverse is thus $\frac{1}{m}(\mathbf{1}\otimes\mathbf{1})\times\frac{1}{m}(\mathbf{1}\otimes\mathbf{1})^\top\otimes\big(\boldsymbol{I}\otimes\bar{A}+\bar{A}\otimes\boldsymbol{I}\big)^{-1}$, and we deduce that

$$\Big(\boldsymbol{\mathcal{P}}\otimes(\boldsymbol{\mathcal{P}}\nabla^2F(\Theta^\star)\boldsymbol{\mathcal{P}})+(\boldsymbol{\mathcal{P}}\nabla^2F(\Theta^\star)\boldsymbol{\mathcal{P}})\otimes\boldsymbol{\mathcal{P}}\Big)^\dagger=\Pi^\top\Big(\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\big(\boldsymbol{I}\otimes\bar{A}+\bar{A}\otimes\boldsymbol{I}\big)^{-1}\Big)\Pi \ .$$

Therefore the result from Lemma B.4 can be rewritten as

$$\Pi\int_{\mathbb{R}^{md}}(\Theta-\Theta_{\mathrm{det}})^{\otimes 2}\,\pi_\gamma(d\Theta)=\gamma\Big(\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\frac{1}{m}(\mathbf{1}\mathbf{1}^\top)\otimes\big(\boldsymbol{I}\otimes\bar{A}+\bar{A}\otimes\boldsymbol{I}\big)^{-1}\Big)\times\Pi\boldsymbol{\mathcal{C}}(\Theta^\star)+\mathcal{O}(\gamma^{3/2}) \ . \tag{28}$$

For $x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} = \sum_{k=1}^m e_k \otimes x_k$, we have $\Pi(x \otimes x) = \sum_{k,\ell=1}^m e_k \otimes e_\ell \otimes x_k \otimes x_\ell$, so (28) becomes

$$\begin{aligned}
&\sum_{k,\ell=1}^m e_k \otimes e_\ell \otimes \left(\int_{\mathbb{R}^{md}} (\theta_k - \theta_k^{\mathrm{det}}) \otimes (\theta_\ell - \theta_\ell^{\mathrm{det}})\, \pi_\gamma(d\Theta) \right) \\
&= \gamma\Big(\frac{1}{m}(\mathbf{1}\mathbf{1}^\top) \otimes \frac{1}{m}(\mathbf{1}\mathbf{1}^\top) \otimes \big(\boldsymbol{I} \otimes \bar{A} + \bar{A} \otimes \boldsymbol{I}\big)^{-1}\Big) \times \left(\sum_{i=1}^m e_i \otimes e_i \otimes \mathbb{E}\left[\left(\varepsilon^{(i)}(\theta)\right)^{\otimes 2} \right] \right) + \mathcal{O}(\gamma^{3/2}) \\
&= \frac{\gamma}{m} \mathbf{1} \otimes \mathbf{1} \otimes \left(\big(\boldsymbol{I} \otimes \bar{A} + \bar{A} \otimes \boldsymbol{I}\big)^{-1} \mathcal{C}(\theta^\star) \right) + \mathcal{O}(\gamma^{3/2}) \\
&= \frac{\gamma}{m} \sum_{k,\ell=1}^m e_k \otimes e_l \otimes \left(\big(\boldsymbol{I} \otimes \bar{A} + \bar{A} \otimes \boldsymbol{I}\big)^{-1} \mathcal{C}(\theta^\star) \right) + \mathcal{O}(\gamma^{3/2}) \ ,
\end{aligned}$$

from which we deduce, for any $k, \ell \in \{1, \dots, m\}$,

$$\int_{\mathbb{R}^{md}} (\theta_k - \theta_k^{\mathrm{det}}) \otimes (\theta_\ell - \theta_\ell^{\mathrm{det}})\, \pi_\gamma(d\Theta) = \frac{\gamma}{m} \big(\boldsymbol{I} \otimes \bar{A} + \bar{A} \otimes \boldsymbol{I}\big)^{-1} \mathcal{C}(\theta^\star) + \mathcal{O}(\gamma^{3/2}) \ .$$

□

Proof of Proposition 4.2. Proposition 4.2 is obtained by combining Lemma B.4 and Lemma B.9. □

Proof of Proposition 4.5. Similarly, Lemma B.5 and Lemma B.9 yield Proposition 4.5. □

C NON-ASYMPTOTIC ANALYSIS

We now give more explicit upper-bounds on the variance of the stationary distribution, and use them to establish non-asymptotic upper-bounds for the iterates of DSGD.

