Stochastic Gradient Descent under Markovian Sampling Schemes

Mathieu Even [1]

Abstract

We study a variation of vanilla stochastic gradient descent where the optimizer only has access to a "Markovian sampling scheme". These schemes encompass applications that range from decentralized optimization with a random walker (*token algorithms*), to RL and online system identification problems. We focus on obtaining rates of convergence under the least restrictive assumptions possible on the underlying Markov chain and on the functions optimized. We first unveil the theoretical lower bound for methods that sample stochastic gradients along the path of a Markov chain, making appear a dependency in the hitting time of the underlying Markov chain. We then study *Markov chain SGD* (MC-SGD) under much milder regularity assumptions than prior works. We finally introduce MC-SAG, an alternative to MC-SGD with variance reduction, that only depends on the hitting time of the Markov chain, therefore obtaining a communication-efficient token algorithm.

1. Introduction

In this paper, we consider a stochastic optimization problem that takes root in decentralized optimization, estimation problems, and Reinforcement Learning. Consider a function f defined as:

$$f(\mathbf{x}) = \mathbb{E}_{v\sim\pi}\left[f_v(\mathbf{x})\right], \quad \mathbf{x}\in\mathbb{R}^d, \tag{1}$$

where π is a probability distribution over a set $\mathcal{V}$, and f_v are smooth functions on $\mathbb{R}^d$ for all v in $\mathcal{V}$. Classicaly, this represents the loss of a model parameterized by $\mathbf{x}$ on data parameterized by v. If *i.i.d.* samples $(v_t)_{t\geqslant 0}$ of law π and their corresponding gradient estimates (∇f_{v_t}) were accessible, one could directly apply SGD-like algorithms, that have proved to be efficient in large scale machine learning problems (Bottou et al., 2018). We however consider in this paper a different setting: we assume the existence of a Markov chain (v_t) of state space $\mathcal{V}$ and stationary distribution π. The optimizer may then use biased stochastic gradients along the path of this Markov chain to perform incremental updates. She may for instance use the *Markov chain SGD* (MC-SGD) algorithm, defined through the following recursion:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma\nabla f_{v_t}(\mathbf{x}_t)\,. \tag{2}$$

Being "ergodically unbiased", such iterates should behave closely to those of vanilla SGD. The analysis is however notoriously difficult, since in (2), variable $\mathbf{x}_t$ and the current state of the Markov chain v_t are not independent, so that $\mathbb{E}\left[\nabla f_{v_t}(\mathbf{x}_t)|\mathbf{x}_t\right]$ can be arbitrarily far from $\nabla f(\mathbf{x}_t)$. This paper focuses on analyzing algorithms that incrementally sample stochastic gradients alongside the Markov chain (v_t), motivated by the following applications.

1.1. Token algorithms

Traditional machine learning optimization algorithms require data centralization, raising scalability and privavy issues, hence the alternative of Federated Learning, where users' data is held on device, and the training is orchestrated at a server level. Decentralized optimization goes further, by removing the dependency over a central entity, leading to increased scalability, privacy and robustness to node failures, broadening the range of applications to IoT (Internet of Things) networks. In decentralized optimization, users (or agents) are represented as nodes of a connected graph $G=(\mathcal{V},\mathcal{E})$ over a finite set of users $\mathcal{V}$ (of cardinality n). The problem considered is then the minimization of

$$f(\mathbf{x}) = \frac{1}{n}\sum_{v\in\mathcal{V}} f_v(\mathbf{x})\,, \quad \mathbf{x}\in\mathbb{R}^d\,, \tag{3}$$

where each f_v is locally held by user $v\in\mathcal{V}$, using only communications between neighboring agents in the graph. There are several known decentralized algorithmic approaches to minimize f under these constrains. The prominent one consists in alternating between communications using gossip matrices (Boyd et al., 2006; Dimakis et al., 2010) and local gradient computations, until a consensus is reached. These gossip approaches suffer from a high synchronization cost (nodes in the graph are required to perform simultaneous

[*]Equal contribution [1]Inria - ENS Paris. Correspondence to: Mathieu Even <mathieu.even@inria.fr>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

operations, or to be aware of operations at the other end of the communication graph) that can be prohibitive if we aim at removing the dependency on a centralized orchestrator. Further, a high number of communications are required to reach consensus, whether all nodes in the graph (as in synchronous gossip) or only two neighboring ones (as in randomized gossip) communicate at each iteration. To alleviate these communication burdens, based on the original works of Lopes & Sayed (2007); Johansson et al. (2007; 2010), we study algorithms based on Markov chain SGD: a variable $\mathbf{x}$ performs a random walk on graph G, and is incrementally updated at each step of the random walk, using the local function available at its location. This approach thus boils down to the one presented above with the function defined in (1), where $\mathcal{V}$ is the (finite) set of agents, π is the uniform distribution over $\mathcal{V}$, and (v_t) is the Markov chain consisting of the consecutive states of the random walk performed on graph G. The random walk guarantees that every communications are spent on updating the global model, as opposed to gossip-based algorithms, where communications are used to reach a running consensus while locally performing gradient steps.

These algorithms are referred to as *token algorithms*: a token (that represents the model estimate) randomly walks the graph and performs updates during its walk. There are two directions to design and analyze token algorithms. Johansson et al. (2007) designed and analyzed its algorithm using, based on SGD with subdifferentials and a Markov chain sampling (consisting of the random walk). Following works (Duchi et al., 2011; Sun et al., 2018) tried to improve convergence guarantees of such stochastic gradient algorithms with Markov chain sampling, under various scenarii (mirror SGD *e.g.*). However, all these analyses rely on overly strong assumption: bounded gradients and/or bounded domains are assumed, and the rates obtained are of the form $\tau_{\rm mix}/T + \sqrt{\tau_{\rm mix}/T}$ for a number T of steps, where $\tau_{\rm mix}$ is the mixing time of the underlying Markov chain. More recently, Dorfman & Levy (2022) obtained similar rates under similar assumptions (bounded losses and gradients), but without requiring any prior knowledge of $\tau_{\rm mix}$, using adaptive stepsizes.

A more recent approach consists in deriving token algorithms from Lagrangian duality and from variants of coordinate gradient methods or ADMM algorithms with Markov chain sampling. Mao et al. (2020) introduce the *Walkman* algorithm, whose analysis works on any graph, and obtain rates of $\frac{\tau_{\rm mix}^2 n}{T}$ to reach approximate-stationary points, while Hendrikx (2022) introduced a more general framework, but whose analysis only works on the complete graph (and is thus equivalent to an *i.i.d.* sampling). Yet, Hendrikx (2022) extend their analysis to arbitrary graph, by performing gradient updates every $\tau_{\rm mix}$ steps of the random walk, obtaining a a dependency on $\tau_{\rm mix} n$, making their algorithm state of the art for these problems. Altenatively, Wang et al. (2022) studies the algorithm stability of MC-SGD in order to derive generalization upper-bounds for this algorithm, and Sun et al. (2022) provides and studies adaptive token algorithms. Recently, and concurrently to this work, Doan (2023) also studies MC-SGD without smoothness; however, their dependency on the mixing time of the random walk (in their Theorem 1) scales as $\exp(c\tau_{\rm mix})$: this is prohibitive as soon as the mixing time becomes larger than $\mathcal{O}(1)$.

In summary, current token algorithms and their analyses either rely on strong noise and regularity assumptions (*e.g.* bounded gradients), or suffer from an overly strong dependency on Markov chain-related quantities (as in (Mao et al., 2020; Hendrikx, 2022)).

The token algorithms we consider are to be put in contrast with *consensus-based decentralized* algorithms, or *gossip* algorithms (with fixed gossip matrices (Dimakis et al., 2010) or with randomized pairwise communications (Boyd et al., 2006)). They originally were introduced to compute the global average of local vectors through peer-to-peer communication. Among the classical decentralized optimization algorithms, some alternate between gossip communications and local steps (Nedic & Ozdaglar, 2009; Koloskova et al., 2019; 2020), others use dual formulations and formulate the consensus constraint using gossip matrices to obtain decentralized dual or primal-dual algorithms (Scaman et al., 2017; Hendrikx et al., 2019; Even et al., 2021a; Kovalev et al., 2021; Alghunaim & Sayed, 2019), and benefit from natural privacy amplification mechanisms (Cyffers et al., 2022). Other approaches include non-symetric communication matrices (Assran & Rabbat, 2021) that are more scalable. We refer the reader to Nedic et al. (2018) for a broader survey on decentralized optimization. The works we relate to in this line of research are Koloskova et al. (2020), where a unified analysis of decentralized SGD is performed (the "gossip equivalent" of our algorithm MC-SGD), and in particular contains rates for convex-non-smooth functions, and Yu et al. (2019), that performs an analysis of decentralized SGD with momentum in the smooth-non-convex case, which is the "gossip equivalent" of our algorithm MC-SAG.

1.2. Reinforcement Learning problems and online system identification

In several applications (RL, time-series analysis *e.g.*), a statistician may have access to values $(X_t)_{t\geqslant 0}$ generated sequentially along the path of a Markov chain, observations from which she wishes to estimate a parameter For instance, Kowshik et al. (2021) consider a sequence of observations $X_{t+1} = A_\star X_t + \xi_t$ for ξ_t *i.i.d.* centered noise, and $A_\star$ to estimate, and aim at finding $\hat{A}$ minimizing the MSE $\mathbb{E}\left[\left\|\sum_{t<T}(\hat{A} - A_\star)X\right\|\right]$ where $X \sim \pi$ is the stationary distribution. Studying this problem under the lens

of stochastic optimization, this boils down to building efficient strategis for SGD under Markov chain sampling, beyond the case of linear mean-squared regressions studied in (Kowshik et al., 2021). While optimal offline policies have extensively been studied in this setting (Jedra & Proutiere, 2019; Simchowitz et al., 2018), online algorithms that take the form of SGD-like algorithms have received little attention, and only focus on the case of quadratic losses with Markov chain as described above. However, these analyses only focus on least squares and Markov chains of the form $X_{t+1} = A_\star X_t + \eta_t$. Under these specific assumptions, Nagaraj et al. (2020) prove that a dependency on τ_{mix} for MC-SGD is inevitable, while using reverse-experince replay, Kowshik et al. (2021) breaks this and obtain sample-optimal online algorithms. Their algorithm however require to store a number of iterates that grow linearly with τ_{mix}.

The convergence guarantees we prove in the sequel for SGD under Markov chain sampling fit in this online framework, and refine previous analyses by removing strong regularity assumptions such as bounded iterates or bounded gradients (Sun et al., 2018), or strong assumptions on the Markovian structure data and least-squares problems (Nagaraj et al., 2020; Kowshik et al., 2021).

Finally, note that in our setting, the iterates of the algorithms considered (denoted as $(\mathbf{x}_k)_{k\geqslant 0}$) and the Markov chain $(v_k)_{k\geqslant 0}$ are dependent of each other. More precisely, $(v_k)_{k\geqslant 0}$ is a Markov chain whose states do not depend on the iterate sequence, while $\mathbf{x}_k$ is $(v_\ell)_{\ell\leqslant k-1}$-measurable. This setting is sometimes referred to as *exogenous Markov noise* (Rust, 1986). Another line of works, pioneered by (Benveniste et al., 1990), considers Markov transitions for (v_k) where $v_{k+1}|v_k$ is sampled using a Markov transition kernel $P_{\mathbf{x}_0,\dots,\mathbf{x}_{k-1}}$ that is directly linked to the iterates. This orthogonal line of work of stochastic approximation with Markovian noise is related to sampling (through the MCMC algorithm), adaptive filtering, and other related problems that involve exploration (Brown & Rutan, 1985; Andrieu et al., 2005; Andrieu & Moulines, 2006; Fort et al., 2016; Blanke & Lelarge, 2023). Our work aims at finding precise rates of convergence as in the convex or non-convex optimization literature (Bubeck, 2015; Carmon et al., 2021), under the mildest assumptions on the exogenous Markov-chain $(v_k)_{k\geqslant 0}$.

Before unrolling our results, we start by reminding definitions and results related to Markov chain theory in Section 2, before presenting our contributions in Section 3

2. Markov chains preliminaries

We refer the interested reader to Levin et al. (2006) for a thorough introduction to Markov chain theory. In this section, we define mixing, hitting and cover times for a Markov chain on a finite space set $\mathcal{V}$ of cardinality n. However, note that these definitions can be extended to the more general setting where $\mathcal{V}$ is infinite (either countable or not). We focus on Markov chain on finite state spaces, but note that the mixing time of a Markov chain can similarly be defined on inifinite state spaces (countable and continuous state spaces). In this paper, all results that involve only the mixing time of the Markov chain (the results from Section 5) easily generalize to infinite state spaces.

Definition 2.1. Let $P \in \mathbb{R}^{\mathcal{V}\times\mathcal{V}}$ be a stochastic matrix (*i.e.* $P_{v,w} \geqslant 0$ for all $v, w \in \mathcal{V}$ and $\sum_{w\in\mathcal{V}} P_{v,w} = 1$ for all $v \in \mathcal{V}$). A time-homogeneous Markov chain on $\mathcal{V}$ of transition matrix P is a stochastic process $(X_t)_{t\geqslant 0}$ with values in $\mathcal{V}$ such that, for any $t \geqslant 0$ and $w, v_0, \dots, v_{t-1}, v \in \mathcal{V}$,

$$\mathbb{P}\left(X_{t+1}= w|X_t=v, X_{t-1}=v_{t-1}, \dots, X_0=v_0\right) = P_{v,w}\,.$$

A Markov chain of transition matrix P is irreducible if, for any $v, w \in \mathcal{V}$, there exists $t \geqslant 0$ such that $(P^t)_{vw} > 0$. A Markov chain of transition matrix P is aperiodic if there exists $t_0 > 0$ such that for all $t \geqslant t_0$ and $v, w \in \mathcal{V}$, $(P^t)_{v,w} > 0$. Any irreducible and aperiodic Markov chain on $\mathcal{V}$ admits a stationary distribution π, that verifies $\pi P = \pi$. It finally holds that, if P is reversible ($\pi_v P_{v,w} = \pi_w P_{w,v}$ for all $v, w \in \mathcal{V}$), denoting as $\lambda_P = 1 - \max_{\lambda\in \mathrm{Sp}(P)\setminus\{1\}} |\lambda| > 0$ the absolute spectral gap of P, where $\mathrm{Sp}(P)$ is the spectrum of P, for any stochastic vector $\pi_0 \in \mathbb{R}^{\mathcal{V}}$: [1]

$$\left\|\pi_0 P^t - \pi\right\|_\pi \leqslant (1-\lambda_P)^t \|\pi_0 - \pi\|_\pi\,.$$

If the chain is not reversible, there is still a linear decay, but in terms of total variation distance rather than in the norm $\|\cdot\|_\pi$ (Chapter 4.3 of Levin et al. (2006)). In the sequel, $(v_t)_{t\geqslant 0}$ is any irreducible aperiodic Markov chain of transition matrix P on $\mathcal{V}$ of stationary distribution π (not necessarily the uniform distribution on $\mathcal{V}$).

Furthermore, we define the graph $G = (\mathcal{V}, \mathcal{E})$ over the state space $\mathcal{V}$ through the relation $\{v, w\} \in \mathcal{E} \iff P_{v,w} > 0$ for v and w two distinct states. Consequently, the Markov chain $(v_t)_t$ can also be seen as a random walk on graph G, with transition probability P. In the random walk decentralized optimization case, this graph coincides with the communication graph. In the sequel, for $t \geqslant 0$ and $v \in \mathcal{V}$, $\mathbb{E}\left[\cdot|v_t = v\right]$ and $\mathbb{P}\left(\cdot|v_t = v\right)$ respectively denote the expectation and probability conditioned on the event $v_t = v$. Similarly for π_t a probability distribution on $\mathcal{V}$, $\mathbb{E}\left[\cdot|v_t \sim \pi_t\right]$ and $\mathbb{P}\left(\cdot|v_t \sim \pi_t\right)$ refers to conditioning on the law of v_t.

Definition 2.2 (Mixing, hitting and cover times)**.** For $w \in \mathcal{V}$, let $\tau_w = \inf\left\{t \geqslant 1 \,|\, v_t = w\right\}$ be the time the chain reaches w (or returns to w, in the case $v_0 = w$). We define the following quantities.

[1] we write $\|z\|_\pi^2 = \sum_{v\in\mathcal{V}} \pi_v z_v^2$ for $z \in \mathbb{R}^{\mathcal{V}}$, and $\|\cdot\|$ always stands for the Euclidean norm

1. *Mixing time.* For $\varepsilon > 0$, the mixing time $\tau_{\text{mix}}(\varepsilon)$ of (v_t) is defined as, where d_{TV} is the total-variation distance:

$$\tau_{\text{mix}}(\varepsilon) = \inf\left\{t \geqslant 1 \,|\, \forall \pi_0\,, \mathrm{d}_{\text{TV}}(P^t\pi_0, \pi) \leqslant \varepsilon\right\},$$

and we define the mixing time τ_{mix} as $\tau_{\text{mix}} = \tau_{\text{mix}}(\pi_{\min}/2)$ [2] where $\pi_{\min} = \min_{v\in\mathcal{V}} \pi_v$.

2. *Hitting and cover times.* The hitting time τ_{hit} and cover time τ_{cov} of (v_t) are defined as:

$$\tau_{\text{hit}} = \max_{(v,w)\in\mathcal{V}^2} \mathbb{E}\left[\tau_w | v_0 = v\right],$$

$$\tau_{\text{cov}} = \max_{v\in\mathcal{V}} \mathbb{E}\left[\max_{w\in\mathcal{V}} \tau_w | v_0 = v\right].$$

The mixing time is the number of steps of the Markov chain required for the distribution of the current state to be close to the stationary probability π. Starting from any arbitrary v_0, the hitting time bounds the time it takes to reach any fixed w, while the cover time bounds the number of steps required to visit all the nodes in the graph.

