Accelerated Stochastic Gradient Method with Applications to Consensus Problem in Markov-Varying Networks [*]

Vladimir Solodkin[1,2], Savelii Chezhegov[1,2], Ruslan Nazikov[1,2], Aleksandr Beznosikov[1,2,3], and Alexander Gasnikov[3,1,2]

[1] Moscow Institute of Physics and Technology, Moscow, Russian Federation
[2] Institute for Information Transmission Problems RAS, Moscow, Russian Federation
[3] Innopolis University, Innopolis, Russian Federation

Abstract. Stochastic optimization is a vital field in the realm of mathematical optimization, finding applications in diverse areas ranging from operations research to machine learning. In this paper, we introduce a novel first-order optimization algorithm designed for scenarios where Markovian noise is present, incorporating Nesterov acceleration for enhanced efficiency. The convergence analysis is performed using an assumption on noise depending on the distance to the solution. We also delve into the consensus problem over Markov-varying networks, exploring how this algorithm can be applied to achieve agreement among multiple agents with differing objectives during changes in the communication system. To show the performance of our method on the problem above, we conduct experiments to demonstrate the superiority over the classic approach.

Keywords: convex optimization · stochastic optimization · Markovian noise · accelerated methods · decentralized communications

1 Introduction

Stochastic optimization encompasses a suite of methodologies aimed at minimizing or maximizing an objective function in the presence of randomness. These methods have evolved into indispensable tools across a spectrum of disciplines including science, engineering, business, computer science, and statistics. Applications are diverse, ranging from refining the placement of acoustic sensors on a beam through simulations, to determining optimal release times for reservoir water to maximize hydroelectric power generation, to fine-tuning the parameters of statistical models based on given datasets. The introduction of randomness typically occurs through the cost function or the constraint set. While the term "stochastic optimization" may encompass any optimization approach that incorporates randomness within certain communities, our focus here is on scenarios

---

where the objective function is stochastic.

As with deterministic optimization, no universal solution method generally excels across all problems. Structural assumptions play a pivotal role in making problems tractable. Given that solution methodologies are intricately linked to problem structures, our analysis relies heavily on problem type, with a detailed exposition of associated solution approaches.

Related work. A considerable body of research has documented substantial advancements achieved by accelerating gradient descent in a Nesterov manner [37]. Building upon this foundation, Nesterov-accelerated stochastic gradient descent [2,5] emerged as a powerful tool for optimizing different objectives in stochastic settings. In the earlier works [39,43], the proof of convergence was done using an assumption on bounded variance, which significantly narrows the application perspective. Later, [40] succeeded in relaxing this assumption to strong growth condition, which partially solved the aforementioned problem. At the same time, several papers delved into applying acceleration to specific stochastic cases, e.g., coordinate descent [38], heavy tailed noise [41], distributed learning [42]. However, all of these works investigate i.i.d. noise setup, while a more general case could be considered.

As of late, there has been an emergence of scholarly works aimed at addressing the existing gap in the analysis of Markovian noise configuration. Nonetheless, it is noteworthy that this domain continues to be a dynamically evolving field of study. Specifically, [14] examined a variant of the Ergodic Mirror Descent algorithm yielding optimal convergence rates for smooth and nonconvex problems. More recently, [18] proposed a random batch size algorithm tailored for nonconvex optimization within a compact domain. In the Markovian noise domain, the finite-time analysis of non-accelerated SGD-type algorithms has been investigated in [19] and [21]. However, [19] relies heavily on the assumption of a bounded domain and uniformly bounded stochastic gradient oracles and [21] achieves only suboptimal dependence on initial conditions for strongly convex problems when employing SGD. In the exploration of accelerated SGD in the presence of Markovian noise, [22] achieved an optimal rate of initial condition forgetting, but suboptimal variance terms. Recently, [1] proposed the accelerated version of SGD achieving linear dependence on the mixing time.

The aforementioned studies predominantly address general Markovian noise optimization. Recently, a surge of papers has emerged, focusing on the specialized scenario of distributed optimization [24,25]. [26] investigates the generalization and stability of Markov SGD with specific emphasis on excess variance guarantees. Simultaneously, specific results such as those from [27] offer lower bounds for particular finite-sum problems within the Markovian setting.

Our contributions. We present the analysis of an accelerated version of SGD in the Markovian noise setting under the assumption of a gradient estimator bounded by the distance to the optimum. We obtain sharp convergence rate and optimal dependence in terms of the mixing time of the underlying Markov chain.