C.1 Upper bound on Limiting Variance

We first study the variance of the limiting distribution. To this end, we decompose the error in the limit distribution in a consensus and a disagreement part, and study them separately.

Define

$$B = \frac{2000\big(\tau_4^2 + \gamma^2 \frac{L^4}{\mu^2} \Lambda^2 \zeta_\star^2\big)}{\mu} \ .$$

According to Lemma B.2 and Hölder's inequality,

$$\int \|\Theta - \Theta_{\mathrm{det}}\|^2 \pi_\gamma(\mathrm{d}\Theta) \le B\gamma \ , \qquad \Big(\int \|\Theta - \Theta_{\mathrm{det}}\|^4 \pi_\gamma(\mathrm{d}\Theta) \Big)^{1/2} \le B\gamma \ . \tag{29}$$

Starting from (18), we have

$$\Theta_1 - \Theta_{\mathrm{det}} = \boldsymbol{W}(\boldsymbol{I} - \gamma \nabla^2 F(\Theta_{\mathrm{det}}))(\Theta_0 - \Theta_{\mathrm{det}}) - \gamma \boldsymbol{W}\big(\varepsilon_1(\Theta_0) + R(\Theta_0)\big) \ , \tag{30}$$

where, by (19), the remainder satisfies $\|R(\Theta)\| \le \frac{K_3}{2} \|\Theta - \Theta_{\mathrm{det}}\|^2$.

Bound on Consensus. Defining $\bar{\theta}_0$, $\bar{\theta}_1, \bar{\theta}_{det}$ the average of all agents parameters for Θ_0, Θ_1, and Θ_{det} respectively, and projecting (30) on the consensus space, we obtain

$$\bar{\theta}_1 - \bar{\theta}_{det} = (\boldsymbol{I} - \gamma \bar{A}(\Theta_{\mathrm{det}}))(\bar{\theta}_0 - \bar{\theta}_{det}) - \gamma\big(\bar{\varepsilon}_1(\Theta_0) + \bar{R}(\Theta_0)\big) \ , \tag{31}$$

where we also defined $\bar{A}(\Theta_{\text{det}}) \in \mathbb{R}^{d\times d}, \bar{\varepsilon}_1(\Theta_0) \in \mathbb{R}^d, \bar{R}(\Theta_0) \in \mathbb{R}^d$ as the averaged Hessian, noise and remainder. Taking the tensor square of (31) and expectation over the noise at step 1, we have

$$
\begin{aligned}
(\bar{\theta}_1 - \bar{\theta}_{det})^{\otimes 2} = (\boldsymbol{I} - \gamma \bar{A}(\Theta_{\text{det}}))(\bar{\theta}_0 - \bar{\theta}_{det})^{\otimes 2}(\boldsymbol{I} - \gamma \bar{A}(\Theta_{\text{det}})) - \gamma(\boldsymbol{I} - \gamma \bar{A}(\Theta_{\text{det}}))(\bar{\theta}_0 - \bar{\theta}_{det})\bar{R}(\Theta_0)^\top \\
- \gamma \bar{R}(\Theta_0)(\bar{\theta}_0 - \bar{\theta}_{det})^\top(\boldsymbol{I} - \gamma \bar{A}(\Theta_{\text{det}})) + \gamma^2 \mathbb{E}[\bar{\varepsilon}_1(\Theta_0)^{\otimes 2}] \ . \qquad (32)
\end{aligned}
$$

Denote by $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{P}}\theta}$ (respectively $\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Q}}\theta}$) the covariance matrix of the consensus part (respectively disagreement part) under the stationary distribution, and $\bar{\boldsymbol{\Sigma}}_\varepsilon$ the covariance matrix of the average noise under the stationary distribution.

Assuming $\Theta_0 \sim \pi_\gamma$ started from the stationary distribution, we can integrate this equality over this distribution to obtain

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{P}}\theta} &= (\boldsymbol{I} - \gamma \bar{A}(\Theta_{\text{det}}))\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{P}}\theta}(\boldsymbol{I} - \gamma \bar{A}(\Theta_{\text{det}})) + \gamma^2 \bar{\boldsymbol{\Sigma}}_\varepsilon \\
&\quad - \gamma \int \Big\{ (\boldsymbol{I} - \gamma \bar{A}(\Theta_{\text{det}}))(\bar{\theta}_0 - \bar{\theta}_{det})\bar{R}(\Theta_0)^\top + \bar{R}(\Theta_0)(\bar{\theta}_0 - \bar{\theta}_{det})^\top(\boldsymbol{I} - \gamma \bar{A}(\Theta_{\text{det}})) \Big\} \pi_\gamma(\mathrm{d}\Theta_0) \ .
\end{aligned}
$$