Note that if the chain is reversible, $\tau_{\text{mix}}(\varepsilon)$ is closely related to λ_P through $\tau_{\text{mix}}(\varepsilon) \leqslant \lceil \lambda_P^{-1} \ln(\pi_{\min}^{-1}\varepsilon^{-1})\rceil$. Under reversibility assumptions, we defined the *relaxation time* of the Markov chain as $\tau_{\text{rel}} = 1/\lambda_P$. More generally without reversibility, $\tau_{\text{mix}}(\varepsilon) \leqslant \lceil \log_2(1/\varepsilon)\rceil \tau_{\text{mix}}(1/4)$. Then, as we prove in Appendix A, τ_{hit} always satisfies $\tau_{\text{hit}} \leqslant 2\pi_{\min}^{-1}\tau_{\text{mix}}$. Finally, using Matthews (1988)' method (detailed in Chapter 11.4 of Levin et al. (2006)), we have $\tau_{\text{cov}} \leqslant \ln(n)\tau_{\text{hit}}$.

3. Contributions

In our paper, we analyze theoretically stochastic gradient methods with Markov chain sampling (such as MC-SGD in Equation (2)), and aim at deriving complexity bounds under the mildest assumptions possible. We first derive in Section 4 complexity lower bounds for such methods, making appear τ_{hit} as the Markov chain quantity that slows down such algorithms.

We then study MC-SGD under various regularity assumptions in Section 5: we remove the bounded gradient assumption of all previous analyses, obtain rates under a μ-PL assumption, and prove a linear convergence in the interpolation regime, where noise and function dissimilarities only need to be bounded at the optimum.

In the data-heterogeneous setting (functions f_v that can be arbitrarily dissimilar) and in the case where $\mathcal{V}$ (the state space of the Markov chain) is finite, we introduce MC-SAG in Section 6, a variance-reduced alternative to MC-SGD, that is perfectly suited to decentralized optimization. Using time adaptive stepsizes, this algorithm has a rate of convergence of τ_{hit}/T and thus matches that of our lower bound, up to acceleration.

We discuss in Section 7 the implications of our results. In particular, we prove that random-walk based decentralization is more communication efficient than consensus-based approaches; prior to our analysis, this was only shown empirically (Mao et al., 2020; Johansson et al., 2010). Further, our results formally prove that using all gradients along the Markov chain trajectory leads to faster rates; as in the previous case, this was only empirically observed before (Sun et al., 2018). These two consequences are derived from the fact that MC-SAG depends only on τ_{hit} rather than the traditionally used quantity $n\tau_{\text{mix}}$, that can be arbitrarily bigger (Table 2).

4. Oracle complexity lower bounds under Markov chain sampling

In this section, we provide oracle complexity lower bounds for finding stationary points of the function f defined in (3), for a class of algorithms that satisfy a "Markov sampling scheme". For a given Markov chain (v_t) on $\mathcal{V}$, we consider algorithms verifying the following procedural constraints, for some fixed initialization $\mathcal{M}_0 = \{\mathbf{x}_0\}$ an then for $t \geqslant 0$,

1. A iteration t, the algorithm has access to function f_{v_t} and may extend its memory:

$$\mathcal{M}_{t+1} = \mathrm{Span}(\{\mathbf{x}\,, \nabla f_{v_t}(\mathbf{x})\,, \quad \mathbf{x} \in \mathcal{M}_t\})\,.$$

2. **Output:** the algorithm specifies an output value $\mathbf{x}_t \in \mathcal{M}_t$.

We call algorithms verifying such constraints "black box procedures with Markov sampling (v_t)". Such procedures as well as the result below are inspired by the distributed black-box procedures defined in Scaman et al. (2017). We use the notation $a(\cdot) = \Omega(b(\cdot))$ for $\exists C > 0$ such that $a(\cdot) \geqslant Cb(\cdot)$ in the theorem below, and classically consider the limiting situation $d \to \infty$, by assuming we are working in $\ell_2 = \left\{(\theta_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}} : \sum_k \theta_k^2 < \infty\right\}$.

Theorem 4.1. *Assume that τ_v (see Definition 2.2) has finite second moment for any $v \in \mathcal{V}$. Let $\Delta, B > 0$, $L > 0$ and $\mu > 0$, denote $\kappa = L/\mu$. Let $\mathbf{x}_0$ be fixed.*

1. Non-convex lower bound: *there exist functions $(f_v)_{v\in\mathcal{V}}$ such that $f = \sum_{v\in\mathcal{V}} \pi_v f_v$ is L-smooth, and $f(\mathbf{x}_0) - \min_{\mathbf{x}} f(\mathbf{x}) \leqslant \Delta$ and such that for any T and any Markov black-box algorithm that outputs $\mathbf{x}_T$ after*

[2] this definition of mixing time is not standard: Levin et al. (2006) define it as $\tau_{\text{mix}}(1/4)$, Mao et al. (2020) define it as we do; however, as explained in Chapter 4.5 of Levin et al. (2006), these definitions are equivalent up to a factor $\ln(1/\pi_{\min})$

T steps, we have:

$$\|\nabla f(\mathbf{x}_T)\|^2 = \Omega\left(L\Delta\left(\frac{\tau_{\text{hit}}}{T}\right)^2\right) .$$

2. Convex lower bound: *there exist functions $(f_v)_{v\in\mathcal{V}}$ such that $f = \sum_{v\in\mathcal{V}} \pi_v f_v$ is convex and L-smooth and minimized at some $\mathbf{x}^\star$ that verifies $\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2 \leqslant B^2$, and such that for any T and any Markov black-box algorithm that outputs $\mathbf{x}_T$ after T steps, we have:*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) = \Omega\left(LB^2\left(\frac{\tau_{\text{hit}}}{T}\right)^2\right) .$$

3. Strongly convex lower bound: *there exist functions $(f_v)_{v\in\mathcal{V}}$ such that $f = \sum_{v\in\mathcal{V}} \pi_v f_v$ is μ-strongly convex and L-smooth and minimized at some $\mathbf{x}^\star$ that verifies $\left\|\mathbf{x}^0 - \mathbf{x}^\star\right\|^2 \leqslant B^2$, and such that for any T and any Markov black-box algorithm that outputs $\mathbf{x}^T$ after T steps, we have:*

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) = \Omega\left(LB^2\exp\left(-\frac{T}{\sqrt{\kappa}\tau_{\text{hit}}}\right)\right) .$$

A complete proof can be found in Appendix B. The hitting time of the Markov chain bounds, starting from any point in $\mathcal{V}$, the mean time it takes to reach any other state in the graph. Making no other assumptions than smoothness, having rates that depend on this hiting time is thus quite intuitive.

5. Analysis of Markov-Chain SGD

We have shown in last subsection that, in order to reach an ε-stationary point with Markov sampling, the optimizer is slowed down by the hitting time of the Markov chain; this lower bound being worst-case on the functions (f_v), we here add additional similarity assumptions, that are still milder than classical ones in this setting (Sun et al., 2018).Studying the iterates generated by (2), we obtain in this section a dependency on τ_{mix}, provided bounded gradient dissimilarities (Assumptions 5.1 and 5.4).

We here assume that $(v_t)_{t\geqslant 0}$ is a Markov chain on $\mathcal{V}$ of invariant probability π (not necessarily the uniform measure on $\mathcal{V}$). In this section, the function f studied is defined as

$$f(\cdot) = \mathbb{E}_{v\sim\pi}[f_v(\cdot)] ,$$

as in (1). Consequently, for the MC-SGD algorithm for decentralized optimization over a given graph G to minimize the averaged function over all nodes (as in (3)), π needs to be the uniform probability over $\mathcal{V}$.

We first derive convergence rates under smoothness assumptions with or without a μ-PL inequality that holds, before improving our results under strong convexity assumptions, under which we prove a linear convergence rate in the interpolation regime. We finally add local noise (due to sampling, or additive gaussian noise to enforce privacy) in the final paragraph of this Section.

5.1. Analysis under bounded gradient dissimilarities

Assumption 5.1. There exists $(\sigma_v^2)_{v\in\mathcal{V}}$ such that for all $v \in \mathcal{V}$ and all $\mathbf{x} \in \mathbb{R}^d$, we have[3]:

$$\|\nabla f_v(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leqslant \sigma_v^2 ,$$

and we denote $\bar{\sigma}^2 = \mathbb{E}_{v\sim\pi}\left[\sigma_v^2\right]$ and $\sigma_{\max}^2 = \max_{v\in\mathcal{V}} \sigma_v^2$.

Assumption 5.2. Each f_v is L-smooth, f is lower bounded, its minimum is attained at some $\mathbf{x}^\star \in \mathbb{R}^d$.

Theorem 5.3 (MC-SGD)**.** *Assume that Assumptions 5.1 and 5.2 hold, and let $\Delta \geqslant f(\mathbf{x}_0) - f(\mathbf{x}^\star) + \sigma_{\max}^2/L$.*

1. *For a constant time-horizon dependent step size γ (i.e., γ is a functiion of T), the iterates generated by Equation (2) satisfy, for $T \geqslant 2\tau_{\text{mix}}\ln(\tau_{\text{mix}})$:* [4]

$$\mathbb{E}\|\nabla f(\hat{\mathbf{x}}_T)\|^2 = \tilde{\mathcal{O}}\left(\frac{\Delta L\tau_{\text{mix}}}{T} + \frac{\sqrt{L\Delta\bar{\sigma}^2\tau_{\text{mix}}} + \bar{\sigma}^2}{\sqrt{T}}\right) ,$$

where $\hat{\mathbf{x}}_T$ is drawn uniformly at random amongst $\mathbf{x}_0, \dots, \mathbf{x}_{T-1}$.

2. *If f additionally verifies a μ-PL inequality (for any $\mathbf{x} \in \mathbb{R}^d$, $\|\nabla f(\mathbf{x})\|^2) \geqslant 2\mu(f(\mathbf{x}) - f(\mathbf{x}^\star))$), for a constant time-horizon dependent step size γ, the iterates generated by Equation (2) satisfy, for $T \geqslant 2\tau_{\text{mix}}\ln(\tau_{\text{mix}})$ a numerical constant $c > 0$, and $\kappa = L/\mu$, with $F_T = \mathbb{E}\left[f(\mathbf{x}_T) - f(\mathbf{x}^\star)\right]$:*

$$F_T \leqslant e^{-\frac{cT}{\kappa\tau_{\text{mix}}\ln(T)}}\Delta + \tilde{\mathcal{O}}\left(\frac{\tau_{\text{mix}}\bar{\sigma}^2}{\mu T}\right) ,$$

Theorem 5.3 is proved in Appendix C, by enforcing a delay of order τ_{mix} and relying on recent analyses of delayed SGD and SGD with biased gradients. As explained in the introduction, removing the bounded gradient assumption present in previous works (Johansson et al., 2010; Sun et al., 2018; Duchi et al., 2011) that study Markov chain SGD (in the mirror setting, or with subdifferentials), and replacing it by a much milder and classical assumption of bounded gradient dissimilarities (Karimireddy et al., 2020), we thus still managed to obtain similar rates. Further, if f verifies a μ-PL inequality (if for any $\mathbf{x} \in \mathbb{R}^d$, $\|\nabla f(\mathbf{x})\|^2) \geqslant 2\mu(f(\mathbf{x}) - f(\mathbf{x}^\star))$), we have an almost-linear rate of convergence: this is the first rate under μ-PL or strong convexity assumptions for MC-SGD-like algorithms, that we even refine further in next subsection.

[3] this assumption could be replaced by a more relaxed noise assumption of the form $\|\nabla f_v(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leqslant M\|\nabla f(\mathbf{x})\|^2 + \sigma_v^2$

[4] $\tilde{\mathcal{O}}$ hides logarithmic factors

5.2. Tight rates and linear convergence in the interpolation regime

We now study MC-SGD under the following assumptions, to derive faster rates, that only depend on the sampling noise at the optimum. The interpolation regime – often related to overparameterization – refers to the case where there exists a model $\mathbf{x}^\star \in \mathbb{R}^d$ minimizing all f_v for $v \in \mathcal{V}$, leading to $\sigma_\star^2 = 0$ in Assumption 5.5, and to a linear convergence rate below.

Assumption 5.4. Functions f_v are L-smooth and μ-strongly convex. We denote $\kappa = L/\mu$.

Assumption 5.5 (Noise at the optimum)**.** Let $\mathbf{x}^\star$ be a minimizer of f. We assume that for some $\sigma_\star \geqslant 0$, we have for all $v \in \mathcal{V}$:

$$\|\nabla f_v(\mathbf{x}^\star)\|^2 \leqslant \sigma_\star^2 \,.$$

Theorem 5.6 (Unified analysis)**.** *Under Assumptions 5.4 and 5.5, the sequence generated by (2) satisfies, if $\gamma L < 1$:*

$$\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}^\star\|^2\right] \leqslant 2(1-\gamma\mu)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + 2\frac{\gamma^3 T L}{\mu} \sum_{0 \leqslant s \leqslant T} (1-\gamma\mu)^{T-s} \mathbb{E}\left[\left\|\sum_{s \leqslant t < T} \nabla f_{v_t}(\mathbf{x}^\star)\right\|^2\right] .$$

In the interpolation regime, $\nabla f_v(\mathbf{x}^\star) = 0$ for all $v \in \mathcal{V}$, so that:

$$\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}^\star\|^2\right] \leqslant 2(1-\gamma\mu)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \,.$$

The result in Theorem 5.6 is in fact true irrespectively of the sequence (v_t) chosen: it does not require (v_t) to specifically be a Markov chain. This property is used in the next Corollary, that also highlights the fact that by studying distance to the optimum, a condition number is lost in the process. This is the case in many previous analyses of other different algorithms (*e.g.*, Bregman/Mirror-SGD (Dragomir et al., 2021) or SGD with random-resfhuffling (Mishchenko et al., 2020), which is in fact a particular instance of MC-SGD, that our analysis recovers), that study distances to the optimum (with respect to some mirror map, in the case of Mirror SGD), and therefore obtain an extra κ factor in the noise term. Theorem 5.6 is proved by generalizing the proof technique of (Mishchenko et al., 2020) to arbitrary orderings and for unbounded time horizons.

Remark 5.7 (Random resfhuffling). A special case of Theorem 5.6 is SGD with random reshuffling. By analyzing SGD with random-reshuffling as SGD with a Markovian ordering (on an extended state space), Theorem 1,2 also recover rates for SGD with random reshuffling for which we have $\tau_{\text{mix}} = n$. Moreover, since Theorem 5.6 generalizes Theorem 1 of (Mishchenko et al., 2020), we also recover their rate as a special case by bounding each term $\mathbb{E}\left[\left\|\sum_{s \leqslant t < T} \nabla f_{v_t}(\mathbf{x}^\star)\right\|^2\right]$.

We specify Theorem 5.6 under a Markovian sampling scheme in next corollary: the noise term at the optimum takes the form τ_{mix}/T.

Corollary 5.8 (MC-SGD, interpolation)**.** *Under Assumptions 5.4 and 5.5, for $T \geqslant 1$ and for a well chosen stepsize $\gamma > 0$, the iterates generated by (2) satisfy:*

$$\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}^\star\|^2\right] \leqslant 2e^{-\frac{T}{\kappa}} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \tilde{\mathcal{O}}\left(\frac{L\tau_{\text{mix}}(\frac{1}{4})\sigma_\star^2}{\mu^3 T}\right) .$$

This result is stronger than Theorem 5.3.2, for *(i)* noise amplitude and gradient dissimilarities only need to be bounded at the optimum; *(ii)* the "optimization term" (the first one) is not slowed down by the mixing time. This comes at the cost of strong convexity assumptions, stronger than a μ-PL inequality for f. The term $\frac{\tau_{\text{mix}}\sigma_\star}{T}$ cannot be removed in the general case, as next proposition shows. Hence, since the two other terms have optimal dependency in terms of Markov-chain and noise related quantities, our analysis ends up being sharp.

Corollary 5.6 together with the following proposition are an extension of Nagaraj et al. (2020), who proved similar results for MC-SGD with constant stepsize on least square problems on Markovian data of a certain form (for linear online system identification).

Proposition 5.9. *For any $\mathcal{V}$ (such that $|\mathcal{V}| \geqslant 2$) and $\tau > 1$, there exists a Markov chain on $\mathcal{V}$ of relaxation time τ, functions $(f_v)_{v\in\mathcal{V}}$ and $\mathbf{x}_0 \in \mathbb{R}^d$ such that given any stepsize γ, the iterates of Equation (2) output $\mathbf{x}_T$ for any $T > 0$ verifying $\|\mathbf{x}_T - \mathbf{x}_0\|^2 = \tilde{\Omega}(\tau\sigma_\star^2/T)$, and the assumptions of Theorem 5.6 hold.*

5.3. MC-SGD with local noise

In the two previous subsections, we analyzed SGD with Markovian sampling schemes, where the stochasticity only came from the Markov chain $(v_k)_{k\geqslant 0}$. We now generalize the analysis and results to SGD with both Markovian sampling, and local noise, by studying the sequence:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \mathbf{g}_t \,. \tag{4}$$

We now formulate the form stochastic gradients $\mathbf{g}_t$ can take.

Assumption 5.10. For all $v \in \mathcal{V}$, the function f_v satisfies $f_v(\mathbf{x}) = \mathbb{E}\left[F_v(\mathbf{x}, \xi_v)\right]$ for all $\mathbf{x} \in \mathbb{R}^d$, where $\xi_v \sim \mathcal{D}_v$. Furthermore, there exists a Markov-chain $(v_t)_{t\geqslant 0}$ such that for all $t \geqslant 0$,

$$\mathbf{g}_t = \nabla_{\mathbf{x}} F_{v_t}(\mathbf{x}_t, \xi_t) \,,$$

where $\xi_t \sim \mathcal{D}_{v_t} | v_t$ is independent from $v_0, \ldots, v_{t-1}$ and $\xi_0, \ldots, \xi_{t-1}$.