Moreover, for $k = 1$ Markovian scheme reduces to a classical *i.i.d.* noise setup. To the best of our knowledge, analysis in this case (even for *i.i.d.* stochasticity) under suggested assumptions has not been presented in the literature before. To show the practicality of our method, we perform numerical experiments on the consensus search problem on time-varying networks and show a better convergence rate compared to classical approaches for solving this problem.

1.1   Technical Preliminaries

Let $(\mathsf{Z}, \mathsf{d_Z})$ be a complete separable metric space endowed with its Borel σ-field $\mathcal{Z}$. Let $(\mathsf{Z}^{\mathbb{N}}, \mathcal{Z}^{\otimes \mathbb{N}})$ be the corresponding canonical process. Consider the Markov kernel Q defined on $\mathsf{Z} \times \mathcal{Z}$, and denote by $\mathbb{P}_\xi$ and $\mathbb{E}_\xi$ the corresponding probability distribution and the expected value with initial distribution ξ. Without loss of generality, we assume that $(Z_k)_{k \in \mathbb{N}}$ is the corresponding canonical process. By construction, for any $A \in \mathcal{Z}$, it holds that $\mathbb{P}_\xi(Z_k \in A | Z_{k-1}) = \mathsf{Q}(Z_{k-1}, A)$, $\mathbb{P}_\xi$-a.s. If $\xi = \delta_z$, $z \in \mathsf{Z}$, we write $\mathbb{P}_z$ and $\mathbb{E}_z$ instead of $\mathbb{P}_{\delta_z}$ and $\mathbb{E}_{\delta_z}$, respectively. For $x^1, \dots, x^k$ being the iterates of any stochastic first-order method, we denote $\mathcal{F}_k = \sigma(x^j, j \leq k)$ and write $\mathbb{E}_k$ as an alias for $\mathbb{E}[\cdot | \mathcal{F}_k]$.

Lemma 1 (*Cauchy Schwartz inequality*). *For any $a, b, x_1, \dots, x_n \in \mathbb{R}^d$ and $c > 0$ the following inequalities hold:*

$$2\langle a, b \rangle \leq \frac{\|a\|^2}{c} + c\|b\|^2, \tag{1}$$

$$\|a + b\|^2 \leq \left(1 + \frac{1}{c}\right)\|a\|^2 + (1+c)\|b\|^2. \tag{2}$$

2   Problem and Assumptions

In this paper, we study the minimization problem

$$\min_{x \in \mathbb{R}^d}\left[f(x) := \mathbb{E}_{Z \sim \pi}[F(x, Z)]\right], \tag{3}$$

where access to the function f and its gradient are available only through the noisy oracle $F(x, Z)$ and $\nabla F(x, Z)$ respectively. We start by presenting two classical regularity constraints on the target function f:

Assumption 1 *The function f is L-smooth on $\mathbb{R}^d$ with $L > 0$, i.e. it is continuously differentiable and there exists a constant $L > 0$ such that the following inequality holds for all $x, y \in \mathbb{R}^d$:*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

Assumption 2 *The function f is μ-strongly convex on $\mathbb{R}^d$, i.e. it is continuously differentiable, and there exists a constant $\mu > 0$ such that the following inequality holds for all $x, y \in \mathbb{R}^d$:*

$$\frac{\mu}{2}\|x - y\|^2 \leq f(x) - f(y) - \langle \nabla f(y), x - y \rangle.$$

Next we specialize our assumption on the sequence of noise variables $\{Z_i\}_{i=0}^{\infty}$. Assumption 3 is also considered to be classical in the case of stochastic optimization with the Markovian noise [18,19,22]. It allows us to deal with finite state-space Markov chains with irreducible and aperiodic transition matrix.