Taking the operator norm and using Hölder's inequality gives

$$
\|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{P}}\theta}\| \le (1-\gamma\mu)^2 \|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{P}}\theta}\| + \gamma^2 \|\bar{\boldsymbol{\Sigma}}_\varepsilon\| + 2\gamma \Big(\int \|\bar{\theta}_0 - \bar{\theta}_{det}\|^2 \pi_\gamma(\mathrm{d}\Theta_0) \Big)^{1/2} \Big(\int \|\bar{R}(\Theta_0)\|^2 \pi_\gamma(\mathrm{d}\Theta_0) \Big)^{1/2} \ . \qquad (33)
$$

Using Jensen's inequality:

$$
\begin{aligned}
\Big(\int \|\bar{\theta}_0 - \bar{\theta}_{det}\|^2 \pi_\gamma(\mathrm{d}\Theta_0) \Big)^{1/2} &\le \Big(\int \frac{1}{m} \|\Theta_0 - \Theta_{\text{det}}\|^2 \pi_\gamma(\mathrm{d}\Theta_0) \Big)^{1/2} \\
&\le \frac{1}{\sqrt{m}} B^{1/2} \gamma^{1/2} \ .
\end{aligned}
$$

Similarly:

$$
\begin{aligned}
\Big(\int \|\bar{R}(\Theta_0)\|^2 \pi_\gamma(\mathrm{d}\Theta_0) \Big)^{1/2} &\le \Big(\int \frac{1}{m} \|R(\Theta_0)\|^2 \pi_\gamma(\mathrm{d}\Theta_0) \Big)^{1/2} \\
&\le \frac{K_3}{2\sqrt{m}} \Big(\int \|\Theta_0 - \Theta_{\text{det}}\|^4 \pi_\gamma(\mathrm{d}\Theta_0) \Big)^{1/2} \\
&\le \frac{K_3 B}{2\sqrt{m}} \gamma \ .
\end{aligned}
$$

Hence

$$
\Big(\int \|\bar{\theta}_0 - \bar{\theta}_{det}\|^2 \pi_\gamma(\mathrm{d}\Theta_0) \Big)^{1/2} \Big(\int \|\bar{R}(\Theta_0)\|^2 \pi_\gamma(\mathrm{d}\Theta_0) \Big)^{1/2} \le \frac{K_3 B^{3/2}}{2m} \gamma^{3/2} \ ,
$$

and

$$
\|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{P}}\theta}\| \le (1-\gamma\mu)^2 \|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{P}}\theta}\| + \gamma^2 \|\bar{\boldsymbol{\Sigma}}_\varepsilon\| + \gamma^{5/2} \frac{K_3 B^{3/2}}{m} \ ,
$$

which gives, after reorganizing and bounding the covariance matrix,

$$
\|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{P}}\theta}\| \le \frac{\gamma}{\mu m} \|\boldsymbol{\Sigma}_\varepsilon\| + \gamma^{3/2} \frac{K_3 B^{3/2}}{\mu m} \ .
$$

Bound on Disagreement. First, we recall that for any vector $u \in \mathbb{R}^{md}$, since $\boldsymbol{W}$ is doubly stochastic,

$$
\boldsymbol{\mathcal{Q}}\boldsymbol{W}u = \boldsymbol{W}u - \boldsymbol{\mathcal{P}}\boldsymbol{W}u = \boldsymbol{W}u - \boldsymbol{W}\boldsymbol{\mathcal{P}}u = \boldsymbol{W}\boldsymbol{\mathcal{Q}}u \ . \qquad (34)
$$

Starting from (18) and applying $\boldsymbol{\mathcal{Q}}$, we thus have

$$
\boldsymbol{\mathcal{Q}}(\Theta_1 - \Theta_{\text{det}}) = \boldsymbol{W}\boldsymbol{\mathcal{Q}}(\Theta_0 - \Theta_{\text{det}}) - \gamma \boldsymbol{W}\boldsymbol{\mathcal{Q}}(\nabla^2 F(\Theta_{\text{det}})(\Theta_0 - \Theta_{\text{det}})) - \gamma \boldsymbol{W}\boldsymbol{\mathcal{Q}}\big(\varepsilon_1(\Theta_0) + R(\Theta_0)\big) \ .
$$