A direct consequence of Assumption 5.4 is that $\mathbb{E}\left[\mathbf{g}_t|\mathbf{x}_t, v_t\right] = \nabla f_{v_t}(\mathbf{x}_t)$. Two main applications of Assumption 5.4 are:

1. **Local sampling.** If $f_v(\mathbf{x}) = \frac{1}{m}\sum_{i=1}^m f_{v,i}(\mathbf{x})$ (agent v has m local samples), agent m may use only a batch $\mathcal{B} \subset [m]$ of its samples, leading to stochastic gradients $\mathbf{g}_t$ in (4) of the form:

$$\mathbf{g}_t = \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t} \nabla f_{v_t,i}(\mathbf{x}_t)\,,$$

for random batches $(\mathcal{B}_t)_{t\geqslant 0}$.

2. **Differential privacy.** Adding local noise (*e.g.*, additive Gaussian random noise) enforces differential privacy under suitable assumptions. A private decentralized token algorithm is then **Differentially Private MC-SGD** (DP-MC-SGD), with iterates (4) where $\mathbf{g}_t$ satisfies

$$\mathbf{g}_t = \nabla f_{v_t}(\mathbf{x}_t) + \eta_t\,, \tag{5}$$

where (v_t) is the Markov chain (random walk performed by the token on the communication graph), and $\eta_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$ is sampled independently from the past, to enforce differential privacy.

Under Assumption 5.4, a direct generalization of Theorem 5.6 and Corollary 5.8 is the following.

Theorem 5.11 (MC-SGD with local noise)**.** *Assume that Assumptions 5.10,5.5 holds, each $F_v(\cdot,\xi)$ is μ-strongly convex L-smooth, and there exists $\sigma^2_{\rm local}$ such that:*

$$\mathbb{E}\left[\|\mathbf{g}_t - \nabla f_{v_t}(\mathbf{x}_t)\|^2|\mathbf{x}_t, v_t\right] \leqslant \sigma^2_{\rm local}\,.$$

Then, for a well chosen $\gamma > 0$, the iterates generated by (4) satisfy:

$$\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}^\star\|^2\right] \leqslant 2e^{-\frac{T}{\kappa}}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \tilde{\mathcal{O}}\left(\frac{L}{\mu^3 T}\big(\sigma^2_{\rm local} + \tau_{\rm mix}\big(\frac{1}{4}\big)\sigma^2_\star\big)\right)\,.$$

Importantly, and as one would have expected, local noise is **not** impacted by the mixing time of the underlying random walk. While we did not pursue in this direction, this observation could easily be made under other regularity assumptions, and such a result would hold for instance under the assumptions of Theorem 1 or 2. While Differentially Private MC-SGD sounds appealing for performing decentralized and differentially private optimization, we here only provided a utility analysis, the privacy analysis being left for future works.

Algorithm 1 Markov Chain SAG (MC-SAG)

Input: $\mathbf{x}_0 \in \mathbb{R}^d$, $\mathbf{h}_v \in \mathbb{R}^d$ for $v \in \mathcal{V}$ and $\bar{\mathbf{h}}_0 \in \mathbb{R}^d$, stepsizes $\gamma_t > 0$, $v_0 \in \mathcal{V}$
for $t = 0, 1, \dots$ **do**
  Compute $\nabla f_{v_t}(\mathbf{x}_t)$
  $\bar{\mathbf{h}}_{t+1} = \bar{\mathbf{h}}_t + \frac{1}{n}\big(\nabla f_{v_t}(\mathbf{x}_t) - \mathbf{h}_{v_t}\big)$
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \bar{\mathbf{h}}_{t+1}$
  $\mathbf{h}_{v_t} \longleftarrow \nabla f_{v_t}(\mathbf{x}_t)$
  Sample $v_{t+1} \sim P_{v_t,\cdot}$
end for

6. Analysis of Markov-Chain SAG

After providing convergence guarantees for the most natural algorithm (MC-SGD) under a Markov chain sampling on the set $\mathcal{V}$, we prove that one can achieve a rate of order $1/T$ (rather than the $1/\sqrt{T}$ previously obtained) in the smooth setting, by applying the variance reduction techniques present in Schmidt et al. (2017), that first introduced the Stochastic Averaged Gradient algorithm, together with a time-adaptive stepsize policy described below. Our faster rate with variance reduction leads of a dependency on $\tau_{\rm hit}$ instead of $\tau_{\rm mix}$; since we do not make any other assumption other than smoothness, this is unavoidable in light of our lower bound (Theorem 4.1).

MC-SAG The MC-SAG algorithm is described in Algorithm 1. The recursion leading to the iterate $\mathbf{x}_t$ can then be summarized as, for stepsizes $(\gamma_t)_{t\geqslant 0}$, under the initialization $\mathbf{h}_v = \nabla f_v(\mathbf{x}_0)$ and $\bar{\mathbf{h}} = \nabla f(\mathbf{x}_0)$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\gamma_t}{n}\sum_{v\in\mathcal{V}} \nabla f_v(\mathbf{x}_{d_v(t)})\,, \tag{6}$$

where for $v \in \mathcal{V}$, we define $d_v(t) = \sup\{s \leqslant t \,|\, v_s = v\}$ as the last previous iterate at which v was the current state of the Markov chain. By convention, if the set over which the supremum is taken is empty, we set $d_v(t) = 0$. We handle both the initialization described just above for $\mathbf{h}_v, \bar{\mathbf{h}}$ and arbitrary initialization in our analysis below.

In the same way that MC-SGD reduces to vanilla SGD if (v_t) is an *i.i.d.* uniform sampling over $\mathcal{V}$, MC-SAG boils down to the SAG algorithm (Schmidt et al., 2017) in that case and under the initialization $\mathbf{h}_v = \nabla f_v(\mathbf{x}_0)$ and $\bar{\mathbf{h}} = \nabla f(\mathbf{x}_0)$. In a decentralized setting, nodes keep in mind their last gradient computed (variable $\mathbf{h}_v$ at node v). At all times, $\bar{\mathbf{h}}_t$ is an average of these $\mathbf{h}_v$ over the graph, and is, in the same way as $\mathbf{x}_t$, updated along the random walk. The MC-SAG algorithm is thus perfectly adapted to decentralized optimization.

Time-adaptive stepsize policy To obtain our convergence guarantees, a time-adaptive stepsize policy (γ_t) is used, as

in Asynchronous SGD (Mishchenko et al., 2022) to obtain delay-independent guarantees. For $t \geqslant 0$, let the stepsize γ_t be defined as:

$$\gamma_t = \frac{1}{2L\big(\tau_{\text{hit}} + \max_{v\in\mathcal{V}}(t - d_v(t))\big)} \,. \tag{7}$$

Denoting $\tau_t = \max_{v\in\mathcal{V}}(t - d_v(t))$, this quantity can be tracked down during the optimization process. Indeed, if agent v_t receives τ_{t-1} together with $(\mathbf{x}_t, \bar{\mathbf{h}}_t)$, she may compute τ_t as:

$$\tau_t = \max\left(\tau_{t-1} + 1\,,\, t - d_{v_t}(t)\right),$$

where $t - d_{v_t}(t)$ is the number of iterations that took place since the last time the Markov chain state was v_t. Hence, if agents keep track of the number of iterations, the adaptive stepsize policy (7) can be used in Algorithm 1, as long as agent v_t sends (τ_t, t) to v_{t+1}, yielding the following result.

We now present the convergence results for MC-SAG. (v_t) is in this section assumed to be a Markov chain on $\mathcal{V}$ of finite hitting time τ_{hit}. Importantly, the next Theorem does not require any additional assumption on (v_t) such as reversibility, or even that it has a stationary probability that is the uniform distribution: the non-symmetric but easily implementable transition probabilities $P_{v,w} = 1/(d_v + 1)$ for $w = v$ or $w \sim v$ can be used here, as well as non-reversible random walks than can have much smaller mixing and hitting times. The function f studied is here independent of the Markov chain, and is defined as in (3), the uniformly averaged function over all states $v \in \mathcal{V}$ (or over all agents in the network).

Theorem 6.1 (MC-SAG)**.** *Assume that Assumption 5.2 holds and that the Markov chain has a finite hitting time (for an arbitrary invariant probability).*

1. *Under the initialization:*

$$\mathbf{h}_v = \nabla f_v(\mathbf{x}_0)\,,$$
$$\bar{\mathbf{h}} = \nabla f(\mathbf{x}_0)\,,$$

using the adaptive stepsize policy defined in Equation (7), the sequence generated by Algorithm 1 satisfies, for any $T > 0$:

$$\mathbb{E}\left[\min_{t<T} \|\nabla f(\mathbf{x}_t)\|^2\right] \leqslant 8L\big(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\big)\frac{\tau_{\text{hit}}\ln(n)}{T}\,.$$

2. *Under any arbitrary initialization that satisfies $\bar{\mathbf{h}} = \frac{1}{n}\sum_{v\in\mathcal{V}}\mathbf{h}_v$, using the adaptive stepsize policy defined in Equation (7), the sequence generated by Algorithm 1 satisfies, for any $T > 0$:*

$$\mathbb{E}\left[\min_{t<T} \|\nabla f(\mathbf{x}_t)\|^2\right] \leqslant 16L\Delta\frac{\tau_{\text{hit}}\ln(n)}{T}\,,$$

where

$$\Delta = f(\mathbf{x}_0) - f(\mathbf{x}^\star) + \frac{1}{8n}\sum_{v\in\mathcal{V}}\|\nabla f_v(\mathbf{x}_0) - \mathbf{h}_v\|^2\,.$$

Theorem 6.1 is proved in Appendix 6.1. Up to the logarithmic factor in n, the rates in Theorem 6.1 are the non-accelerated versions of the lower-bound in Theorem 4.1.

7. Discussion of our results

7.1. Communication efficiency: comparison of our results with consensus-based approaches

We summarize the communication efficiencies in Table 1 (in terms of total number of communications required to reach an ε-stationary point), of classical gossip-based decentralized gradient methods (non-accelerated, since no accelerated method is known under our regularity assumptions). We consider the algorithm of (Yu et al., 2019) (decentralized SGD with momemtum, state of the art decentralized gossip-based algorithm for this problem) with fixed communication matrix W on the graph G together with the Walkman algorithm (Mao et al., 2020) and our algorithms, for a Markov chain with transition matrix P. For the sake of comparison, we take as gossip matrix $W = P$.Consequently as shown in Table 1, *our algorithm (MC-SAG) always outperforms non-accelerated gossip-based decentralized gradient descent algorithms in terms of number of communications required to reach ε-stationary points*. Note that we do not claim the "overall superiority" of our approach over classical decentralized optimization algorithms (the latter benefit from parallelization while ours do not), but a superiority in terms of communication efficiency.

Table 1: Number of communications required (# comm. below) to obtain an ε-stationary point $\mathbf{x}$ (verifying $\|\nabla f(\mathbf{x})\|^2 \leqslant \varepsilon$). Logarithmic/constant factors hidden.

	A	B	Our work
# comm.	$\varepsilon^{-1}\lvert\mathcal{E}\rvert\tau_{\text{mix}}$	$\varepsilon^{-1}n\tau_{\text{mix}}^2$	$\varepsilon^{-2}\tau_{\text{mix}}$ [c] $\varepsilon^{-1}\tau_{\text{hit}}$ [d]

A: (Yu et al., 2019).
B: (Mao et al., 2020).
[c] MC-SGD under Assumption 5.1.
[d] MC-SAG.

The dependency on the quantity τ_{hit} we obtain (under no other assumptions than smoothness) is always better than the dependency on $n\tau_{\text{mix}}$ of previous works (using gossip communications or a random walker), since $\tau_{\text{hit}} \leqslant 2n\tau_{\text{mix}}$ always holds. As illustrated in Table 2 on some known graphs, this inequality is rather loose when the connectivity decreases (*i.e.* the mixing time increases), so that the speedup our results lead to is even more effective on ill-connected graphs; the difference between the two can scale

Table 2: Hitting and mixing times of some known graphs, for the simple random walk.

	Cycle	d-dim. torus	Complete graph
τ_{hit}	$\mathcal{O}(n^2)$	$\mathcal{O}(n^{1+\frac{1}{d}})$	$\mathcal{O}(n)$
$n\tau_{\text{mix}}$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^{1+\frac{2}{d}})$	$\mathcal{O}(n)$

up to a factor n. In fact, we prove in Appendix A that for d-regular and symetric graphs, we have:

$$\tau_{\text{hit}} \leqslant \frac{2|\mathcal{E}|\text{Diam}(G)}{d} ,$$

where Diam is the diameter of G. The dependency $n\tau_{\text{mix}}^2$ obtained in (Mao et al., 2020) (the only work that does not make bounded gradient assumptions) is prohibitive when graph connectivity decreases (n^3 on the grid, n^5 on the cycle). Our analysis does not rely on a reversibility assumption of the Markov chain, so that non symetric random walks can be used, therefore accelerating mixing; on the cycle for a non-symmetric random walk for instance, the hitting time decreases to $\mathcal{O}(n)$.

7.2. Using all gradient along the trajectory of (v_t) is provably more efficient

Sun et al. (2018) empirically motivated through empirical evidence the use of all gradients ∇f_{v_t} sampled along the trajectory of the Markov chain rather than waiting for the chain to mix before every stochastic gradient step in order to mimic the behavior of vanilla SGD. However, their rates (as well as those of (Johansson et al., 2010; Duchi et al., 2011) and ours for MC-SGD) are functions of $S = T/\tau_{\text{mix}}$, and of order $1/S + 1/\sqrt{S}$. These are exactly what one would obtain, by waiting for τ_{mix} steps of the chain in order to have an approximate uniform sampling before each update! Consequently, there are no theoretical ground or evidence for using all the gradients along the trajectory of the Markov chain with these results, other than by doing so, one does not do worse than waiting for the chain to mix to mimic vanilla SGD. This is exactly the approach taken by Hendrikx (2022): a gradient step is performed every τ_{mix} random walk steps. This is where MC-SAG and its guarantees that depend on T/τ_{hit} come in place. Under our assumptions, the rate of SAG for finding approximate stationary points when waiting for the chain to mix before using a stochastic gradient is of order $n/S = n\tau_{\text{mix}}/T$ where $S = T/\tau_{\text{mix}}$ is the number of stochastic gradients used. We obtain τ_{hit}/T instead: hence, in cases where $\tau_{\text{hit}} = o(n\tau_{\text{mix}})$, using all stochastic gradients along the trajectory of the Markov chain - instead of waiting for mixing before performing a stochastic gradient step - provably helps. Hence, we here provided a realistic scenario where using all stochastic gradients proves to accelerate the rate; this was previously noticed in another setting with RER-SGD (SGD with reverse-experience replay, (Kowshik et al., 2021)).

7.3. Running-time complexity and robustness to "stragglers"

The total time it takes to run random walk-based decentralized algorithms depends on $T_{v\to w}$, the time it takes to compute a gradient at v, and then the communication time to send it to w. Using ergodicity of the Markov chain, the time $\text{time}^{\text{MC}}(T)$ it takes to run MC-SAG or MC-SGD for T iterations verifies:

$$\frac{\text{time}^{\text{MC}}(T)}{T} \to \sum_{(v,w)\in\mathcal{V}^2} \pi_v P_{v,w} T_{v\to w} ,$$

where the limit is a weighted sum of the local computation/communication times, with weights summing to 1. Random-walk based decentralized algorithms are therefore robust to slow edges or nodes ("stragglers"), a property that synchronous gossip algorithms do not verify (their time complexity depends on $\max_{v,w} T_{v\to w}$), while studying asynchronous gossip is notoriously difficult (Even et al., 2021b).

Numerical illustration of our theory We present in Appendix G two experiments on synthetic problems, comparing MC-SAG and MC-SGD to gossip-based and token baselines. We consider two settings (a well-connected graph with homogeneous functions, an ill-connected graph with heterogeneous functions) in an effort to illustrate how these two difficulties (graph connectivity and data-heterogeneity) are both bypassed by the MC-SAG algorithm.

Conclusion Without variance reduction and under bounded data-heterogeneity assumptions, SGD under MC sampling is slowed down by a factor τ_{mix}, due to increased sampling variance. Using variance-reduction techniques, we obtain faster rates, that depend on τ_{hit} rather than $n\tau_{\text{mix}}$, which one would have expected by directly extending known results in the *i.i.d.* setting to our MC sampling schemes. Leveraging such a dependency yields a fast token algorithm (MC-SAG), robust to both ill-connectivity of the graph and data-heterogeneity.

Aknowledgments I deeply thank Hadrien Hendrikx for many interesting discussions on the subject and for all his help in writing this paper, as well as Anastasiia Koloskova and Edwige Cyffers for helpful discussions.

References

Alghunaim, S. A. and Sayed, A. H. Linear convergence of primal-dual gradient methods and their performance in distributed optimization, 2019.

Andrieu, C. and Moulines, É. On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16(3), August 2006.

Andrieu, C., Moulines, É., and Priouret, P. Stability of stochastic approximation under verifiable conditions. *SIAM Journal on Control and Optimization*, 44(1):283–312, January 2005.

Assran, M. S. and Rabbat, M. G. Asynchronous gradient push. *IEEE Transactions on Automatic Control*, 66(1): 168–183, 2021.

Benveniste, A., Métivier, M., and Priouret, P. *Adaptive Algorithms and Stochastic Approximations*. Springer Berlin Heidelberg, 1990.

Blanke, M. and Lelarge, M. Flex: an adaptive exploration algorithm for nonlinear systems. In *International Conference on Machine Learning*, 2023.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, January 2018.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

Brown, S. and Rutan, S. Adaptive kalman filtering. *Journal of Research of the National Bureau of Standards*, 90(6): 403, November 1985.

Bubeck, S. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn.*, 8(3–4):231–357, November 2015.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points ii: First-order methods. *Math. Program.*, 185(1–2):315–355, jan 2021. ISSN 0025-5610.

Cyffers, E., Even, M., Bellet, A., and Massoulié, L. Muffliato: Peer-to-peer privacy amplification for decentralized optimization and averaging. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Dimakis, A. G., Kar, S., Moura, J. M. F., Rabbat, M. G., and Scaglione, A. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.