Assumption 3 $\{Z_i\}_{i=0}^{\infty}$ *is a stationary Markov chain on* $(\mathsf{Z}, \mathcal{Z})$ *with Markov kernel* Q *and unique invariant distribution* π. *Moreover,* Q *is uniformly geometrically ergodic with mixing time* $\tau \in \mathbb{N}$, *i.e., for every* $k \in \mathbb{N}$,

$$\Delta(\mathsf{Q}^k) = \sup_{z,z' \in \mathsf{Z}} (1/2)\|\mathsf{Q}^k(z,\cdot) - \mathsf{Q}^k(z',\cdot)\|_{\mathsf{TV}} \le (1/4)^{\lfloor k/\tau \rfloor}.$$

Now we specify our assumption on the stochastic gradient estimator. The majority of existing literature on stochastic first order methods for solving (3) utilizes *strong growth condition* [40] or *uniformly bounded variance* [39] as they allow to prove the convergence quite straightforwardly. However, these assumptions narrow down the set of target functions that can be considered rather strongly and there are several kinds of relaxation of it [9], where gradient differences are bounded by the norm of the true gradient and a certain bias. Instead of this, we propose to use the following assumption:

Assumption 4 *For all* $x \in \mathbb{R}^d$ *it holds that* $\mathbb{E}_{\pi}[\nabla F(x,Z)] = \nabla f(x)$. *Moreover, for all* $z \in \mathsf{Z}$ *and* $x \in \mathbb{R}^d$ *it holds that*

$$\|\nabla F(x,z) - \nabla f(x)\|^2 \le \sigma^2 + \delta^2\|x - x^*\|^2. \tag{4}$$

It is one way or another much weaker, then strong grows condition and uniformly bounded variance and, to the best of our knowledge, seem to be new for analyzing accelerated methods for solving stochastic optimization problems. One can notice that unlike the i.i.d. case, we are forced to require the almost sure bound in (4) rather than in expectation. This issue inevitably arises when dealing with Markovian stochasticity due to the impossibility of using the expectation trick [20], and has not yet been solved by any authors dealing with such type of stochasticity [18,19,21]. Either way, there are advantages to this approach as well. If we additionally require our noisy oracle $F(x,Z)$ to be $\tilde{L}-$Lipschiz, then Assumption 4 is automatically satisfied. Formally, if for any $x, y \in \mathbb{R}^d$,

$$\|\nabla F(x,z) - \nabla F(y,z)\| \le \tilde{L}(z)\|x - y\|,$$

for $\tilde{L} : \mathsf{Z} \to \mathbb{R}^+$ with $\sup |\tilde{L}| < \infty$, then

$$\begin{aligned}
\|\nabla F(x,z) - \nabla f(x)\|^2 &\le 3\|\nabla F(x,z) - \nabla F(x^*,z)\|^2 + 3\|\nabla F(x^*,z) - \nabla f(x^*)\|^2 \\
&\quad + 3\|\nabla f(x) - \nabla f(x^*)\|^2 \\
&\le 3(\|\tilde{L}\|^2 + L^2)\|x - x^*\|^2 + 3\|\nabla F(x^*,z) - \nabla f(x^*)\|^2,
\end{aligned}$$

taking $\sigma = \sqrt{3\|\nabla F(x^*,z) - \nabla f(x^*)\|}$ and $\delta = \sqrt{6\max(L,\|\tilde{L}\|)}$ gives Assumption 4.

3   Main results

We start by introducing our version of Nesterov accelerated SGD. It utilizes the idea from [1] of using exactly the number of samples that comes from the truncated geometric distribution with truncation parameter to be specified later (see Theorem 1) in order to obtain optimal computational complexity of the algorithm.

Algorithm 1 Markov Accelerated GD

1: **Parameters:** stepsize $\gamma > 0$, momentums θ, η, number of iterations N, batchsize limit M
2: **Initialization:** choose $x^0 = x_f^0$
3: **for** $k = 0, 1, 2, \ldots, N-1$ **do**
4:     $x_g^k = \theta x_f^k + (1-\theta)x^k$
5:     Sample $J_k \sim \mathrm{Geom}\,(1/2)$
6:     $g^k = g_0^k + \begin{cases} 2^{J_k}\big(g_{J_k}^k - g_{J_k-1}^k\big), \text{if } 2^{J_k} \leq M \\ 0, \qquad\qquad\qquad \text{otherwise} \end{cases}$ with $g_j^k = \frac{1}{2^j}\sum_{i=1}^{2^j} \nabla F(x_g^k, Z_{T^k+i})$
7:     $x_f^{k+1} = x_g^k - \gamma g^k$
8:     $x^{k+1} = \eta x_f^{k+1} + (1-\eta)x_f^k$
9:     $T^{k+1} = T^k + 2^{J_k}$
10: **end for**