Computing the tensor square and taking the expectation on the noise, we obtain

$$\begin{aligned}\Big(\boldsymbol{\mathcal{Q}}(\Theta_1-\Theta_{\mathrm{det}})\Big)^{\otimes 2} &= \boldsymbol{W}\Big(\boldsymbol{\mathcal{Q}}(\Theta_0-\Theta_{\mathrm{det}})\Big)^{\otimes 2}\boldsymbol{W}\\ &\quad -\gamma\Big(\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\Theta_0-\Theta_{\mathrm{det}})\Big)\Big(\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\nabla^2 F(\Theta_{\mathrm{det}})(\Theta_0-\Theta_{\mathrm{det}}))+\boldsymbol{W}\boldsymbol{\mathcal{Q}}R(\Theta_0)\Big)^{\top}\\ &\quad -\gamma\Big(\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\nabla^2 F(\Theta_{\mathrm{det}})(\Theta_0-\Theta_{\mathrm{det}}))+\boldsymbol{W}\boldsymbol{\mathcal{Q}}R(\Theta_0)\Big)\Big(\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\Theta_0-\Theta_{\mathrm{det}})\Big)^{\top}\\ &\quad +\gamma^2\Big(\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\nabla^2 F(\Theta_{\mathrm{det}})(\Theta_0-\Theta_{\mathrm{det}}))+\boldsymbol{W}\boldsymbol{\mathcal{Q}}R(\Theta_0)\Big)^{\otimes 2}+\gamma^2\mathbb{E}[\boldsymbol{W}\boldsymbol{\mathcal{Q}}\varepsilon_1(\Theta_0)^{\otimes 2}\boldsymbol{\mathcal{Q}}\boldsymbol{W}]\ .\end{aligned}$$

Integrating over the stationary distribution gives

$$\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Q}}\theta}=\boldsymbol{W}\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Q}}\theta}\boldsymbol{W}-\gamma\boldsymbol{B_1}+\gamma^2\boldsymbol{B_2}\ , \tag{35}$$

where we introduced

$$\begin{aligned}\boldsymbol{B_1}&=\int\Big(\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\Theta_0-\Theta_{\mathrm{det}})\Big)\Big(\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\nabla^2 F(\Theta_{\mathrm{det}})(\Theta_0-\Theta_{\mathrm{det}}))+\boldsymbol{W}\boldsymbol{\mathcal{Q}}R(\Theta_0)\Big)^{\top}\pi_\gamma(\mathrm{d}\Theta_0)\\ &\quad+\int\Big(\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\nabla^2 F(\Theta_{\mathrm{det}})(\Theta_0-\Theta_{\mathrm{det}}))+\boldsymbol{W}\boldsymbol{\mathcal{Q}}R(\Theta_0)\Big)\Big(\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\Theta_0-\Theta_{\mathrm{det}})\Big)^{\top}\pi_\gamma(\mathrm{d}\Theta_0)\ ,\\ \boldsymbol{B_2}&=\int\Big(\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\nabla^2 F(\Theta_{\mathrm{det}})(\Theta_0-\Theta_{\mathrm{det}}))+\boldsymbol{W}\boldsymbol{\mathcal{Q}}R(\Theta_0)\Big)^{\otimes 2}\pi_\gamma(\mathrm{d}\Theta_0)+\int\mathbb{E}[\boldsymbol{W}\boldsymbol{\mathcal{Q}}\varepsilon_1(\Theta_0)^{\otimes 2}\boldsymbol{\mathcal{Q}}\boldsymbol{W}]\ .\end{aligned}$$

Note that these matrices can be upper bounded, in operator norm, as

$$\begin{aligned}\|\boldsymbol{B_1}\|&\le 2\lambda_2^2\int\Big(L\|\Theta_0-\Theta_{\mathrm{det}}\|+\|R(\Theta_0)\|\Big)\|\Theta_0-\Theta_{\mathrm{det}}\|\pi_\gamma(\mathrm{d}\Theta_0)\\ &=2\lambda_2^2\int\Big(L\|\Theta_0-\Theta_{\mathrm{det}}\|^2+\|R(\Theta_0)\|\|\Theta_0-\Theta_{\mathrm{det}}\|\Big)\pi_\gamma(\mathrm{d}\Theta_0)\\ &\le 2\lambda_2^2\int\Big(L\|\Theta_0-\Theta_{\mathrm{det}}\|^2+\frac{K_3}{2}\|\Theta_0-\Theta_{\mathrm{det}}\|^3\Big)\pi_\gamma(\mathrm{d}\Theta_0)\\ &\le 2\lambda_2^2\Big(LB\gamma+\frac{K_3}{2}B^{3/2}\gamma^{3/2}\Big)\\ &=2\lambda_2^2LB\gamma+\lambda_2^2K_3B^{3/2}\gamma^{3/2}\ ,\end{aligned}$$