Doan, T. T. Finite-time analysis of markov gradient descent. *IEEE Transactions on Automatic Control*, 68(4):2140–2153, 2023.

Dorfman, R. and Levy, K. Y. Adapting to mixing time in stochastic optimization with Markovian data. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5429–5446. PMLR, 17–23 Jul 2022.

Dragomir, R. A., Even, M., and Hendrikx, H. Fast stochastic bregman gradient methods: Sharp analysis and variance reduction. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2815–2825. PMLR, 18–24 Jul 2021.

Duchi, J. C., Agarwal, A., Johansson, M., and Jordan, M. I. Ergodic mirror descent. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 701–706, 2011. doi: 10.1109/Allerton.2011.6120236.

Even, M., Berthier, R., Bach, F., Flammarion, N., Hendrikx, H., Gaillard, P., Massoulié, L., and Taylor, A. Continuized accelerations of deterministic and stochastic gradient descents, and of gossip algorithms. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28054–28066. Curran Associates, Inc., 2021a.

Even, M., Hendrikx, H., and Massoulié, L. Decentralized optimization with heterogeneous delays: a continuous-time approach. Technical report, arXiv:2106.03585, 2021b.

Even, M., Massoulié, L., and Scaman, K. On sample optimality in personalized collaborative and federated learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Fort, G., Moulines, E., Schreck, A., and Vihola, M. Convergence of markovian stochastic approximation with discontinuous dynamics. *SIAM Journal on Control and Optimization*, 54(2):866–893, January 2016.

Hendrikx, H. A principled framework for the design and analysis of token algorithms, 2022.

Hendrikx, H., Bach, F., and Massoulié, L. An accelerated decentralized stochastic proximal algorithm for finite sums. In *Advances in Neural Information Processing Systems*, 2019.

Jedra, Y. and Proutiere, A. Sample complexity lower bounds for linear system identification. In *IEEE Conference on decision and control (CDC) 2019*, pp. 2676–2681, 12 2019.

Johansson, B., Rabi, M., and Johansson, M. A simple peer-to-peer algorithm for distributed optimization in sensor networks. In *2007 46th IEEE Conference on Decision and Control*, pp. 4705–4710, 2007. doi: 10.1109/CDC.2007.4434888.

Johansson, B., Rabi, M., and Johansson, M. A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3):1157–1170, 2010.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for federated learning. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5132–5143. PMLR, 13–18 Jul 2020.

Koloskova, A., Stich, S., and Jaggi, M. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, volume 97, pp. 3478–3487. PMLR, 2019.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized SGD with changing topology and local updates. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5381–5393. PMLR, 13–18 Jul 2020.

Kovalev, D., Gasanov, E., Gasnikov, A., and Richtarik, P. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22325–22335. Curran Associates, Inc., 2021.

Kowshik, S., Nagaraj, D., Jain, P., and Netrapalli, P. Streaming linear system identification with reverse experience replay. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30140–30152. Curran Associates, Inc., 2021.

Levin, D. A., Peres, Y., and Wilmer, E. L. *Markov chains and mixing times*. American Mathematical Society, 2006.

Lopes, C. G. and Sayed, A. H. Incremental adaptive strategies over distributed networks. *IEEE Transactions on Signal Processing*, 55(8):4064–4077, 2007. doi: 10.1109/TSP.2007.896034.

Mania, H., Pan, X., Papailiopoulos, D., Recht, B., Ramchandran, K., and Jordan, M. I. Perturbed iterate analysis for asynchronous stochastic optimization. *SIAM Journal on Optimization*, 27(4):2202–2229, January 2017.

Mao, X., Yuan, K., Hu, Y., Gu, Y., Sayed, A. H., and Yin, W. Walkman: A communication-efficient random-walk algorithm for decentralized optimization. *IEEE Transactions on Signal Processing*, 68:2513–2528, 2020. doi: 10.1109/TSP.2020.2983167.

Matthews, P. Covering Problems for Markov Chains. *The Annals of Probability*, 16(3):1215 – 1228, 1988.

Mei, S., Bai, Y., and Montanari, A. The landscape of empirical risk for nonconvex losses. *The Annals of Statistics*, 46(6A):2747–2774, 2018.

Mishchenko, K., Khaled, A., and Richtarik, P. Random reshuffling: Simple analysis with vast improvements. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17309–17320. Curran Associates, Inc., 2020.

Mishchenko, K., Bach, F., Even, M., and Woodworth, B. Asynchronous SGD beats minibatch SGD under arbitrary delays. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Nagaraj, D., Wu, X., Bresler, G., Jain, P., and Netrapalli, P. Least squares regression with markovian data: Fundamental limits and algorithms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 16666–16676. Curran Associates, Inc., 2020.

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

Nedic, A., Olshevsky, A., and Rabbat, M. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5): 953–976, May 2018.

Nesterov, Y. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

Rao, S. Finding hitting times in various graphs, 2012.

Rust, J. Structural estimation of markov decision processes. In Engle, R. F. and McFadden, D. (eds.), *Handbook of Econometrics*, volume 4, chapter 51, pp. 3081–3143. Elsevier, 1 edition, 1986.

Scaman, K., Bach, F., Bubeck, S., Lee, Y. T., and Massoulié, L. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3027–3036. PMLR, 06–11 Aug 2017.

Schmidt, M., Le Roux, N., and Bach, F. Minimizing finite sums with the stochastic average gradient. *Math. Program.*, 162(1–2):83–112, mar 2017. ISSN 0025-5610.

Simchowitz, M., Mania, H., Tu, S., Jordan, M. I., and Recht, B. Learning without mixing: Towards a sharp analysis of linear system identification. In Bubeck, S., Perchet, V., and Rigollet, P. (eds.), *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 439–473. PMLR, 06–09 Jul 2018.

Stich, S. U. and Karimireddy, S. P. The error-feedback framework: Better rates for sgd with delayed gradients and compressed communication, 2021.

Sun, T., Sun, Y., and Yin, W. On markov chain gradient descent. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Sun, T., Li, D., and Wang, B. Adaptive random walk gradient descent for decentralized optimization. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20790–20809. PMLR, 17–23 Jul 2022.

Wang, P., Lei, Y., Ying, Y., and Zhou, D.-X. Stability and generalization for markov chain stochastic gradient methods. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022.

Yu, H., Jin, R., and Yang, S. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7184–7193. PMLR, 09–15 Jun 2019.

A. Preliminary results

A.1. Mixing time and relaxation time, mixing time and hitting time

We first begin by the two following lemmas. The first one is very classical, and bounds the mixing time in terms of $1/\lambda_P$, in the case where the chain is reversible; we provide a proof for completeness. Note that if the chain is reversible, *we still have a linear decay* (Levin et al., 2006). The second lemma we provide bounds the hitting time of the Markov chain with the mixing time. This result is somewhat less classical, and is not present in the classical Markov chain literature surveys.

Lemma A.1 ($\tau_{\rm mix}$ and λ_P)**.** *For any $\varepsilon > 0$, if the chain is reversible:*

$$\tau_{\rm mix}(\varepsilon) \leqslant \left\lceil \frac{1}{\lambda_P} \ln(\varepsilon^{-1}\pi_{\min}^{-1}) \right\rceil ,$$

so that $\tau_{\rm mix} \leqslant \left\lceil \frac{1}{\lambda_P} \ln(\pi_{\min}^{-2}/2) \right\rceil$.

Proof. We have:

$$\begin{aligned}
\mathrm{d}_{\rm TV}(P^t\pi_0, \pi) &= \frac{1}{2}\sum_{w\in\mathcal{V}} |(P^t\pi_0)_w - \pi_w| \\
&\leqslant \frac{1}{2\pi_{\min}} \sum_{w\in\mathcal{V}} \pi_w |(P^t\pi_0)_w - \pi_w| \\
&\leqslant \frac{1}{2\pi_{\min}} \sqrt{\|P^t\pi_0 - \pi\|_\pi^2} \\
&\leqslant \frac{(1-\lambda_P)^t}{2\pi_{\min}} \|\pi_0 - \pi\|_\pi \\
&\leqslant \frac{(1-\lambda_P)^t}{\pi_{\min}} ,
\end{aligned}$$

so that $|(P^t)_{v,w} - \pi_w| \leqslant \varepsilon$ for $t \geqslant \lambda_P^{-1}\ln(\pi_{\min}^{-1}\varepsilon^{-1}/2)$. □

Lemma A.2 (Mixing times and hitting times)**.**

$$\tau_{\rm hit} \leqslant 2\pi_{\min}^{-1}\tau_{\rm mix} ,$$

so that if π is the uniform distribution over $\mathcal{V}$, $\tau_{\rm hit} \leqslant 2n\tau_{\rm mix}$.

Proof. for any $v, w \in \mathcal{V}$,

$$\begin{aligned}
\mathbb{E}\left[\tau_w | v_0 = v\right] &= \sum_{k\geqslant 1} \mathbb{P}\left(\tau_w \geqslant k | v_0 = v\right) \\
&\leqslant \sum_{\ell\geqslant 0} \mathbb{P}\left(\tau_w > \ell\tau_{\rm mix} | v_0 = v\right) .
\end{aligned}$$

Then, for $\ell \geqslant 0$,

$$\mathbb{P}\left(\tau_w > (\ell+1)\tau_{\rm mix} | v_0 = v\right) = \mathbb{P}\left(\tau_w > (\ell+1)\tau_{\rm mix} | \tau_w > \ell\tau_{\rm mix}, v_0 = v\right)\mathbb{P}\left(\tau_w > \ell\tau_{\rm mix} | v_0 = v\right) ,$$

and, conditioning on $v_{\ell\tau_{\rm mix}}$, $\mathbb{P}\left(\tau_w > (\ell+1)\tau_{\rm mix} | \tau_w > \ell\tau_{\rm mix}, v_{\ell\tau_{\rm mix}}\right) \leqslant \mathbb{P}\left(v_{(\ell+1)\tau{\rm mix}} \neq w | v_{\ell\tau_{\rm mix}}\right)$. By definition of $\tau_{\rm mix}$, we have that $\mathbb{P}\left(v_{(\ell+1)\tau{\rm mix}} \neq w | v_{\ell\tau_{\rm mix}}\right) \leqslant (1-\pi_w/2)$, so that:

$$\mathbb{P}\left(\tau_w > (\ell+1)\tau_{\rm mix} | v_0 = v\right) \leqslant (1-\pi_w/2)\mathbb{P}\left(\tau_w > \ell\tau_{\rm mix} | v_0 = v\right) ,$$

and $\mathbb{P}\left(\tau_w > \ell\tau_{\rm mix} | v_0 = v\right) \leqslant (1-\pi_w/2)^\ell$ by recursion. Finally,

$$\begin{aligned}
\mathbb{E}\left[\tau_w | v_0 = v\right] &\leqslant \tau_{\rm mix} \sum_{\ell\geqslant 0} (1-\pi_w/2)^\ell \\
&\leqslant \frac{2\tau_{\rm mix}}{\pi_w} ,
\end{aligned}$$

concluding the proof by taking the maximum over w. □

A.2. Matthews' bound for cover times

The following result bounds the cover time of the Markov chain: it is in fact closely related to its hitting time, and the two differ with a most a factor $\ln(n)$. This surprising result is proved in a very elegant way in the survey Levin et al. (2006), using the famous Matthews' method (Matthews, 1988).

Theorem A.3 (Matthews' bound for cover times)**.** *The hitting and cover times of the Markov chain verify:*

$$\tau_{\text{cov}} \leqslant \left(\sum_{k=1}^{n-1} \frac{1}{k}\right) \tau_{\text{hit}} \,.$$

A.3. A bound on the hitting time of regular and symetric graphs

Using results from Rao (2012), we relate the hitting time of symmetric regular graphs (in a sense that we define below) to well-known graph-related quantities: number of edges $|\mathcal{E}|$, diameter δ and degree d.

Lemma A.4 (Bounding hitting times of regular graphs)**.** *Let (v_t) be the simple random walk on a d-regular graph G of diameter δ, that satisfies the following symetry property: for any $\{u, v\}, \{v, w\} \in \mathcal{E}$, there exists a graph automorphism that maps v to w. Then, we have:*

$$\tau_{\text{hit}} \leqslant \frac{2|\mathcal{E}|\delta}{d}$$

Proof. Using Theorem 2.1 of Rao (2012), for $\{v, w\} \in \mathcal{E}$, we have

$$\mathbb{E}\left[\tau_w | v_0 = v\right] = \frac{2|\mathcal{E}|}{d} \,,$$

where $|\mathcal{E}|$ is the number of edges in the graph. Let v and w in $\mathcal{V}$, at distance $\delta' \leqslant \delta$. There exists nodes $v = v(0), v(1) \ldots, v(\delta' - 1), v(\delta') = w$ such that for all $0 \leqslant s < \delta'$, $\{v(s), v(s+1)\} \in \mathcal{E}$, and by using the Markov property:

$$\mathbb{E}\left[\tau_w | v_0 = v\right] \leqslant \sum_{s<\delta'} \mathbb{E}\left[\tau_{v(s+1)} v_0 = v(s)\right] \leqslant \delta \frac{2|\mathcal{E}|}{d} \,.$$

□

A.4. Two miscellaneous lemmas

We finally end this "preliminary results" section with the two following lemmas, that we help us conclude the proof of Theorem 6.1. The first lemma will lead to a bound on $\left(\sum_{t<T} \gamma_t\right)^{-1}$ where γ_t is the adaptive stepsize policy defined in Equation (7), while the second one is used to conclude the proof of Theorem 6.1 to show that a remaining term is non-positive.

Lemma A.5. *For $t \geqslant 0$ and $v \in \mathcal{V}$, let $p_v(t) = \inf\{s > t | v_s = v\}$ and $d_v(t) = \sup\{s < t | v_s = v\}$ be the next and the last previous iterates for which $v_t = v$ ($d_v(t) = 0$ by convention, if v has not yet been visited). Assume that (v_t) has stationary distribution π. For $t \geqslant 0$, let $A_t = \sup_{v \in \mathcal{V}} \big(t - d_v(t)\big)$ and $B_t = \sup_{v \in \mathcal{V}} \big(p_v(t) - t\big)$. We have:*

$$\mathbb{E}\left[B_t | v_t = v\right] \leqslant \tau_{\text{cov}} \,, \quad \forall v \in \mathcal{V} \,,$$

and for $T \geqslant 1$:

$$\sum_{t<T} \mathbb{E}\left[A_t\right] \leqslant T \tau_{\text{cov}} \,.$$

Proof. The first bound on B_t is obtained using the Markov property of the chain, and by definition of τ_{cov}. We have:

$$\mathbb{E}\left[A_t\right] = \sum_{k=0}^{t-1} \mathbb{P}\left(A_t \geqslant k\right) \,.$$

For $t \geqslant 0$ fixed, we denote $d = \inf_v d_v(t)$, so that $A_t = t - d$. We then have the equality between the following events:

$$\{A_t \geqslant k\} = \{t - d \geqslant k\} = \{d \leqslant t - k\} = \{B_{t-k} \geqslant k\},$$

that all coincide with the event "there exists some $v \in \mathcal{V}$ such that for all $t - k \leqslant s \leqslant t$, $v_s \neq v$". Summing over $t < T$:

$$\begin{aligned}
\sum_{t<T} \mathbb{E}[A_t] &= \sum_{t<T} \sum_{k<t} \mathbb{P}(B_{t-k} \geqslant k) \\
&= \sum_{\ell<T} \sum_{s<T-\ell} \mathbb{P}(B_\ell \geqslant s) \\
&\leqslant \sum_{\ell<T} \sum_{s\geqslant 0} \mathbb{P}(B_\ell \geqslant s) \\
&\leqslant \sum_{\ell<T} \mathbb{E}[B_\ell] \\
&\leqslant T\tau_{\mathrm{cov}}.
\end{aligned}$$

□

Lemma A.6. *Let $(a_t)_{t\geqslant 0}, (b_t)_{t\geqslant 0}$ be two sequences of real-valued random variables. Let $(\mathcal{F}_t)_{t\geqslant 0}$ be a filtration. Assume that b_t is positive and $\mathcal{F}_t$-measurable for all t, and that $\mathbb{E}[a_t|\mathcal{F}_t] \leqslant 0$. Then, denoting $H_T = \frac{\sum_{t=0}^T a_t}{\sum_{t=0}^T b_t}$, the sequence $(\mathbb{E}[H_T])_{T\geqslant 0}$ is non-increasing, so that $\mathbb{E}[H_T] \leqslant 0$ for all T.*

Proof. For fixed $T \geqslant 1$, we have, using the fact tha b_t is $\mathcal{F}_T$ measurable for $t \leqslant T$

$$\begin{aligned}
\mathbb{E}[H_T|\mathcal{F}_T] &= \frac{\mathbb{E}[a_T|\mathcal{F}_T] + \sum_{t<T} \mathbb{E}[a_t|\mathcal{F}_T]}{\sum_{t\leqslant T} b_t} \\
&\leqslant \frac{\sum_{t<T} \mathbb{E}[a_t|\mathcal{F}_T]}{\sum_{t\leqslant T} b_t} \\
&\leqslant \frac{\sum_{t<T} \mathbb{E}[a_t|\mathcal{F}_T]}{\sum_{t<T} b_t},
\end{aligned}$$

using $\mathbb{E}[a_T|\mathcal{F}_T] \leqslant 0$ and $b_T > 0$. Consequently, taking the mean, we obtain $\mathbb{E}[H_T] \leqslant \mathbb{E}[H_{T-1}]$. □

B. Lower bound

We prove the smooth non-convex version of Theorem 4.1; the convex cases are proved in a similar way using exactly the same arguments, and the "most difficult function in the world", as defined by Nesterov (2014), rather than the one used by Carmon et al. (2021), albeit the two are closely related.