The key idea behind randomized batch size is to reduce the bias of the stochastic gradient estimator. Motivation for this is irrefutably natural as under the Markovian stochastic gradients oracles this bias appears by itself. Indeed, one can easily show the fact that:

$$\mathbb{E}_k[\nabla F(x^k, Z_{T^k+i})] \neq \nabla f(x^k)\,.$$

In a subsequent part, we show how the bias of the gradient estimator introduced in line 6 of Algorithm 1 scales with the truncation parameter M. To obtain proper dependence, we first need to introduce auxiliary Lemma 2, which is to constrain the gradient estimator with a simpler structure. In particular, we bound MSE for sample average approximation computed over batch size n under arbitrary initial distribution. We emphasise that it is extremely essential to have the bound for MSE under arbitrary initial distribution ξ, because in the proof of our Theorem 1 we will unavoidably manage the conditional expectations w.r.t. the previous iterate.

Lemma 2. *Consider Assumptions 3 and 4. Then, for any $n \geq 1$ and $x \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}_\pi\Big[\|\frac{1}{n}\sum_{i=1}^n \nabla F(x, Z_i) - \nabla f(x)\|^2\Big] \leq \frac{8\tau}{n}\left(\sigma^2 + \delta^2\|x - x^*\|^2\right)\,. \qquad (5)$$

Moreover, for any initial distribution ξ on $(\mathsf{Z}, \mathcal{Z})$, that

$$\mathbb{E}_\xi \left[\| \frac{1}{n} \sum_{i=1}^n \nabla F(x, Z_i) - \nabla f(x) \|^2 \right] \leq \frac{C_1 \tau}{n} \left(\sigma^2 + \delta^2 \|x - x^*\|^2 \right), \qquad (6)$$

where $C_1 = 16(1 + \frac{1}{\ln^2 4})$.

Proof. By [31, Lemma 19.3.6 and Theorem 19.3.9], for any two probabilities ξ, ξ' on $(\mathsf{Z}, \mathcal{Z})$ there is a *maximal exact coupling* $(\Omega, \mathcal{F}, \tilde{\mathbb{P}}_{\xi,\xi'}, Z, Z', T)$ of $\mathbb{P}_\xi^{\mathrm{Q}}$ and $\mathbb{P}_{\xi'}^{\mathrm{Q}}$, that is,

$$\|\xi \mathrm{Q}^n - \xi' \mathrm{Q}^n\|_{\mathrm{TV}} = 2\tilde{\mathbb{P}}_{\xi,\xi'}(T > n). \qquad (7)$$

We write $\tilde{\mathbb{E}}_{\xi,\xi'}$ for the expectation with respect to $\tilde{\mathbb{P}}_{\xi,\xi'}$. Using the coupling construction (7),

$$\mathbb{E}_\xi^{1/2} \left[\| \sum_{i=1}^n \{\nabla F(x, Z_i) - \nabla f(x)\} \|^2 \right] \leq \mathbb{E}_\pi^{1/2} \left[\| \sum_{i=0}^{n-1} \nabla F(x, Z_i) - \nabla f(x) \|^2 \right] +$$

$$\tilde{\mathbb{E}}_{\xi,\pi}^{1/2} \left[\| \sum_{i=0}^{n-1} \{\nabla F(x, Z_i) - \nabla F(x, Z_i')\} \|^2 \right].$$

The first term is bounded with (5). Moreover, with (7) and Assumption 4, we get

$$\| \sum_{i=0}^{n-1} \{\nabla F(x, Z_i) - \nabla F(x, Z_i')\} \|^2 \leq 8 \left(\sigma^2 + \delta^2 \|x - x^*\|^2 \right) \left(\sum_{i=0}^{n-1} \mathbb{1}_{\{Z_i \neq Z_i'\}} \right)^2$$

$$= 8 \left(\sigma^2 + \delta^2 \|x - x^*\|^2 \right) \left(\sum_{i=0}^{n-1} \mathbb{1}_{\{T > i\}} \right)^2$$

$$\leq 16 \left(\sigma^2 + \delta^2 \|x - x^*\|^2 \right) \sum_{i=1}^\infty i \, \mathbb{1}_{\{T > i\}}.$$

Thus, using the Assumption 3, we bound

$$\tilde{\mathbb{E}}_{\xi,\pi} \left[\sum_{i=1}^\infty i \, \mathbb{1}_{\{T > i\}} \right] = \sum_{i=1}^\infty i \tilde{\mathbb{P}}_{\xi,\xi'}(T > i) = \sum_{i=1}^\infty i(1/4)^{\lfloor i/\tau \rfloor} \leq 4 \sum_{i=1}^\infty i(1/4)^{i/\tau}.$$