and

$$\begin{aligned}\|\boldsymbol{B_2}\|&\le\int\Big\|\boldsymbol{W}\boldsymbol{\mathcal{Q}}(\nabla^2 F(\Theta_{\mathrm{det}})(\Theta_0-\Theta_{\mathrm{det}}))-\gamma\boldsymbol{W}\boldsymbol{\mathcal{Q}}R(\Theta_0)\Big\|^2\pi_\gamma(\mathrm{d}\Theta_0)+\|\boldsymbol{W}\boldsymbol{\mathcal{Q}}\boldsymbol{\Sigma}_\varepsilon\boldsymbol{\mathcal{Q}}\boldsymbol{W}\|\\ &\le 2\lambda_2^2\int L^2\|\Theta_0-\Theta_{\mathrm{det}}\|^2\pi_\gamma(\mathrm{d}\Theta_0)+2\lambda_2^2\int\|R(\Theta_0)\|^2\pi_\gamma(\mathrm{d}\Theta_0)+\lambda_2^2\|\boldsymbol{\Sigma}_\varepsilon\|\\ &\le 2\lambda_2^2L^2B\gamma+\frac{1}{2}\gamma^2\lambda_2^2K_3^2B^2+\lambda_2^2\|\boldsymbol{\Sigma}_\varepsilon\|\ ,\end{aligned}$$

where we used $\|\boldsymbol{W}\boldsymbol{\mathcal{Q}}\|\le\lambda_2$.

Remark that (35) is a Sylvester equation, whose solution is

$$\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Q}}\theta}=\sum_{k\ge 0}\boldsymbol{W}^k(-\gamma\boldsymbol{B_1}+\gamma^2\boldsymbol{B_2})\boldsymbol{W}^k\ . \tag{36}$$

Taking the norm of (36) gives

$$\begin{aligned}\|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Q}}\theta}\|&\le\frac{\gamma}{1-\lambda_2^2}\|\boldsymbol{B_1}+\gamma\boldsymbol{B_2}\|\\ &\le\frac{\gamma^2}{1-\lambda_2^2}\big(2\lambda_2^2LB+\lambda_2^2K_3B^{3/2}\gamma^{1/2}+2\lambda_2^2L^2B\gamma+\frac{1}{2}\gamma^2\lambda_2^2K_3^2B^2+\lambda_2^2\|\boldsymbol{\Sigma}_\varepsilon\|\big)\\ &\le\frac{\gamma^2\lambda_2^2}{1-\lambda_2^2}\big(4LB+K_3B^{3/2}\gamma^{1/2}+\frac{1}{2}\gamma^2K_3^2B^2+\|\boldsymbol{\Sigma}_\varepsilon\|\big)\ ,\end{aligned}$$

where we used $\gamma L\le 1$ in the last inequality.

Bound on Variance. It holds that $\|\boldsymbol{\Sigma}_\theta\| \leq \|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Q}}\theta}\| + \|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{P}}\theta}\| + 2\|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Q}}\boldsymbol{\mathcal{P}}}\| \leq 2(\|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{Q}}\theta}\| + \|\boldsymbol{\Sigma}_{\boldsymbol{\mathcal{P}}\theta}\|)$. Based on the two bounds above, we then have the following bound on the variance:

$$\|\boldsymbol{\Sigma}_\theta\| \leq 2\Big(\frac{\gamma}{\mu m}\|\boldsymbol{\Sigma}_\varepsilon\| + \gamma^{3/2}\frac{K_3 B^{3/2}}{\mu m} + \frac{\gamma^2\lambda_2^2}{1-\lambda_2^2}\big(4LB + K_3 B^{3/2}\gamma^{1/2} + \frac{1}{2}\gamma^2 K_3^2 B^2 + \|\boldsymbol{\Sigma}_\varepsilon\|\big)\Big) \ . \tag{37}$$

C.2 Non-Asymptotic Bounds for the Iterates of DSGD.