Proof of Theorem 4.1. For $\mathbf{x} \in \ell_2$ and $k \in \mathbb{N}$, denote by $\mathbf{x}(k)$ its kth coordinate. We split the function defined in Section 3.2 of Carmon et al. (2021) (inspired by the "most difficult function in the world" of Nesterov (2014)) between two nodes $v, w \in \mathcal{V}$ maximizing $\mathbb{E}[\tau_w|v_0 = v]$, by setting $\pi_v f_v(\mathbf{x}) = \frac{1}{2}\sum_{k\geqslant 1} 2\mathbf{x}(2k)^2 - 2\mathbf{x}(2k-1)\mathbf{x}(2k) + \frac{1}{2}\alpha\mathbf{x}(0)^2 - b\mathbf{x}(0) + \frac{\alpha}{2}$ and $\pi_w f_w(\mathbf{x}) = \frac{1}{2}\sum_{k\geqslant 0} 2\mathbf{x}(2k+1)^2 - 2\mathbf{x}(2k+1)\mathbf{x}(2k)$ for some $b, \alpha > 0$. Then, we define $T_0 = \tau_v$ and for $t \geqslant 0$,

$$\begin{aligned}
T_{2k+1} &= \inf\{t \geqslant T_{2k} \,|\, v_t = w\} \quad \text{and} \\
T_{2k+2} &= \inf\{t \geqslant T_{2k+1} \,|\, v_t = w\}.
\end{aligned}$$

The second step of the proof is somewhat classical, and consists in observing that the black-box constraints of the algorithm

together with the construction of the functions f_v and f_w defined in the proof sketch of Section 4 imply that:

$$\begin{cases} \text{if} \quad v_t = v \quad \text{and} \quad \begin{cases} \mathcal{M}_t \supset \mathrm{Span}(\mathbf{e}_i, i \leqslant 2k-1) \quad \text{then} \quad \mathcal{M}_{t+1} \supset \mathrm{Span}(\mathbf{e}_i, i \leqslant 2k)\,, \\ \mathcal{M}_t \subset \mathrm{Span}(\mathbf{e}_i, i \leqslant 2k) \quad \text{then} \quad \mathcal{M}_{t+1} \subset \mathrm{Span}(\mathbf{e}_i, i \leqslant 2k)\,, \end{cases} \\ \text{if} \quad v_t = w \quad \text{and} \quad \begin{cases} \mathcal{M}_t \supset \mathrm{Span}(\mathbf{e}_i, i \leqslant 2k) \quad \text{then} \quad \mathcal{M}_{t+1} \supset \mathrm{Span}(\mathbf{e}_i, i \leqslant 2k+1)\,, \\ \mathcal{M}_t \subset \mathrm{Span}(\mathbf{e}_i, i \leqslant 2k+1) \quad \text{then} \quad \mathcal{M}_{t+1} \subset \mathrm{Span}(\mathbf{e}_i, i \leqslant 2k+1)\,, \end{cases} \\ \text{if} \quad v_t \notin \{v, w\}, \quad \text{then} \quad \mathcal{M}_t = \mathcal{M}_{t+1}\,. \end{cases}$$

In other words, even dimensions are discovered by node v, while odd ones are discovered by node w. The dimension $\mathbb{R}\mathbf{e}_0$ is discovered by node v thanks to the term $-b\mathbf{e}_0$. Using Theorem 1 of Carmon et al. (2021), for a right choice of parameters $\alpha, b > 0$, f is L-smooth and satisfies $f(\mathbf{x}_0) - \inf_{\mathbf{x}} f(\mathbf{x}) \leqslant \Delta$, together with, any k and any $\mathbf{x} \in \mathcal{M}_t \subset \mathrm{Span}(\mathbf{e}_i, i \leqslant 2k)$,

$$\|\nabla f(\mathbf{x})\|^2 = \frac{L\Delta}{16k^2}\,.$$

This lower bound proof technique is explained in a detailed and enlightening fashion in Chapter 3.5 of Bubeck (2015).

Then, the final and more technical step of the proof consists in upper bounding $\mathbb{E}k(t)$. If $(T_{k+1} - T_k)_{k \geqslant 0}$ were independent from $k(t)$, using $\mathbb{E}\left[T_{k+1} - T_k\right] = \tau_{\mathrm{hit}}$ for k even, we would directly obtain $t \geqslant \mathbb{E}\left[T_{k(t)}\right] \geqslant \mathbb{E}\left[k(t)\right] - 1)\tau_{\mathrm{hit}}/2$. However, these random variables are not independent: since tail effects can happen, we need a finite second moment for hitting times, and the proof is a bit trickier. First, note that:

$$\mathbb{E}\left[k(t)\right] = \sum_{0 \leqslant k \leqslant t} \mathbb{P}\left(k(t) \geqslant k\right) = \sum_{0 \leqslant k \leqslant t} \mathbb{P}\left(T_k \leqslant t\right)\,.$$

Let $(X_\ell)_{\ell \geqslant 0}$ be *i.i.d.* random variables of same law as τ_w conditioned on $v_0 = v$. We have $\mathbb{E}\left[X_\ell\right] = \mathbb{E}\left[\tau_w | v_0 = v\right] = \tau_{\mathrm{hit}}$, and $\mathrm{var}\left(X_\ell\right) < \infty$ (by assumption). Let $S_k = \sum_{\ell=0}^{k-1} X_\ell$ (S_k has the same law as $\sum_{\ell=0}^{k-1} T_{2k+1} - T_{2k}$), so that, using the Markov property of (v_t), T_k stochastically dominates $S_{\lfloor k/2 \rfloor}$. Hence, $\mathbb{P}\left(T_k \leqslant t\right) \leqslant \mathbb{P}\left(S_{\lfloor k/2 \rfloor} \leqslant t\right)$. Then, using Chebychev inequality, for any $\ell \geqslant 0$ and for t such that $\ell\tau_{\mathrm{hit}} \geqslant t$, we have:

$$\begin{aligned} \mathbb{P}\left(S_\ell \leqslant t\right) &= \mathbb{P}\left(S_\ell - \ell\tau_{\mathrm{hit}} \leqslant t - \ell\tau_{\mathrm{hit}}\right) \\ &= \mathbb{P}\left((S_\ell - \ell\tau_{\mathrm{hit}})^2 \leqslant (t - \ell\tau_{\mathrm{hit}})^2\right) \\ &\leqslant \frac{\ell \mathrm{var}\left(X_0\right)}{(t - \ell\tau_{\mathrm{hit}})^2}\,. \end{aligned}$$

We then have:

$$\begin{aligned} \mathbb{E}\left[k(t)\right] &\leqslant 2 \sum_{0 \leqslant \ell \leqslant t/2} \mathbb{P}\left(S_\ell \leqslant t\right) \\ &= 2 \sum_{0 \leqslant \ell \leqslant 2t/\tau_{\mathrm{hit}}} \mathbb{P}\left(S_\ell \leqslant t\right) + 2 \sum_{2t/\tau_{\mathrm{hit}} \leqslant \ell \leqslant t/2} \mathbb{P}\left(S_\ell \leqslant t\right) \\ &\leqslant \frac{4t}{\tau_{\mathrm{hit}}} + 2 \sum_{2t/\tau_{\mathrm{hit}} \leqslant \ell \leqslant t/2} \frac{\ell \mathrm{var}\left(X_0\right)}{(t - \ell\tau_{\mathrm{hit}})^2}\,. \end{aligned}$$

We finally show that the second term stays bounded:

$$\begin{aligned} \sum_{2t/\tau_{\mathrm{hit}} \leqslant \ell \leqslant t/2} \frac{\ell}{(t - \ell\tau_{\mathrm{hit}})^2} &= \frac{1}{\tau_{\mathrm{hit}}^2} \sum_{0 \leqslant \ell \leqslant t/2 - 2t/\tau_{\mathrm{hit}}} \frac{\ell + 2t/\tau_{\mathrm{hit}}}{(\ell + t/\tau_{\mathrm{hit}})^2} \\ &= \frac{1}{\tau_{\mathrm{hit}}^2} \sum_{0 \leqslant \ell \leqslant t/2 - 2t/\tau_{\mathrm{hit}}} \frac{\ell}{(\ell + t/\tau_{\mathrm{hit}})^2} + \frac{2}{\tau_{\mathrm{hit}}^2} \sum_{0 \leqslant \ell \leqslant t/2 - 2t/\tau_{\mathrm{hit}}} \frac{t/\tau_{\mathrm{hit}}}{(\ell + t/\tau_{\mathrm{hit}})^2}\,. \end{aligned}$$

First, using a comparison with a continuous sum, we have:

$$\sum_{0 \leqslant \ell \leqslant t} \frac{\ell}{(\ell + t/\tau_{\mathrm{hit}})^2} \leqslant \sum_{0 \leqslant \ell \leqslant t} \frac{1}{(\ell + t/\tau_{\mathrm{hit}})} \leqslant \ln(\frac{t}{t/\tau_{\mathrm{hit}}}) = \ln(\tau_{\mathrm{hit}})\,,$$

since for $a, x > 0$, $\int_0^{ax} \frac{y\mathrm{d}y}{(y+a)^2} \leqslant \int_0^{ax} \frac{\mathrm{d}y}{(y+a)} = \ln(x)$. Finally, using $\sum_{\ell \geqslant 1} \frac{1}{(a+\ell)^2} \leqslant \int_0^\infty \frac{\mathrm{d}y}{(y+a)^2} = \frac{1}{a}$, we bound the second sum as:

$$\sum_{0 \leqslant \ell \leqslant t/2 - 2t/\tau_{\mathrm{hit}}} \frac{t/\tau_{\mathrm{hit}}}{(\ell + t/\tau_{\mathrm{hit}})^2} \leqslant \frac{\tau_{\mathrm{hit}}}{t} + 1 \,.$$

Wrapping our arguments together, we end up with:

$$\mathbb{E}\left[k(t)\right] \leqslant \frac{4t}{\tau_{\mathrm{hit}}} + \frac{2\mathrm{var}\left(\tau_v\right)}{\tau_{\mathrm{hit}}^2}(\ln(\tau_{\mathrm{hit}}) + 1 + \frac{\tau_{\mathrm{hit}}}{t}) \,.$$

For t big enough, we end up with $\mathbb{E}\left[k(t)\right] \leqslant 5t/\tau_{\mathrm{hit}}$, so that since $\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \geqslant L\Delta/(16\mathbb{E}\left[k(t)\right]^2)$ as explained in the main text, we have:

$$\|\nabla f(\mathbf{x}_t)\|^2 = \Omega\left(\frac{L\Delta\tau_{\mathrm{hit}}^2}{t^2}\right) \,.$$

□

C. Markov chain stochastic gradient descent: proof of Theorem 5.3

We start by proving the following bound on $\mathbb{E}\left[\|\nabla f_{v_t}(\mathbf{x}_t)\|^2\right]$. Note that this bound can be used for any $t \geqslant \tau_{\mathrm{mix}}$.

Lemma C.1. *For $t \geqslant 0$ and if $v_t \sim \pi_t$ for $\mathrm{d}_{\mathrm{TV}}(\pi_t, \pi) \leqslant \pi_{\min}/2$, we have:*

$$\mathbb{E}\left[\|\nabla f_{v_t}(\mathbf{x}_t)\|^2\right] \leqslant 3\bar{\sigma}^2 + 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \,.$$

Proof of the Lemma. We have for any $v \in \mathcal{V}$ that $\mathbb{P}\left(v_t = v\right) \leqslant \pi_v + \pi_v/2 = 3\pi_v/2$, so that

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f_{v_t}(\mathbf{x}_t)\|^2\right] &\leqslant 2\mathbb{E}\left[\|\nabla f_{v_t}(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2\right] + 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \\
&\leqslant 2\sum_{v \in \mathcal{V}} \mathbb{P}\left(v_t = v\right)\sigma_v^2 + 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \\
&= 3\bar{\sigma}^2 + 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \,.
\end{aligned}$$

□

The proof borrows ideas from both the analyses of delayed SGD (Mania et al., 2017) and SGD with biased gradients (Even et al., 2022), thus refining MC-SGD initial analysis (Johansson et al., 2010). While a biased gradient analysis would not yield convergence to an ε-stationary point for arbitrary ε (at every iterations, biases are non-negligible and can be arbitrary high), by enforcing a delay τ (of order τ_{mix}) in the analysis, we manage to take advantage of the ergodicity of the biases.

C.1. Smooth non-convex case of Theorem 5.3

Proof of Theorem 5.3.1. Denoting $F_t = \mathbb{E}f(\mathbf{x}_t) - f(\mathbf{x}^\star)$, we have using smoothness:

$$F_{t+1} - F_t \leqslant -\gamma\mathbb{E}\left[\langle \nabla f_{v_t}(\mathbf{x}_t), \nabla f(\mathbf{x}_t)\rangle\right] + \frac{\gamma^2 L}{2}\mathbb{E}\left[\|\nabla f_{v_t}(\mathbf{x}_t)\|^2\right] \,.$$

For the first term on the righthandside of the inequality, assuming that $t \geqslant \tau$ for some $\tau > 0$ we explicit later in the proof:

$$\begin{aligned}
\mathbb{E}\left[-\gamma\langle \nabla f_{v_t}(\mathbf{x}_t), \nabla f(\mathbf{x}_t)\rangle\right] &= \mathbb{E}\left[-\gamma\langle \nabla f_{v_t}(\mathbf{x}_{t-\tau}), \nabla f(\mathbf{x}_{t-\tau})\rangle\right] + \mathbb{E}\left[-\gamma\langle \nabla f_{v_t}(\mathbf{x}_t), \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-\tau})\rangle\right] \\
&\quad + \mathbb{E}\left[-\gamma\langle \nabla f_{v_t}(\mathbf{x}_t) - \nabla f_{v_t}(\mathbf{x}_{t-\tau}), \nabla f(\mathbf{x}_{t-\tau})\rangle\right] \,.
\end{aligned}$$

First, we condition the first term on the filtration up to time $t - \tau$:

$$\begin{aligned}
\mathbb{E}\left[-\gamma\langle \nabla f_{v_t}(\mathbf{x}_{t-\tau}), \nabla f(\mathbf{x}_{t-\tau})\rangle\right] &= \mathbb{E}\left[-\gamma\langle \mathbb{E}_{t-\tau}\nabla f_{v_t}(\mathbf{x}_{t-\tau}), \nabla f(\mathbf{x}_{t-\tau})\rangle\right] \\
&\leqslant -\frac{\gamma}{2}\mathbb{E}\left[\left\|\mathbb{E}_{t-\tau}\nabla f_{v_{t-\tau}}(\mathbf{x}_t)\right\|^2\right] + \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau}) - \mathbb{E}_{t-\tau}\nabla f_{v_t}(\mathbf{x}_{t-\tau})\|\right] \\
&\quad - \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] \,.
\end{aligned}$$

Then, for $\tau \geqslant \tau_{\text{mix}}(\pi_{\min}\varepsilon)$, using the following lemma, we have, for $\varepsilon < 1/2$:

$$\mathbb{E}\left[-\gamma\langle \mathbb{E}_{t-\tau}\nabla f_{v_t}(\mathbf{x}_{t-\tau}), \nabla f(\mathbf{x}_{t-\tau})\rangle\right] \leqslant -\frac{\gamma}{4}\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + \gamma\varepsilon^2\bar{\sigma}^2 \,.$$

Lemma C.2. *For $\tau \geqslant \tau_{\text{mix}}(\varepsilon\pi_{\min})$ and $t \geqslant \tau$,*

$$\mathbb{E}\left[\|\mathbb{E}_{t-\tau}\nabla f_{v_t}(\mathbf{x}_{t-\tau}) - \nabla f(\mathbf{x}_{t-\tau})\|^2\right] \leqslant 2\varepsilon^2\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + 2\varepsilon^2\bar{\sigma}^2 \,.$$

Proof of the Lemma. We have:

$$\begin{aligned}\mathbb{E}\left[\|\mathbb{E}_{t-\tau}\nabla f_{v_t}(\mathbf{x}_{t-\tau}) - \nabla f(\mathbf{x}_{t-\tau})\|^2\right] &= \mathbb{E}\left[\left\|\sum_{v\in\mathcal{V}}(\mathbb{P}\left(v_t = v|\mathbf{x}_{t-\tau}\right) - \pi_v)\nabla f_v(\mathbf{x}_{t-\tau})\right\|^2\right] \\ &\leqslant \varepsilon^2\sum_{v\in\mathcal{V}}\pi_v\mathbb{E}\left[\|\nabla f_v(\mathbf{x}_{t-\tau})\|^2\right] \,,\end{aligned}$$

where we used $|\mathbb{P}\left(v_t = v|\mathbf{x}_{t-\tau}\right) - \pi_v| \leqslant \varepsilon\pi_v$ and convexity of the squared Euclidean norm. For that last term,

$$\begin{aligned}\sum_{v\in\mathcal{V}}\pi_v\mathbb{E}\left[\|\nabla f_v(\mathbf{x}_{t-\tau})\|^2\right] &\leqslant \sum_{v\in\mathcal{V}}2\pi_v\left(\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + \sigma_v^2\right) \\ &= 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + 2\bar{\sigma}^2 \,,\end{aligned}$$

concluding the proof of the Lemma. □

Using gradient Lipschitzness and writing $\mathbf{x}_t - \mathbf{x}_{t-\tau} = -\gamma\sum_{s=\max(t-\tau,0)}^{t-1}\nabla f_{v_s}(\mathbf{x}_s)$, we have:

$$\begin{aligned}\mathbb{E}\left[-\gamma\langle\nabla f_{v_t}(\mathbf{x}_t), \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-\tau})\rangle\right] &\leqslant \gamma^2 L\mathbb{E}\left[\|\nabla f_{v_t}(\mathbf{x}_t)\|\left\|\sum_{s=\max(t-\tau,0)}^{t-1}\nabla f_{v_s}(\mathbf{x}_s)\right\|\right] \\ &\leqslant \frac{\gamma^2 L}{2}(\tau\mathbb{E}\left[\|\nabla f_{v_t}(\mathbf{x}_t)\|^2\right] + \sum_{s=\max(t-\tau,0)}^{t-1}\mathbb{E}\left[\|\nabla f_{v_s}(\mathbf{x}_s)\|^2\right]) \,.\end{aligned}$$