Now we set $\rho = (1/4)^{1/\tau}$ and use an upper bound

$$\sum_{k=1}^\infty k\rho^k \leq \rho^{-1} \int_0^{+\infty} x^p \rho^x \, dx \leq \rho^{-1} \left(\ln \rho^{-1} \right)^{-2} \Gamma(2)$$

$$= \rho^{-1} \left(\ln \rho^{-1} \right)^{-2} = \frac{\tau^2}{(1/4)^{1/\tau} \ln^2 4}.$$

Combining the bounds above yields

$$\mathbb{E}_\xi\left[\|\frac{1}{n}\sum_{i=1}^n \nabla F(x, Z_i) - \nabla f(x)\|^2\right] \leq \left(\frac{c_1\tau}{n} + \frac{c_2\tau^2}{n^2}\right)\left(\sigma^2 + \delta^2\|x - x^*\|^2\right),$$

where $c_1 = 16$, $c_2 = \frac{128(1/4)^{-1/\tau}}{\ln^2 4}$. Now we consider the two cases. If $n < c_1\tau$, we get from Minkowski's inequality that

$$\mathbb{E}_\xi\left[\|\frac{1}{n}\sum_{i=1}^n \nabla F(x, Z_i) - \nabla f(x)\|^2\right] \leq 2\sigma^2 + 2\delta^2\|x - x^*\|^2,$$

and (6) holds. If $n > c_1\tau$, it holds that

$$\frac{c_2\tau^2}{n^2}\left(\sigma^2 + \delta^2\|\nabla f(x)\|^2\right) \leq \frac{c_2\tau^2}{nc_1\tau}\left(\sigma^2 + \delta^2\|x - x^*\|^2\right),$$

and we gain (6) too. $\square$

We are now ready to bound the MSE for the gradient estimator introduced in line 6 of Algorithm 1. From Lemma 3, we obtain a desired linear dependence of the error reduction on the parameter M.

Lemma 3. *Consider Assumptions 3 and 4. Then for the gradient estimates g^k from line 6 Algorithm 1 it holds that $\mathbb{E}_k[g^k] = \mathbb{E}_k[g^k_{\lfloor \log_2 M\rfloor}]$. Moreover,*

$$\mathbb{E}_k[\|\nabla f(x_g^k) - g^k\|^2] \leq 13C_1\tau\log_2 M(\sigma^2 + \delta^2\|x_g^k - x^*\|^2), \tag{8}$$
$$\|\nabla f(x_g^k) - \mathbb{E}_k[g^k]\|^2 \leq 2C_1\tau M^{-1}(\sigma^2 + \delta^2\|x_g^k - x^*\|^2),$$

where C_1 is defined in (6).

Proof. To show that $\mathbb{E}_k[g^k] = \mathbb{E}_k[g^k_{\lfloor \log_2 M\rfloor}]$ we simply compute conditional expectation w.r.t. J_k:

$$\mathbb{E}_k[g^k] = \mathbb{E}_k\left[\mathbb{E}_{J_k}[g^k]\right] = \mathbb{E}_k[g_0^k] + \sum_{i=1}^{\lfloor \log_2 M\rfloor} \mathbb{P}\{J_k = i\}\cdot 2^i\mathbb{E}_k[g_i^k - g_{i-1}^k]$$

$$= \mathbb{E}_k[g_0^k] + \sum_{i=1}^{\lfloor \log_2 M\rfloor} \mathbb{E}_k[g_i^k - g_{i-1}^k] = \mathbb{E}_k[g^k_{\lfloor \log_2 M\rfloor}]. \tag{9}$$