If $(\Theta'_t)_{t\geq 0}$ is another sequence, such that $\Theta'_0 \sim \pi_\gamma$, and defined using the same noise sequence as $(\Theta_t)_{t\geq 0}$, then

$$\mathbb{E}[\|\Theta_t - \Theta^\star\|^2] \leq 3\mathbb{E}[\|\Theta_t - \Theta'_t\|^2] + 3\|\Theta_{\mathrm{det}} - \Theta^\star\|^2 + 3\int \|\Theta - \Theta_{\mathrm{det}}\|^2 \pi_\gamma(\mathrm{d}\Theta) \ .$$

Bounding each term using Proposition 4.1, Lemma 3.4 and the bounds established previously, we obtain

$$\begin{aligned}\mathbb{E}[\|\Theta_t - \Theta^\star\|^2] \leq 3(1-\gamma\mu)^t\mathbb{E}[\|\Theta_0 - \Theta'_0\|^2] + 3\frac{\gamma^2 L^2\Lambda^2}{\mu^2}\zeta_\star^2 \\ + 6\Big(\frac{\gamma}{\mu m}\|\boldsymbol{\Sigma}_\varepsilon\| + \gamma^{3/2}\frac{K_3 B^{3/2}}{\mu m}\Big) \\ + 6\Big(\frac{\gamma^2\lambda_2^2}{1-\lambda_2^2}\big(4LB + K_3 B^{3/2}\gamma^{1/2} + \frac{1}{2}\gamma^2 K_3^2 B^2 + \|\boldsymbol{\Sigma}_\varepsilon\|\big)\Big) \ .\end{aligned}$$

Proof of Corollary 4.7. The proof is straightforward from Theorem 4.6: the conditions on γ ensure that all terms in the upper-bound except the first one are of order ε, and the condition on T ensures that the first term is of order ε. □

D RICHARDSON-ROMBERG FOR DECENTRALIZED LEARNING

We now prove the results stated in Section 5 on Richardson-Romberg extrapolation.

Proof of Proposition 5.1. This result is a direct consequence of Proposition 3.6. Indeed, first write

$$\begin{aligned}\Theta_{\mathrm{det}}^{\gamma} &= \Theta^\star - \gamma(\boldsymbol{I} - \boldsymbol{\mathcal{H}})\boldsymbol{\mathcal{G}}\nabla F(\Theta^\star) + R(\gamma) \\ \Theta_{\mathrm{det}}^{\gamma/2} &= \Theta^\star - \frac{\gamma}{2}(\boldsymbol{I} - \boldsymbol{\mathcal{H}})\boldsymbol{\mathcal{G}}\nabla F(\Theta^\star) + R(\gamma/2) \ ,\end{aligned}$$

where

$$\|R(\gamma)\| \leq \gamma^2\frac{L^2}{2\mu^2}\Lambda^2\Big(\tfrac{K_3}{\mu\sqrt{m}}\zeta_\star^2 + L\cdot\zeta_\star\Big) \ .$$

Hence, using $2\Theta_{\mathrm{det}}^{\gamma/2} - \Theta_{\mathrm{det}}^{\gamma} - \Theta^\star = 2R(\gamma/2) - R(\gamma)$, we have

$$\|\Theta_{\mathrm{det}}^{\mathrm{RR},\gamma} - \Theta^\star\| = \|2\Theta_{\mathrm{det}}^{\gamma/2} - \Theta_{\mathrm{det}}^{\gamma} - \Theta^\star\| \leq 2\|R(\gamma/2)\| + \|R(\gamma)\| \leq \gamma^2\frac{3L^2}{4\mu^2}\Lambda^2\Big(\tfrac{K_3}{\mu\sqrt{m}}\zeta_\star^2 + L\cdot\zeta_\star\Big) \ .$$