Similarly,

$$\mathbb{E}\left[-\gamma\langle\nabla f_{v_t}(\mathbf{x}_t) - \nabla f_{v_t}(\mathbf{x}_{t-\tau}), \nabla f(\mathbf{x}_{t-\tau})\rangle\right] \leqslant \frac{\gamma^2 L}{2}(\tau\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + \sum_{s=\max(t-\tau,0)}^{t-1}\mathbb{E}\left[\|\nabla f_{v_s}(\mathbf{x}_s)\|^2\right]) \,.$$

Wrapping things up, we obtain, for $t \geqslant \tau$ and $\tau \geqslant \tau_{\text{mix}}$:

$$\begin{aligned}F_{t+1} - F_t &\leqslant -\frac{\gamma}{4}\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + \gamma\varepsilon^2\bar{\sigma}^2 \\ &\quad + \frac{\gamma^2 L}{2}\big((\tau+1)\mathbb{E}\left[\|\nabla f_{v_t}(\mathbf{x}_t)\|^2\right] + \tau\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + 2\sum_{s=\max(t-\tau,0)}^{t-1}\mathbb{E}\left[\|\nabla f_{v_s}(\mathbf{x}_s)\|^2\right]\big) \\ &\leqslant -\frac{\gamma}{4}\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + \gamma\varepsilon^2\bar{\sigma}^2 + (3\tau+1)\frac{\gamma^2 L}{2} \\ &\quad + \frac{\gamma^2 L}{2}\big((\tau+1)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + \tau\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + 2\sum_{s=\max(t-\tau,0)}^{t-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_s)\|^2\right]\big) \,.\end{aligned}$$

Summing for $\tau \leqslant t < T$:

$$\frac{1}{T}\sum_{\tau\leqslant t<T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] \leqslant \frac{4F_\tau}{\gamma T} + \frac{1}{T}\sum_{\tau\leqslant t<T}6\gamma L\tau\big(\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + \mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right]\big) + 2\big(2\varepsilon^2 + \gamma(3\tau+1)\big)\bar{\sigma}^2 \,,$$

leading to, for $\gamma \leqslant \frac{1}{12L\tau}$:

$$\frac{1}{T}\sum_{t<T-\tau}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \leqslant \frac{4F_\tau}{\gamma T} + \frac{6\gamma L\tau}{T}\sum_{T-\tau\leqslant t<T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + 2\big(2\varepsilon^2+\gamma(3\tau+1)\big)\bar{\sigma}^2\,. \tag{8}$$

We now prove that for any $t \geqslant 0$, we have $\sup_{t\leqslant s\leqslant t+\tau}\mathbb{E}\left[\|\nabla f(\mathbf{x}_s)\|^2\right] \leqslant 4\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + 8\gamma^2L^2\tau^2\sigma^2$. Let $t \leqslant s < t+\tau$.

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(\mathbf{x}_s)\|^2\right] &\leqslant 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_s)-\nabla f(\mathbf{x}_t)\|^2\right]\\
&\leqslant 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + 2L^2\gamma^2\mathbb{E}\left[\sum_{r=t}^{s-1}\|\nabla f_{v_r}(\mathbf{x}_r)\|^2\right]\\
&\leqslant 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + 4L^2\gamma^2\tau\sum_{r=t}^{s-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_r)\|^2\right] + \bar{\sigma}^2\\
&\leqslant 2\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + 4L^2\gamma^2\tau^2(\sup_{t\leqslant s\leqslant t+\tau}\mathbb{E}\left[\|\nabla f(\mathbf{x}_s)\|^2\right] + \bar{\sigma}^2)\,,
\end{aligned}$$

leading to the desired result for $\gamma \leqslant 1/(8L\tau)$. Plugging this in (8):

$$\begin{aligned}
\frac{1}{T}\sum_{t<T-\tau}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \leqslant \frac{4F_\tau}{\gamma T} &+ \frac{24\gamma L\tau}{T}\sum_{T-\tau\leqslant t<T}\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right]\\
&+ \frac{\tau}{T}4L^2\gamma^2\tau^2\bar{\sigma}^2 + 2\big(2\varepsilon^2+\gamma(3\tau+1)\big)L\bar{\sigma}^2\,.
\end{aligned}$$

Now, for $\gamma = \min(1/(48L\tau), \sqrt{F_0/(TL\tau\bar{\sigma}^2)})$, $\varepsilon = 1/\sqrt{T}$, and so $\tau = \tau_{\text{mix}}\ln(T)$, we have:

$$\frac{1}{T}\sum_{t<T-\tau}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] \leqslant \frac{196\tau LF_\tau}{T} + 7\sqrt{\frac{LF_0\bar{\sigma}^2}{T}}\,. \tag{9}$$

We now upper bound F_τ. For any t and $\gamma < 1/(2L)$,

$$F_{t+1} - F_t \leqslant \frac{\gamma}{2}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t) - \nabla f_{v_t}(\mathbf{x}_t)\|^2\right] \leqslant \gamma\sigma_{\max}^2/2\,,$$

where the first inequality is a simplified version of the descent lemma with biased gradient at the beggining of this proof, and the second inequality uses the initialization properties of v_0. Thus, we obtain $F_\tau \leqslant F_0 + \gamma\tau\sigma_{\max}^2/2 \leqslant F_0 + \sigma_{\max}^2/L$ for our choice of γ. We thus conclude by plugging this in (9) applied for $T+\tau$ instead of T, yielding the desired result.

The condition for the upper-bound we proved above to be true, namely $T \geqslant \tau = \tau_{\text{mix}}\ln(T)$, is always satisfied for $T \geqslant 2\tau_{\text{mix}}\ln(\tau_{\text{mix}})$. Indeed, if $T \leqslant \tau_{\text{mix}}^2$, then $\tau_{\text{mix}}\ln(T) \leqslant 2\tau_{\text{mix}}\ln(\tau_{\text{mix}}) \leqslant T$, and otherwise we have $\tau_{\text{mix}}\ln(T) \leqslant \sqrt{T}\ln(T) \leqslant T$. This concludes the proof, and $\tilde{\mathbf{x}}_0$ in the Theorem corresponds to $\mathbf{x}_\tau$. □

C.2. Under a μ-PL inequality

Proof of Theorem 2.2. We start from:

$$\begin{aligned}
F_{t+1} - F_t \leqslant &-\frac{\gamma}{4}\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + \gamma\varepsilon^2\bar{\sigma}^2 + (3\tau+1)\frac{\gamma^2L}{2}\\
&+ \frac{\gamma^2L}{2}\big((\tau+1)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + \tau\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + 2\sum_{s=\max(t-\tau,0)}^{t-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_s)\|^2\right]\big)\,.
\end{aligned}$$

If f satisfies a μ-PL inequality, then $-\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] \leqslant -2\mu F_{t-\tau}$, so that, for some $\alpha \in (0,1)$:

$$\begin{aligned}
F_{t+1} - F_t \leqslant &-\frac{\alpha\gamma\mu}{4}F_{t-\tau} - \frac{(1-\alpha)\gamma}{8}\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + \gamma\varepsilon^2\bar{\sigma}^2 + (3\tau+1)\frac{\gamma^2L}{2}\bar{\sigma}^2\\
&+ \frac{\gamma^2L}{2}\big((\tau+1)\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] + \tau\mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-\tau})\|^2\right] + 2\sum_{s=\max(t-\tau,0)}^{t-1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_s)\|^2\right]\big)\,.
\end{aligned}$$

For $t \geqslant 0$, let $P_t = (1-\alpha\gamma\mu/4)^{-t}$. We multiply the above expression by P_{t+1} and sum for $t < T$, hoping for cancellations. For $T \geqslant \tau$:

$$\begin{aligned}
\sum_{\tau \leqslant t < T} P_{t+1}\big(F_t - F_{t+1} - \frac{\alpha\gamma\mu}{4}F_{t-\tau}\big) &= \sum_{\tau \leqslant t < T} P_{t+1}\big((1-\frac{\alpha\gamma\mu}{4})F_t - F_{t+1} + \frac{\alpha\gamma\mu}{4}(F_t - F_{t-\tau})\big) \\
&= \sum_{\tau \leqslant t < T} P_t F_t - \sum_{\tau+1 \leqslant t \leqslant T} P_t F_t \\
&\quad + \frac{\alpha\gamma\mu}{4}\sum_{\tau \leqslant t < T} P_{t+1}F_t - \frac{\alpha\gamma\mu}{4}\sum_{\tau \leqslant t < T} P_{t+1}F_{t-\tau} \\
&\leqslant P_\tau F_\tau - P_T F_T + \frac{\alpha\gamma\mu}{4}\sum_{\tau \leqslant t < T} P_{t+1}F_t - \frac{P_\tau \alpha\gamma\mu}{4}\sum_{0 \leqslant t < T-\tau} P_{t+1}F_t \\
&\leqslant P_\tau F_\tau - P_T F_T + \frac{\alpha\gamma\mu}{4}\sum_{T-\tau \leqslant t < T} P_{t+1}F_t \\
&\leqslant P_\tau F_\tau - P_T F_T + \frac{\alpha\gamma}{8}\sum_{T-\tau \leqslant t < T} P_{t+1}\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right] ,
\end{aligned}$$

using the μ-PL inequality. For $t \geqslant 0$, we denote $R_t = \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2\right]$. We now handle the "$R_t$" terms.

$$\begin{aligned}
-\sum_{\tau \leqslant t < T} \frac{(1-\alpha)\gamma}{8} P_{t+1}R_{t-\tau} &+ \sum_{\tau \leqslant t < T} \frac{\gamma^2 L}{2}\big((\tau+1)P_{t+1}R_t + \tau P_{t+1}R_{t-\tau} + 2\sum_{s=t-\tau}^{t-1} P_{t+1}R_s\big) \\
&\leqslant -\sum_{0 \leqslant t < T-\tau} \frac{(1-\alpha)\gamma}{8} P_{t+\tau+1}R_t \\
&\quad + \frac{\gamma^2 L}{2}\left((\tau+1)\sum_{\tau \leqslant t < T} P_{t+1}R_t + \tau\sum_{0 \leqslant t < T-\tau} P_{t+1}R_t + 2\tau\sum_{t<T} R_t P_{t+\tau}\right) \\
&= -\sum_{0 \leqslant t < T-\tau} P_{t+1}R_t\gamma\Big(\frac{(1-\alpha)}{8}P_\tau - \frac{\gamma L}{2}(2\tau+1+2\tau P_{\tau-1})\Big) \\
&\quad + \frac{\gamma^2 L}{2}\sum_{T-\tau \leqslant t < T}\big((\tau+1+2\tau P_{\tau-1})\big)P_{t+1}R_t \\
&\leqslant -\sum_{0 \leqslant t < T-\tau} \frac{(1-\alpha)\gamma}{16} P_{t+\tau+1}R_t \\
&\quad + \frac{(1-\alpha)\gamma}{16\beta}\sum_{T-\tau \leqslant t < T} P_{t+1}R_t ,
\end{aligned}$$

if γ satisfies $\gamma \leqslant \frac{1-\alpha}{8\beta L(5\tau+1)}$ and $P_\tau \leqslant 2$, for some $\beta \geqslant 1$. Since for $\gamma\mu \leqslant 1$, $P_\tau \leqslant e^{\tau\mu\gamma}$, $P_\tau \leqslant 2$ can be ensured with $\gamma \leqslant \frac{1}{2\tau L}$. All in one, we have:

$$\begin{aligned}
0 \leqslant P_\tau F_\tau - P_T F_T &+ \gamma\Big(\frac{\alpha}{8} + \frac{1-\alpha}{16\beta}\Big)\sum_{T-\tau \leqslant t < T} P_{t+1}R_t \\
&- \sum_{0 \leqslant t < T-\tau} \frac{(1-\alpha)\gamma}{16} P_{t+\tau+1}R_t \\
&+ \Big(\gamma\varepsilon^2\bar{\sigma}^2 + (3\tau+1)\frac{\gamma^2 L}{2}\bar{\sigma}^2\Big)\sum_{\tau \leqslant t < T} P_{t+1} .
\end{aligned}$$

Using what we proved in the previous proof, we have $R_t \leqslant 4R_{t-\tau} + 8\gamma^2 L^2 \tau^2 \sigma^2$ for $T - \tau \leqslant t < T$, so that:

$$\gamma\Big(\frac{\alpha}{8} + \frac{1-\alpha}{16\beta}\Big) \sum_{T-\tau \leqslant t < T} P_{t+1} R_t \leqslant 4\gamma\Big(\frac{\alpha}{8} + \frac{1-\alpha}{16\beta}\Big) \sum_{T-2\tau \leqslant t < T-\tau} P_{t+\tau+1} R_t + 8\gamma^2 L^2 \tau^2 \sigma^2 \gamma \Big(\frac{\alpha}{8} + \frac{1-\alpha}{16\beta}\Big) \sum_{T-2\tau \leqslant t < T-\tau} P_{t+\tau+1} \,.$$

Consequently, for $4\gamma\Big(\frac{\alpha}{8} + \frac{1-\alpha}{16\beta}\Big) \leqslant \frac{1-\alpha}{16}\gamma$, which can be ensured with $\alpha = 1/16$ and $\beta = 8$, we have:

$$0 \leqslant P_\tau F_\tau - P_T F_T + \frac{1}{8}\gamma^3 L^2 \tau^2 \sigma^2 \sum_{T-2\tau \leqslant t < T-\tau} P_{t+\tau+1} + \Big(\gamma\varepsilon^2 + (3\tau+1)\frac{\gamma^2 L}{2}\Big)\bar{\sigma}^2 \sum_{\tau \leqslant t < T} P_{t+1} \,,$$

so that:

$$\begin{aligned} F_T &\leqslant F_\tau / P_{T-\tau} + \gamma^2 \bar{\sigma}^2 L \Big(\frac{\varepsilon^2}{L\gamma} + \frac{3\tau+1}{2} + \frac{\gamma L \tau^2}{8}\Big) \frac{\sum_{t \leqslant T} P_t}{P_T} \\ &\leqslant 2F_\tau / P_T + \frac{2\gamma\bar{\sigma}^2}{\mu} L \Big(\frac{\varepsilon^2}{L\gamma} + \frac{3\tau+1}{2} + \frac{\gamma L \tau^2}{8}\Big) \\ &\leqslant 2F_\tau / P_T + \frac{2\gamma\bar{\sigma}^2}{\mu} L (\frac{\varepsilon^2}{L\gamma} + \frac{1}{2} + 2\tau) \,. \end{aligned}$$

Finally, using $F_\tau \leqslant F_0 + \sigma_{\max}^2 / L$, if $\gamma \leqslant \frac{1}{64(5\tau+1)}$ where $\tau = \tau_{\text{mix}}(\varepsilon)$:

$$F_T \leqslant 2(F_0 + \sigma_{\max}^2 / L)(1 - \frac{\gamma\mu}{8})^T + \frac{2\gamma\bar{\sigma}^2}{\mu}(\frac{\varepsilon^2}{L\gamma} + \frac{1}{2} + 2\tau)$$

We thus choose $\varepsilon = \sqrt{1/T}$ so that $\tau \leqslant \tau_{\text{mix}} \ln(T)$, and stepsize $\gamma = \min(\frac{8\ln(T(F_0+\bar{\sigma}^2)/\bar{\sigma}^2)}{\mu T}, \frac{1}{64(5\tau+1)})$, leading to the desired result for $c = 64 \times 6 = 384$.

The same discussion than in the smooth non-convex proof regarding the condition $T \geqslant \tau$ applies here. □

D. Markov chain SGD: results in the interpolation regime

D.1. Proof of Theorem 5.6

We begin by proving the following lemma.

Lemma D.1. *For any $T \geqslant 1$, we have:*

$$\mathbb{E}\left[\left\|\sum_{t<T} \nabla f_{v_t}(\mathbf{x}^\star)\right\|^2\right] \leqslant T\sigma_\star^2 + \sigma_\star^2 \sum_{t<T} \mathrm{d}_{\mathrm{TV}}(P^t_{v_0,\cdot}, \pi^\star) + 2\sigma_\star^2 \sum_{s<t<T} \mathrm{d}_{\mathrm{TV}}(t-s) \,,$$

where $\mathrm{d}_{\mathrm{TV}}(r) = \sup\{\mathrm{d}_{\mathrm{TV}}((P^r)_{v,\cdot}, \pi^\star), v \in \mathcal{V}\}$ for $r \in \mathbb{N}$, so that:

$$\mathbb{E}\left[\left\|\sum_{t<T} \nabla f_{v_t}(\mathbf{x}^\star)\right\|^2\right] \leqslant \sigma_\star^2\big(4\tau_{\text{mix}}(1/4) + T(1 + 8\tau_{\text{mix}}(1/4))\big) \,.$$

Proof. We have:

$$\begin{aligned} \mathbb{E}\left[\left\|\sum_{t<T} \nabla f_{v_t}(\mathbf{x}^\star)\right\|^2\right] &= \mathbb{E}\left[\left(\sum_{t<T} \nabla f_{v_t}(\mathbf{x}^\star)\right)^\top \left(\sum_{t<T} \nabla f_{v_t}(\mathbf{x}^\star)\right)\right] \\ &= \sum_{t<T} \mathbb{E}\left[\|\nabla f_{v_t}(\mathbf{x}^\star)\|^2\right] + 2\sum_{s<t<T} \mathbb{E}\left[\langle \nabla f_{v_s}(\mathbf{x}^\star), \nabla f_{v_t}(\mathbf{x}^\star)\rangle\right] \,. \end{aligned}$$

Denote $\mathbf{G}_\star = (\nabla f_v(\mathbf{x}^\star))_{v\in\mathcal{V}} \in \mathbb{R}^{\mathcal{V}\times d}$. For the first term above,

$$\begin{aligned}\sum_{t<T}\mathbb{E}\left[\|\nabla f_{v_t}(\mathbf{x}^\star)\|^2\right] &= \sum_{t<T}\mathbb{E}\left[\|\mathbf{G}_{\star,v_s}\|^2\right]\\ &= \sum_{t<T}\left(\sigma_\star^2 + \sum_{v\in\mathcal{V}}\left(\mathbb{P}\left(v_t=v\right)-\pi_v\right)\|\mathbf{G}_{\star,v}\|^2\right)\\ &\leqslant T\sigma_\star^2 + \sigma_\star^2\sum_{t<T}\mathrm{d}_{\mathrm{TV}}(P^t_{v_0,\cdot},\pi^\star)\,.\end{aligned}$$

Finally,

$$\begin{aligned}\sum_{s<t<T}\mathbb{E}\left[\langle\nabla f_{v_s}(\mathbf{x}^\star),\nabla f_{v_t}(\mathbf{x}^\star)\rangle\right] &= \sum_{s<t<T}\mathbb{E}\left[\langle\mathbf{G}_{\star,v_s},\mathbf{G}_{\star,v_t}\rangle\right]\\ &= \sum_{s<t<T}\sum_{v,w\in\mathcal{V}}(P^s)_{v_0,v}(P^{t-s})_{v,w}\mathbf{G}_{\star,v}^\top\mathbf{G}_{\star,w}\\ &= \sum_{s<t<T}\sum_{v,w\in\mathcal{V}}(P^s)_{v_0,v}\big((P^{t-s})_{v,w}-\frac{1}{n}\big)\mathbf{G}_{\star,v}^\top\mathbf{G}_{\star,w}\\ &\leqslant \sum_{s<t<T}\sum_{v,w\in\mathcal{V}}(P^s)_{v_0,v}\big|(P^{t-s})_{v,w}-\frac{1}{n}\big|\sigma_\star^2\\ &= \sigma_\star^2\sum_{s<t<T}\sum_{v\in\mathcal{V}}(P^s)_{v_0,v}\sum_{w\in\mathcal{V}}\big|(P^{t-s})_{v,w}-\frac{1}{n}\big|\\ &\leqslant \sigma_\star^2\sum_{s<t<T}\mathrm{d}_{\mathrm{TV}}(t-s)\,,\end{aligned}$$

where $\mathrm{d}_{\mathrm{TV}}(r) = \sup\left\{\mathrm{d}_{\mathrm{TV}}((P^r)_{v,\cdot},\pi^\star)\,, v\in\mathcal{V}\right\}$ for $r\in\mathbb{N}$.

We finally bound $\sum_{t<T}\mathrm{d}_{\mathrm{TV}}(t)$. Using Chapter 4.5 of Levin et al. (2006), we have $\tau_{\mathrm{mix}}(\varepsilon)\leqslant(\log_2(\varepsilon^{-1})+1)\tau_{\mathrm{mix}}(1/4)$, so that for any $t\geqslant 0$, $d_{\mathrm{TV}}(t)\leqslant 2^{-t/\tau_{\mathrm{mix}}(1/4)+1}$. Hence,

$$\sum_{t<T}\mathrm{d}_{\mathrm{TV}}(t)\leqslant\frac{2}{1-2^{-1/\tau_{\mathrm{mix}}(1/4)}}\leqslant 4\tau_{\mathrm{mix}}(1/4)\,,$$

and

$$\sum_{s<t<T}\mathrm{d}_{\mathrm{TV}}(t-s)\leqslant 4T\tau_{\mathrm{mix}}(1/4)\,,$$

concluding the proof. □

Lemma D.2. *For any $\mathbf{y}_t\in\mathbb{R}^d$ and $t\geqslant 0$, denoting $\mathbf{y}_{t+1}=\mathbf{y}_t-\gamma\nabla f_{v_t}(\mathbf{x}^\star)$, we have:*

$$\|\mathbf{x}_{t+1}-\mathbf{y}_{t+1}\|^2\leqslant(1-\gamma\mu)\|\mathbf{x}_t-\mathbf{y}_t\|^2+\gamma L\|\mathbf{y}_t-\mathbf{x}^\star\|^2\,.$$

Proof. Denote $f_t = f_{v_t}(\cdot)$. We expand:

$$\|\mathbf{x}_{t+1}-\mathbf{y}_{t+1}\|^2 = \|\mathbf{x}_t-\mathbf{y}_t\|^2 - 2\gamma\langle\nabla f_t(\mathbf{x}_t)-\nabla f_t(\mathbf{x}^\star),\mathbf{x}_t-\mathbf{y}_t\rangle + \gamma^2\|\nabla f_t(\mathbf{x}_t)-\nabla f_t(\mathbf{x}^\star)\|^2\,.$$

Using the *three points equality* as Mishchenko et al. (2020), we have:

$$-2\gamma\langle\nabla f_t(\mathbf{x}_t)-\nabla f_t(\mathbf{x}^\star),\mathbf{x}_t-\mathbf{y}_t\rangle = -2\gamma D_{f_t}(\mathbf{y}_t,\mathbf{x}_t) - 2\gamma D_{f_t}(\mathbf{x}_t,\mathbf{x}^\star) + 2\gamma D_{f_t}(\mathbf{y}_t,\mathbf{x}^\star)\,.$$

First, $-2\gamma D_{f_t}(\mathbf{y}_t,\mathbf{x}_t)\leqslant-\gamma\mu\|\mathbf{x}_t-\mathbf{y}_t\|^2$ using strong convexity. Then, $-2\gamma D_{f_t}(\mathbf{x}_t,\mathbf{x}^\star)$ cancels the term $\gamma^2\|\nabla f_t(\mathbf{x}_t)-\nabla f_t(\mathbf{x}^\star)\|^2\leqslant 2\gamma^2 L D_{f_t}(\mathbf{x}_t,\mathbf{x}^\star)$ for $\gamma\leqslant 1/L$, using smoothness of f_t. And finally, using smoothness again, $2\gamma D_{f_t}(\mathbf{y}_t,\mathbf{x}^\star)\leqslant\gamma L\|\mathbf{y}_t-\mathbf{x}^\star\|^2$, concluding the proof. □

Proof of Theorem 5.6. Fix some $y_0 \in \mathbb{R}^d$ and let $(\mathbf{y}_t)$ be defined with the recursion

$$\mathbf{y}_{t+1} = \mathbf{y}_t - \gamma \nabla f_{v_t}(\mathbf{x}^\star, \xi_t)\,.$$

Unrolling the previous Lemma, we have, for a fixed time horizon T:

$$\|\mathbf{x}_T - \mathbf{y}_T\|^2 \leqslant (1-\gamma\mu)^T \|\mathbf{x}_0 - \mathbf{y}_0\|^2 + \gamma L \sum_{t<T} (1-\gamma\mu)^{T-t} \|\mathbf{y}_t - \mathbf{x}^\star\|^2\,.$$

This is possible, since the descent lemma is deterministic, in the sense that no expectations are taken so far. Since we want control over the distance to the optimum, we wish to have $\mathbf{y}_T = \mathbf{x}^\star$, leading to:

$$\mathbf{y}_0 = \mathbf{x}^\star + \gamma \sum_{t<T} \nabla f_{v_t}(\mathbf{x}^\star, \xi_t)\,, \quad \mathbf{y}_s = \mathbf{x}^\star + \gamma \sum_{s \leqslant t < T} \nabla f_{v_t}(\mathbf{x}^\star, \xi_t)\,, \quad s < T\,.$$

We thus have:

$$\begin{aligned}\mathbb{E}\|\mathbf{x}_T - \mathbf{x}^\star\|^2 \leqslant 2(1-\gamma\mu)^T &\left(\mathbb{E}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \gamma^2 \mathbb{E}\left[\left\| \sum_{s<T} \nabla f_{v_s}(\mathbf{x}^\star) \right\|^2 \right] \right) \\ &+ \gamma^3 L \sum_{t<T} (1-\gamma\mu)^{T-t} \mathbb{E}\left[\left\| \sum_{t \leqslant s < T} \nabla f_{v_s}(\mathbf{x}^\star) \right\|^2 \right]\,.\end{aligned}$$

Using Lemma D.1, we have for any $t \leqslant T$:

$$\mathbb{E}\left[\left\| \sum_{t \leqslant s < T} \nabla f_{v_s}(\mathbf{x}^\star) \right\|^2 \right] \leqslant C T \tau_{\text{mix}} \sigma_\star^2\,,$$

for $C = 13$. Hence,

$$\begin{aligned}\mathbb{E}\|\mathbf{x}_T - \mathbf{x}^\star\|^2 &\leqslant 2(1-\gamma\mu)^T \left(\mathbb{E}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \gamma^2 T C \tau_{\text{mix}} \sigma_\star^2 \right) \\ &\quad + \gamma^3 L \sum_{t<T} (1-\gamma\mu)^{T-t} (T-t) C \tau_{\text{mix}} \sigma_\star^2 \\ &\leqslant 2(1-\gamma\mu)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \frac{3\gamma L}{\mu^2} C \tau_{\text{mix}} \sigma_\star^2\,,\end{aligned}$$

using $\sum_{t<T} (1-x)^t t \leqslant 1/x^2$ and $(1-z)^x x \leqslant \frac{1}{eu} \leqslant 1/(2u)$. Finally, for a stepsize choice of

$$\gamma = \min\left(\frac{1}{L}\,, \frac{1}{T\mu} \ln\left(T \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\frac{3\gamma L}{\mu^2} C \tau_{\text{mix}} \sigma_\star^2} \right) \right)\,,$$

we obtain:

$$\mathbb{E}\left[\|\mathbf{x}_T - \mathbf{x}^\star\|\right] \leqslant 2 e^{-\frac{\mu T}{L}} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \mathcal{O}\left(\frac{L \tau_{\text{mix}} \sigma_\star^2}{\mu^3 T} \right)\,.$$

□

D.2. Proof of Proposition 5.9

Proof. Consider the graph G on the set of nodes $\mathcal{V} = \{0, 1\}$, with probability transitions $p_{01} = p_{10} = p$ and $p_{00} = p_{11} = 1 - p$, for some small $p \in (0, 1)$. The relaxation time $\tau_{\text{mix}}(1/4)$ of this graph scales as $1/p$.

Consider now $f_0(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - 1)^2$ and $f_1(\mathbf{x}) = \frac{1}{2}(\mathbf{x} + 1)^2$ for $\mathbf{x} \in \mathbb{R}$, so that $\mathbf{x}^\star = 0$. For (v_t) a Markov chain with the given transition probabilities, started at v_0 following the uniform (stationary) distribution on $\mathcal{V}$, let $\mathbf{x}_t$ be generated with MC-SGD: $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma \nabla f_{v_t}(\mathbf{x}_t)$ and $\mathbf{x}_0 = 1$, *i.e.*,

$$\mathbf{x}_T = (1-\gamma)^T - \gamma \sum_{t<T} (1-\gamma)^{T-t-1} \zeta_t\,,$$

where $\zeta_t \in \{-1, 1\}$ takes value 1 if $v_t = 0$ and value -1 if $v_t = 0$. We have:

$$\mathbb{E}\left[(\mathbf{x}_T - \mathbf{x}^\star)^2\right] = (1-\gamma)^{2T} + \gamma^2 \sum_{s<t<T} (1-\gamma)^{2T-t-s-2} \mathbb{E}\left[\zeta_s \zeta_t\right] .$$

We compute this second term, and show that it is non-negative for $p \leqslant 1/2$ and of order $\frac{\gamma}{4p}$, so that to reach a given precision $\varepsilon > 0$, is required $\gamma \leqslant 4p\varepsilon$ and thus to make the first term small, T must verify $T = \tilde{\Omega}(1/(2p\varepsilon))$, concluding our reasonning.

For $s < t$, we have $\mathbb{E}\left[\zeta_s \zeta_t\right] = 2\mathbb{P}\left(v_{t-s} = v_0 | v_0 \sim \pi^\star\right) - 1$/ Denoting $z_k = \mathbb{P}\left(v_k = v_0 | v_0 \sim \pi^\star\right)$, we have $z_{k+1} = pz_k + (1-p)(1-z_k)$ and $z_0 = 1$, so that $z_k = \frac{1}{2}(1 + (1-2p)^k)$ for $k \geqslant 0$. This leads to:

$$\gamma^2 \sum_{s<t<T} (1-\gamma)^{2T-t-s-2} \mathbb{E}\left[\zeta_s \zeta_t\right] = \gamma^2 \frac{(1-\gamma)(1-2p)}{1-(1-\gamma)(1-2p)} \times \left(\frac{1-(1-\gamma)^{2T}}{1-(1-\gamma)^2} - (1-2p) \frac{(1-\gamma)^T - (1-2p)^T}{2p-\gamma} (1-\gamma)^T \right)$$

For $\varepsilon \to 0$, in order to have $\mathbb{E}\left[(\mathbf{x}_T - \mathbf{x}^\star)^2\right] \leqslant \varepsilon$, is required $(1-\gamma)^T \leqslant \varepsilon$ so that $\gamma T \to \infty$. Under $\gamma T \to \infty$, we have

$$\gamma^2 \sum_{s<t<T} (1-\gamma)^{2T-t-s-2} \mathbb{E}\left[\zeta_s \zeta_t\right] \sim \frac{\gamma}{4p} .$$

Finally, to reach precision ε, this quantity needs to be upper-bounded by $\varepsilon(1 + o(1))$, so that $\gamma \leqslant 4p\varepsilon^{-1}(1 + o(1))$ is necessary. Plugging this in $(1-\gamma)^{2T} \leqslant \varepsilon$ yields $T = \tilde{\Omega}(p\varepsilon^{-1})$, the desired result. □

E. With local noise: proof of Theorem 5.11

The proof follows the exact same steps as the proof of Theorem 5.6.

Proof. First, note that, using Lemma 15 in Stich and Karimireddy (2021), we have:

$$\mathbb{E}\left[\left\| \sum_{t<T} \nabla_{\mathbf{x}} F_{v_t}(\mathbf{x}^\star, \xi_t) \right\|^2\right] \leqslant 2\mathbb{E}\left[\left\| \sum_{t<T} \nabla f_{v_t}(\mathbf{x}^\star) \right\|^2\right] + 2T\sigma^2_{\text{local}} .$$

We then have the following lemma, proved exactly as in the previous section.

Lemma E.1. *For any $\mathbf{y}_t \in \mathbb{R}^d$ and $t \geqslant 0$, denoting $\mathbf{y}_{t+1} = \mathbf{y}_t - \gamma \nabla_{\mathbf{x}} F_{v_t}(\mathbf{x}^\star, \xi_t)$, we have:*

$$\|\mathbf{x}_{t+1} - \mathbf{y}_{t+1}\|^2 \leqslant (1-\gamma\mu)\|\mathbf{x}_t - \mathbf{y}_t\|^2 + \gamma L \|\mathbf{y}_t - \mathbf{x}^\star\|^2 .$$

This leads to:

$$\|\mathbf{x}_T - \mathbf{y}_T\|^2 \leqslant (1-\gamma\mu)^T \|\mathbf{x}_0 - \mathbf{y}_0\|^2 + \gamma L \sum_{t<T} (1-\gamma\mu)^{T-t} \|\mathbf{y}_t - \mathbf{x}^\star\|^2 ,$$

for

$$\mathbf{y}_0 = \mathbf{x}^\star + \gamma \sum_{t<T} \nabla_{\mathbf{x}} F_{v_t}(\mathbf{x}^\star, \xi_t) , \quad \mathbf{y}_s = \mathbf{x}^\star + \gamma \sum_{s \leqslant t < T} \nabla_{\mathbf{x}} F_{v_t}(\mathbf{x}^\star, \xi_t) , \quad s < T .$$

We thus have:

$$\mathbb{E}\|\mathbf{x}_T - \mathbf{x}^\star\|^2 \leqslant 2(1-\gamma\mu)^T \left(\mathbb{E}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 + \gamma^2 \mathbb{E}\left[\left\| \sum_{s<T} \nabla_{\mathbf{x}} F_{v_s}(\mathbf{x}^\star, \xi_s) \right\|^2\right] \right) + \gamma^3 L \sum_{t<T} (1-\gamma\mu)^{T-t} \mathbb{E}\left[\left\| \sum_{t \leqslant s < T} \nabla_{\mathbf{x}} F_{v_s}(\mathbf{x}^\star, \xi_s) \right\|^2\right] .$$

To conclude, we use the first inequality of this proof, and Lemma D.1, and proceed as in the proof of Theorem 5.6 and Corollary 5.8. □

F. MC-SAG: proof of Theorem 6.1

F.1. With perfect initialization: Theorem 6.1.1

We begin classically by proving a descent lemma. This lemma is deterministic, in the sense that not means $\mathbb{E}$ are present, and it therefore does not use the Markovian properties of the Markov chain. MC-SAG uses biased gradients, even in the case where v_t are *i.i.d.*, since the algorithm SAG (Schmidt et al., 2017) is inherently biased (making it unbiased leads to the SAGA iterations (Defazio et al., 2014)).

Let $\mathbf{G}_t = \bar{\mathbf{h}}_{t+1}$ for $t \geqslant 0$, so that $\mathbf{x}_{t+1} = \mathbf{x}_t - \gamma_t \mathbf{G}_t$. We recall that for $v \in \mathcal{V}$ and $t \geqslant 0$, $p_v(t)$ is the next time (strictly) the chain hits node v, while $d_v(t)$ is either the last time the chain was at the state v (if that happened), or 0 in v has not yet been visited.