□

Proof of Corollary 5.2. Using Lemma 3.1 and Proposition 5.1,

$$\begin{aligned}\|\Theta_t^{\mathrm{RR},\gamma} - \Theta^\star\| &\leq \|\Theta_t^{\mathrm{RR},\gamma} - \Theta_{\mathrm{det}}^{\mathrm{RR},\gamma}\| + \|\Theta_{\mathrm{det}}^{\mathrm{RR},\gamma} - \Theta^\star\| \\ &\leq 2\|\Theta_t^{\gamma/2} - \Theta_{\mathrm{det}}^{\gamma/2}\| + \|\Theta_t^{\gamma} - \Theta_{\mathrm{det}}^{\gamma/2}\| + \|\Theta_{\mathrm{det}}^{\mathrm{RR},\gamma} - \Theta^\star\| \\ &\leq 2(1-\gamma\mu/2)^t\|\Theta_0 - \Theta_{\mathrm{det}}^{\gamma/2}\| + (1-\gamma\mu)^t\|\Theta_0 - \Theta_{\mathrm{det}}^{\gamma}\| + \gamma^2\frac{5L^2}{8\mu^2}\Lambda^2\big(\frac{K_3}{\mu\sqrt{m}}\zeta_\star^2 + L\zeta_\star\big) \\ &\leq 3(1-\gamma\mu/2)^t\|\Theta_0 - \Theta^\star\| + 2(1-\gamma\mu/2)^t\|\Theta^\star - \Theta_{\mathrm{det}}^{\gamma/2}\| + (1-\gamma\mu)^t\|\Theta^\star - \Theta_{\mathrm{det}}^{\gamma}\| \\ &\quad + \gamma^2\frac{5L^2}{8\mu^2}\Lambda^2\big(\frac{K_3}{\mu\sqrt{m}}\zeta_\star^2 + L\zeta_\star\big) \ .\end{aligned}$$

We then apply Lemma 3.4,

$$\|\Theta_t^{\mathrm{RR},\gamma} - \Theta^\star\| \le 3(1-\gamma\mu/2)^t\|\Theta_0 - \Theta^\star\| + 2(1-\gamma\mu/2)^t\frac{\gamma L\Lambda}{\mu}\zeta_\star + \gamma^2\frac{5L^2}{8\mu^2}\Lambda^2\big(\frac{K_3}{\mu\sqrt{m}}\zeta_\star^2 + L\zeta_\star\big) \ ,$$

hence the conclusion.

If $K_3/m^{1/2}$ is small enough, and if γ is as in the statement of the corollary, then the second and third terms of the above bound are of order ε.

We thus look for T such that $3(1-\gamma\mu/2)^T\|\Theta_0 - \Theta^\star\| \lesssim \varepsilon$, which yields $T \gtrsim \frac{\log(\varepsilon/\|\Theta_0-\Theta^\star\|)}{\log(1-\gamma\mu/2)} \gtrsim \frac{\log(\|\Theta_0-\Theta^\star\|/\varepsilon)}{\gamma\mu}$, which concludes. □

Proof of Corollary 5.3. We first write

$$\|\Theta_t^{\mathrm{RR},\gamma} - \Theta^\star\|^2 \le 4\|\Theta_t^{\gamma/2} - \Theta_{\mathrm{det}}^{\gamma/2}\|^2 + \|\Theta_t^{\gamma} - \Theta^\star\|^2 \ .$$

Then, we simply need to make sure that the upper-bound given in Theorem 4.6 is of order ε. This can be done similarly to the proof of Corollary 5.2, using the assumptions on both γ and T. □

E USEFUL RESULTS ON MATRICES

Lemma E.1. *Let $A \succeq 0$ and $B \succ 0$. Then, as t approaches 0,*

$$(A+tB)^{-1} = \frac{1}{t}\big(\boldsymbol{\mathcal{P}}B\boldsymbol{\mathcal{P}}\big)^\dagger + \mathcal{O}(1) \ ,$$

where $\boldsymbol{\mathcal{P}}$ is the orthogonal projection onto the kernel of A.

Moreover,

$$\left\|(A+tB)^{-1} - \frac{1}{t}\big(\boldsymbol{\mathcal{P}}B\boldsymbol{\mathcal{P}}\big)^\dagger\right\|_2 \le \frac{1}{\lambda_{\min}^+(A)}\left(1 + \frac{\lambda_{\max}(B)}{\lambda_{\min}(B)}\right)^2 \ ,$$

where $\lambda_{\min}^+(A)$ is A's smallest non-zero eigenvalue.

Proof. Let U be an orthogonal matrix such that

$$A = U\begin{pmatrix} 0 & 0 \\ 0 & D\end{pmatrix}U^\top \ , \qquad B = U\begin{pmatrix} B_{00} & B_{01} \\ B_{10} & B_{11}\end{pmatrix}U^\top \ ,$$

where D is diagonal and non-singular, $B_{00}, B_{11} \succ 0$, and $B_{10} = B_{01}^\top$.