Lemma F.1. *Assume that f is L-smooth. Then, for any $t \geqslant 0$, we have:*

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leqslant -\frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{\gamma_t}{4}\|\mathbf{G}_t\|^2 + \frac{\gamma_t L^2}{2n}\sum_{v\in\mathcal{V}}\left\|\sum_{s=d_v(t)}^{t-1}\gamma_s \mathbf{G}_s\right\|^2 .$$

Proof. For $t \geqslant 0$ and $v \in \mathcal{V}$, let $p_v(t) = \inf\{s > t | v_s = v\}$ and $d_v(t) = \sup\{s \leqslant t | v_s = v\}$ be the next and the last previous iterates for which $v_t = v$ ($d_v(t) = 0$ by convention, if v has not yet been visited). Denote $F_t = \mathbb{E}\left[f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right]$. We have, using smoothness:

$$f(\mathbf{x}_{t+1}) \leqslant f(\mathbf{x}_t) - \gamma_t \langle \nabla f(\mathbf{x}_t), \mathbf{G}_t\rangle + \frac{\gamma_t^2 L}{2}\|\mathbf{G}_t\|^2$$

Together with $\langle \nabla f(\mathbf{x}_t), \mathbf{G}_t\rangle = \frac{1}{2}(\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{G}_t\|^2 - \|\nabla f(\mathbf{x}_t) - \mathbf{G}_t\|^2)$, we obtain:

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leqslant f(\mathbf{x}_t) - \frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{G}_t\|^2 - \|\nabla f(\mathbf{x}_t) - \mathbf{G}_t\|^2 + \frac{\gamma_t^2 L}{2}\|\mathbf{G}_t\|^2 \\ &\leqslant f(\mathbf{x}_t) - \frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{\gamma_t}{4}\|\mathbf{G}_t\|^2 + \frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t) - \mathbf{G}_t\|^2 , \end{aligned}$$

as long as $\gamma_t \leqslant 1/(2L)$. We thus need to upperbound the bias $\|\nabla f(\mathbf{x}_t) - \mathbf{G}_t\|^2$. We have:

$$\begin{aligned} \|\nabla f(\mathbf{x}_t) - \mathbf{G}_t\|^2 &= \left\|\frac{1}{n}\sum_{v\in\mathcal{V}} \nabla f_v(\mathbf{x}_{d_v(t)}) - \nabla f_v(\mathbf{x}_t)\right\|^2 \\ &\leqslant \frac{1}{n}\sum_{v\in\mathcal{V}} \left\|\nabla f_v(\mathbf{x}_{d_v(t)}) - \nabla f_v(\mathbf{x}_t)\right\|^2 . \end{aligned}$$

Fix some v in $\mathcal{V}$. We have $\left\|\nabla f_v(\mathbf{x}_t) - \nabla f_v(\mathbf{x}_{d_v(t)})\right\|^2 \leqslant L^2 \left\|\sum_{s=d_v(t)}^{t-1}\gamma_s \mathbf{G}_s\right\|^2$, leading to:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leqslant -\frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{\gamma_t}{4}\|\mathbf{G}_t\|^2 + \frac{\gamma_t L^2}{2n}\sum_{v\in\mathcal{V}}\left\|\sum_{s=d_v(t)}^{t-1}\gamma_s \mathbf{G}_s\right\|^2 .$$

□

We now proceed with the proof of Theorem 6.1.1.

Proof of Theorem 6.1.1. We begin with

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leqslant -\frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{\gamma_t}{4}\|\mathbf{G}_t\|^2 + \frac{\gamma_t L^2}{2n}\sum_{v\in\mathcal{V}}\left\|\sum_{s=d_v(t)}^{t-1}\gamma_s \mathbf{G}_s\right\|^2 ,$$

as a starting point. For $v \in \mathcal{V}$,

$$\gamma_t \left\| \sum_{s=d_v(t)}^{t-1} L^2 \gamma_s \mathbf{G}_s \right\|^2 \leqslant \sum_{s=d_v(t)}^{t-1} (t - d_v(t)) L^2 \gamma_t \gamma_s^2 \|\mathbf{G}_s\|^2$$
$$\leqslant \sum_{s=d_v(t)}^{t-1} L \gamma_s^2 \|\mathbf{G}_s\|^2 ,$$

since $\gamma_t \leqslant 1/(L(t - d_v(t)))$. Summing for $t < T$, we obtain:

$$\sum_{t<T} \frac{\gamma_t}{2} \|\nabla f(\mathbf{x}_t)\|^2 \leqslant F_0 - \sum_{t<T} \frac{\gamma_t}{2} \|\mathbf{G}_t\|^2 + \sum_{t<T} \frac{1}{2n} \sum_{v \in \mathcal{V}} \sum_{s=d_v(t)}^{t-1} L \gamma_s^2 \|\mathbf{G}_s\|^2 .$$

Then,

$$\sum_{t<T} \sum_{s=d_v(t)}^{t-1} L \gamma_s^2 \|\mathbf{G}_s\|^2 = \sum_{s<T} \|\mathbf{G}_s\|^2 L \gamma_s^2 (p_v(s) - s) .$$

For $\gamma_s \leqslant 1/(2L\tau_{\text{hit}})$, we have $\mathbb{E}\left[\|\mathbf{G}_s\|^2 \gamma_s^2 (p_v(s) - s)\right] \leqslant \frac{1}{2}\mathbb{E}\left[\|\mathbf{G}_s\|^2 \gamma_s\right]$, so that:

$$\sum_{t<T} \frac{\gamma_t}{2} \|\nabla f(\mathbf{x}_t)\|^2 \leqslant F_0 + \sum_{t<T} \left(-\frac{\gamma_t}{2} \|\mathbf{G}_t\|^2 + K_t \right) ,$$

where $K_t = \frac{1}{4n} \sum_{v \in \mathcal{V}} \|\mathbf{G}_t\|^2 L \gamma_t^2 (p_v(t) - t)$ verifies $\mathbb{E}\left[K_t | \mathcal{F}_t\right] \leqslant \|\mathbf{G}_t\|^2 \gamma_t / 4$ since $\gamma_t \leqslant 1/(2\tau_{\text{hit}})$, where $\mathcal{F}_t$ is the filtration up to time t:

$$\begin{aligned}
\mathbb{E}\left[K_t | \mathcal{F}_t\right] &= \mathbb{E}\left[\frac{1}{4n} \sum_{v \in \mathcal{V}} \|\mathbf{G}_t\|^2 L \gamma_t^2 (p_v(t) - t) | \mathcal{F}_t \right] \\
&= \frac{1}{4n} \sum_{v \in \mathcal{V}} \|\mathbf{G}_t\|^2 L \gamma_t^2 \mathbb{E}\left[(p_v(t) - t) | \mathcal{F}_t\right] \quad \left(\mathbf{G}_t \,,\, \gamma_t \text{ are } \mathcal{F}_t\text{-measurable}\right) \\
&\leqslant \frac{1}{4n} \sum_{v \in \mathcal{V}} \|\mathbf{G}_t\|^2 L \gamma_t^2 \tau_{\text{hit}} \quad \left(\text{since } \mathbb{E}\left[(p_v(t) - t) | \mathcal{F}_t\right] = \mathbb{E}\left[(p_v(t) - t) | v_t\right] \leqslant \tau_{\text{hit}}\right) \\
&\leqslant \frac{1}{8} \gamma_t \|\mathbf{G}_t\|^2 \quad \left(\text{since } \gamma_t \leqslant 1/(2\tau_{\text{hit}})\right) .
\end{aligned}$$

Finally,

$$\mathbb{E}\left[\min_{t<T} \|\nabla f(\mathbf{x}_t)\|^2 \right] \leqslant \mathbb{E}\left[\frac{2F_0}{\sum_{t<T} \gamma_t} \right] + \mathbb{E}\left[\frac{\sum_{t<T} \left(-\frac{\gamma_t}{2} \|\mathbf{G}_t\|^2 + K_t \right)}{\sum_{t<T} \gamma_t} \right] .$$

Using Lemma A.6 and the above bound on $\mathbb{E}\left[K_t | \mathcal{F}_t\right]$, that last term is non-positive. Using Jensen inequality, we have:

$$\mathbb{E}\left[\min_{t<T} \|\nabla f(\mathbf{x}_t)\|^2 \right] \leqslant \frac{2F_0}{T^2} \sum_{t<T} \mathbb{E}\left[\gamma_t^{-1}\right] .$$

Since $\gamma_t^{-1} = 2L(\tau_{\text{hit}} + \sup_{v \in \mathcal{V}}(t - d_v(t)))$, we have $\sum_{t<T} \mathbb{E}\left[\gamma_t^{-1}\right] \leqslant 2LT(\tau_{\text{hit}} + \tau_{\text{cov}})$ using Lemma A.5, concluding the proof. □

F.2. With arbitrary initialization: proof of Theorem 6.1.2

We no longer assume that $\mathbf{G}_0 = \nabla f(\mathbf{x}_0)$ or that $\mathbf{h}_v = \nabla f_v(\mathbf{x}_0)$: $\mathbf{h}_v$ are arbitrary and $\mathbf{G}_0 = \frac{1}{n}\sum_{v\in\mathcal{V}} \mathbf{h}_v$. In that case, we have the following descent lemma.

Lemma F.2. *Assume that f is L-smooth. Let*

$$\tilde{\tau}_{\text{cov}} = \inf\left\{t \geqslant 1 \quad \textit{such that} \quad \{v_0, \dots, v_{t-1}\} = \mathcal{V}\right\}.$$

Then, for any $t \geqslant 0$, we have:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leqslant -\frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{\gamma_t}{4}\|\mathbf{G}_t\|^2 + \frac{\gamma_t L^2}{n}\sum_{v\in\mathcal{V}}\left\|\sum_{s=d_v(t)}^{t-1}\gamma_s \mathbf{G}_s\right\|^2 + \frac{\gamma_t \mathbb{1}_{t<\tilde{\tau}_{\text{cov}}}}{n}\sum_{v\in\mathcal{V}}\|\nabla f_v(\mathbf{x}_0) - \mathbf{h}_v\|^2.$$

Proof. As in Lemma F.1, we have

$$f(\mathbf{x}_{t+1}) \leqslant f(\mathbf{x}_t) - \frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{\gamma_t}{4}\|\mathbf{G}_t\|^2 + \frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t) - \mathbf{G}_t\|^2,$$

as long as $\gamma_t \leqslant 1/(2L)$. We thus again need to upperbound the bias $\|\nabla f(\mathbf{x}_t) - \mathbf{G}_t\|^2$. For $t \geqslant \tilde{\tau}_{\text{cov}}$, we have $\mathbf{G}_t = \frac{1}{n}\sum_{v\in\mathcal{V}} \nabla f_v(\mathbf{x}_{d_v(t)})$, and so:

$$\|\nabla f(\mathbf{x}_t) - \mathbf{G}_t\|^2 \leqslant \frac{1}{n}\sum_{v\in\mathcal{V}}\left\|\nabla f_v(\mathbf{x}_{d_v(t)}) - \nabla f_v(\mathbf{x}_t)\right\|^2,$$

hence the result of Lemma F.1 holds for $t \leqslant \tilde{\tau}_{\text{cov}}$ (and so does that of the Lemma we are proving). Then, for $t \leqslant \tilde{\tau}_{\text{cov}}$, $\nabla f(\mathbf{x}_t) - \mathbf{G}_t = \frac{1}{n}\sum_{v\in\mathcal{V}} \nabla f_v(\mathbf{x}_t) - \mathbf{h}_v^t$, where $\mathbf{h}_v^t = \mathbf{h}_v$ if v hasn't been visited yet, or $\mathbf{h}_v^t = \nabla f_v(\mathbf{x}_{d_v(t)})$ otherwise. Hence,

$$\begin{aligned}\|\nabla f(\mathbf{x}_t) - \mathbf{G}_t\|^2 &\leqslant \frac{1}{n}\sum_{v\in\mathcal{V}}\left\|\mathbf{h}_v^t - \nabla f_v(\mathbf{x}_t)\right\|^2 \\ &\leqslant \frac{2}{n}\sum_{v\in\mathcal{V}}\left\|\nabla f_v(\mathbf{x}_{d_v(t)}) - \nabla f_v(\mathbf{x}_t)\right\|^2 + \frac{2}{n}\sum_{v\in\mathcal{V}}\|\nabla f_v(\mathbf{x}_0) - \mathbf{h}_v\|^2\end{aligned}$$

Since $\left\|\nabla f_v(\mathbf{x}_t) - \nabla f_v(\mathbf{x}_{d_v(t)})\right\|^2 \leqslant L^2\left\|\sum_{s=d_v(t)}^{t-1}\gamma_s \mathbf{G}_s\right\|^2$, we have:

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leqslant -\frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{\gamma_t}{4}\|\mathbf{G}_t\|^2 + \frac{\gamma_t L^2}{n}\sum_{v\in\mathcal{V}}\left\|\sum_{s=d_v(t)}^{t-1}\gamma_s \mathbf{G}_s\right\|^2 + \frac{\gamma_t}{n}\sum_{v\in\mathcal{V}}\|\nabla f_v(\mathbf{x}_0) - \mathbf{h}_v\|^2,$$

leading to the desired result. □

Proof of Theorem 6.1.2. Starting from

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leqslant -\frac{\gamma_t}{2}\|\nabla f(\mathbf{x}_t)\|^2 - \frac{\gamma_t}{4}\|\mathbf{G}_t\|^2 + \frac{\gamma_t L^2}{n}\sum_{v\in\mathcal{V}}\left\|\sum_{s=d_v(t)}^{t-1}\gamma_s \mathbf{G}_s\right\|^2 + \frac{\gamma_t \mathbb{1}_{t<\tilde{\tau}_{\text{cov}}}}{n}\sum_{v\in\mathcal{V}}\|\nabla f_v(\mathbf{x}_0) - \mathbf{h}_v\|^2,$$

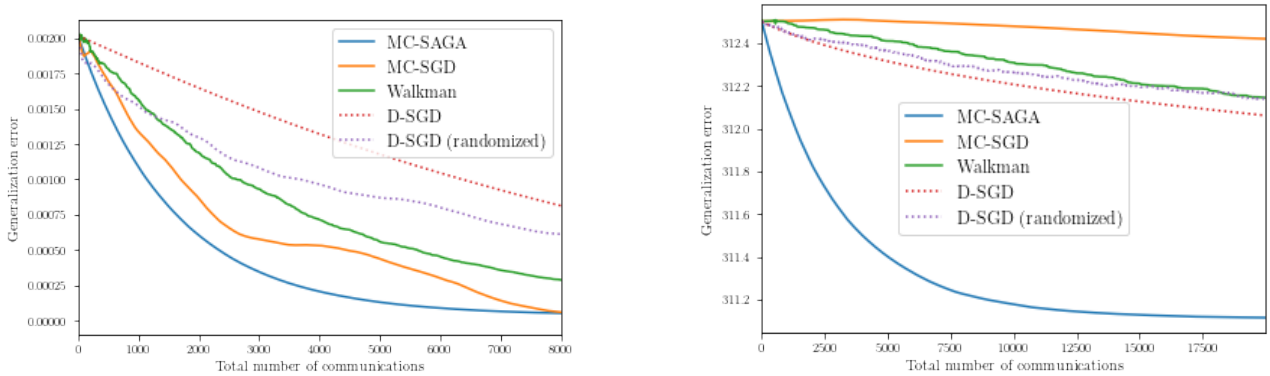
and mimicking the proof with initialization, we obtain that for $\gamma_t^{-1} = 4L(\tau_{\text{hit}} + \sup_{v\in\mathcal{V}}(t - d_v(t)))$, we have

$$\mathbb{E}\left[\min_{t<T}\|\nabla f(\mathbf{x}_t)\|^2\right] \leqslant \frac{2F_0}{T^2}\sum_{t<T}\mathbb{E}\left[\gamma_t^{-1}\right] + \frac{2\mathbb{E}\left[\tilde{\tau}_{\text{cov}}\right]}{T}\frac{1}{n}\sum_{v\in\mathcal{V}}\|\nabla f_v(\mathbf{x}_0) - \mathbf{h}_v\|^2,$$

and we conclude the proof by noticing that $\mathbb{E}\left[\tilde{\tau}_{\text{cov}}\right] \leqslant \tau_{\text{cov}}$. □

G. Numerical illustration of our theory

Setting We place ourselves in the decentralized optimization setting on a graph G with local functions f_v, and build two toy problems. We compare our algorithms MC-SGD and MC-SAGA with Walkman (Mao et al., 2020) and decentralized SGD (D-SGD in Figure 1) (Koloskova et al., 2020; Yu et al., 2019) with both randomized gossip communications and fixed gossip matrix. We consider the non-convex loss function $\ell(\mathbf{x}, a, b) = (\sigma(\mathbf{x}^\top a) - b)^2/2$ where $\sigma(t) = 1/(1 + \exp(-t))$ as in Mei et al. (2018). For $v \in \mathcal{V}$, we take $f_v(x) = \ell(\mathbf{x}, a_v, b_v)$ for a_v and b_v random variables. In Figure 1(a), we take a connected random geometric graph of $n = 50$ nodes in $[0, 1]^2$ with radius parameter $\rho = 0.3$ (nodes are connected if their distance is less than ρ). We consider homogeneous data: a_v, b_v are taken *i.i.d.*, with $a_v \sim \mathcal{N}(0, 1)$ and b_v uniform in $[0, 1]$. In Figure 1(b), we take the cycle of $n = 50$ nodes and consider heterogeneous data: for two opposite nodes in the graph ($v = 0$ and $v = 25$ *e.g.*), a_v are $\mathcal{N}(0, 1)$ and b_v uniform in $[0, 1]$ (and the functions are re-normalized), while they are taken equal to 0 in the rest of the graph. In Figure 1(a), the graph is well connected and Assumption 5.1 is verified for a small enough $\bar{\sigma}^2$, so that all algorithms perform comparatively well; MC-SGD because of the homogeneity, the gossip-based ones and Walkman thanks to the connectivity. However, decreasing the connectivity and increasing data heterogeneity in a pathological way, we obtain Figure 1(b): MC-SGD fails due to function-heterogeneity ($\bar{\sigma}^2$ too big), while the three others are slowed down by their communication inefficiency, illustrating how depending only on τ_{hit}, MC-SAGA outperforms the other algorithms, that rather depend on $n\tau_{\text{mix}}$ or $n\tau_{\text{mix}}^2$.



(a) Geometric graph, homogeneous functions

(b) Cycle graph, heterogeneous functions

Figure 1: Comparison of MC-SAGA and MC-SGD with existing algorithms. In abscissa are the total number of communications, and in ordinate $f(\mathbf{x}) - f(\mathbf{x}^\star)$.