Then

$$A + tB = U\underbrace{\begin{pmatrix} tB_{00} & tB_{01} \\ tB_{10} & D + tB_{11}\end{pmatrix}}_{:=M}U^\top \ .$$

The Schur complement is defined as

$$S(t) = D + t\big(B_{11} - B_{10}B_{00}^{-1}B_{01}\big) \ .$$

Then:

$$\begin{aligned} M^{-1} &= \begin{pmatrix} \frac{1}{t}B_{00}^{-1} + B_{00}^{-1}B_{01}S(t)^{-1}B_{10}B_{00}^{-1} & -B_{00}^{-1}B_{01}S(t)^{-1} \\ -S(t)^{-1}B_{10}B_{00}^{-1} & S(t)^{-1}\end{pmatrix} \\ &= \frac{1}{t}\begin{pmatrix} B_{00}^{-1} & 0 \\ 0 & 0\end{pmatrix} + \begin{pmatrix} B_{00}^{-1}B_{01}S(t)^{-1}B_{10}B_{00}^{-1} & -B_{00}^{-1}B_{01}S(t)^{-1} \\ -S(t)^{-1}B_{10}B_{00}^{-1} & S(t)^{-1}\end{pmatrix} \ , \end{aligned}$$

and we thus have

$$(A+tB)^{-1} = \frac{1}{t} U \begin{pmatrix} B_{00}^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^\top + U \begin{pmatrix} B_{00}^{-1} B_{01} S(t)^{-1} B_{10} B_{00}^{-1} & -B_{00}^{-1} B_{01} S(t)^{-1} \\ -S(t)^{-1} B_{10} B_{00}^{-1} & S(t)^{-1} \end{pmatrix} U^\top \ .$$

If we set

$$\boldsymbol{\mathcal{P}} = U \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} U^\top$$

the orthogonal projection onto the kernel of A, then

$$\frac{1}{t} U \begin{pmatrix} B_{00}^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^\top = \frac{1}{t} \big(\boldsymbol{\mathcal{P}} B \boldsymbol{\mathcal{P}} \big)^\dagger \ .$$

Moreover:

- $\left\| U \begin{pmatrix} B_{00}^{-1} B_{01} S(t)^{-1} B_{10} B_{00}^{-1} & -B_{00}^{-1} B_{01} S(t)^{-1} \\ -S(t)^{-1} B_{10} B_{00}^{-1} & S(t)^{-1} \end{pmatrix} U^\top \right\|_2 \leq \|S(t)^{-1}\|_2 \big(1 + \|B_{00}^{-1}\|_2 \cdot \|B_{01}\|_2\big)^2$
- $S(t) \succeq D$, so $\lambda_{\min}(S(t)) \geq \lambda_{\min}(D)$, and $\|S(t)^{-1}\|_2 \leq \dfrac{1}{\lambda_{\min}(D)} = \dfrac{1}{\lambda_{\min}^+(A)}$
- $\|B_{01}\|_2 \leq \|B\|_2 = \lambda_{\max}(B)$
- $\lambda_{\min}(B_{00}) \geq \lambda_{\min}(B)$, so $\|B_{00}^{-1}\|_2 \leq \dfrac{1}{\lambda_{\min}(B)}$.

Putting everything together yields:

$$\left\| U \begin{pmatrix} B_{00}^{-1} B_{01} S(t)^{-1} B_{10} B_{00}^{-1} & -B_{00}^{-1} B_{01} S(t)^{-1} \\ -S(t)^{-1} B_{10} B_{00}^{-1} & S(t)^{-1} \end{pmatrix} U^\top \right\|_2 \leq \frac{1}{\lambda_{\min}^+(A)} \left(1 + \frac{\lambda_{\max}(B)}{\lambda_{\min}(B)} \right)^2 \ ,$$

hence the conclusion. ☐

Lemma E.2. *Let $A \succ 0, B \succeq 0$. Then it holds that*

$$\|(A+B)^{-1} - A^{-1}\|_2 \leq \|A^{-1}\|_2^2 \cdot \|B\|_2 \ .$$

Proof. We use the identity

$$M^{-1} - N^{-1} = -N^{-1}(M-N)M^{-1}$$

to obtain

$$(A+B)^{-1} - A^{-1} = -A^{-1} B (A+B)^{-1} \ ,$$

hence

$$\|(A+B)^{-1} - A^{-1}\|_2 \leq \|A^{-1}\|_2 \cdot \|B\|_2 \cdot \|(A+B)^{-1}\|_2 \ .$$

And $A + B \succeq A$, so $(A+B)^{-1} \preceq A^{-1}$, which implies $\|(A+B)^{-1}\|_2 \leq \|A^{-1}\|_2$, hence the conclusion. ☐