We start with the proof of the first statement of (8) by taking the conditional expectation for J_k:

$$\mathbb{E}_k[\|\nabla f(x_g^k) - g^k\|^2] \leq 2\mathbb{E}_k[\|\nabla f(x_g^k) - g_0^k\|^2] + 2\mathbb{E}_k[\|g^k - g_0^k\|^2]$$

$$= 2\mathbb{E}_k[\|\nabla f(x_g^k) - g_0^k\|^2] + 2\sum_{i=1}^{\lfloor \log_2 M\rfloor}\mathbb{P}\{J_k = i\}\cdot 4^i\mathbb{E}_k[\|g_i^k - g_{i-1}^k\|^2]$$

$$= 2\mathbb{E}_k[\|\nabla f(x_g^k) - g_0^k\|^2] + 2\sum_{i=1}^{\lfloor \log_2 M\rfloor} 2^i\mathbb{E}_k[\|g_i^k - g_{i-1}^k\|^2]$$

$$\leq 2\mathbb{E}_k[\|\nabla f(x_g^k) - g_0^k\|^2] +$$
$$+ 4\sum_{i=1}^{\lfloor \log_2 M\rfloor} 2^i \left(\mathbb{E}_k[\|\nabla f(x_g^k) - g_{i-1}^k\|^2 + \mathbb{E}_k[\|g_i^k - \nabla f(x_g^k)\|^2]\right).$$

To bound $\mathbb{E}_k[\|\nabla f(x_g^k) - g_0^k\|^2]$, $\mathbb{E}_k[\|\nabla f(x_g^k) - g_{i-1}^k\|^2]$, $\mathbb{E}_k[\|g_i^k - \nabla f(x_g^k)\|^2]$, we apply Lemma 2 and get

$$\mathbb{E}_k[\|\nabla f(x_g^k) - g^k\|^2]$$
$$\leq 2\sigma^2 + 2\delta^2\|x_g^k - x^*\|^2 + 12\sum_{i=1}^{\lfloor \log_2 M\rfloor} 2^i \cdot \frac{C_1\tau}{2^i}(\sigma^2 + \delta^2\|x_g^k - x^*\|^2)$$
$$\leq 13C_1\tau \log_2 M(\sigma^2 + \delta^2\|x_g^k - x^*\|^2).$$

To show the second part of the statement, we use (9) and get

$$\|\nabla f(x_g^k) - \mathbb{E}_k[g^k]\|^2 = \|\nabla f(x^k) - \mathbb{E}_k[g_{\lfloor \log_2 M\rfloor}^k]\|^2.$$

Using Lemma 2 and $2^{\lfloor \log_2 M\rfloor} \geq M/2$ finishes the proof. $\qquad\square$

We also note that our proofs of Lemma 2 and Lemma 3 rely on the proofs of Lemmas 1 and 2 of [1], but for the sake of clarity of the narrative we give them in full.

Now, before we move on to the proof of our major result, we first need to introduce two descent lemmas:

Lemma 4. *Consider Assumptions 1 and 2 be satisfied. Then for the iterates of Algorithm 1 with $\theta = (1 - \eta)/(\beta - \eta)$, $\theta > 0$, $\eta \geq 1$, it holds that*

$$\begin{aligned}
\mathbb{E}_k[\|x^{k+1} - x^*\|^2] \leq &(1 + \alpha\gamma\eta)(1 - \beta)\|x^k - x^*\|^2 + (1 + \alpha\gamma\eta)\beta\|x_g^k - x^*\|^2 \\
&+ (1 + \alpha\gamma\eta)(\beta^2 - \beta)\|x^k - x_g^k\|^2 + \eta^2\gamma^2\mathbb{E}_k[\|g^k\|^2] \\
&- 2\eta\gamma\langle\nabla f(x_g^k), \eta x_g^k + (1 - \eta)x_f^k - x^*\rangle \\
&+ \frac{\eta\gamma}{\alpha}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2,
\end{aligned} \tag{10}$$

where $\alpha > 0$ is any positive constant.

Proof. We start with lines 8 and 7 of Algorithm 1:

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 &= \|\eta x_f^{k+1} + (1 - \eta)x_f^k - x^*\|^2 = \|\eta x_g^k - \eta\gamma g^k + (1 - \eta)x_f^k - x^*\|^2 \\
&= \|\eta x_g^k + (1 - \eta)x_f^k - x^*\|^2 + \gamma^2\eta^2\|g^k\|^2 - 2\gamma\eta\langle g^k, \eta x_g^k + (1 - \eta)x_f^k - x^*\rangle.
\end{aligned}$$

Using straightforward algebra, we get

$$\begin{aligned}
\|x^{k+1} - x^*\|^2 = &\|\eta x_g^k + (1 - \eta)x_f^k - x^*\|^2 - 2\gamma\eta\langle\nabla f(x_g^k), \eta x_g^k + (1 - \eta)x_f^k - x^*\rangle \\
&- 2\gamma\eta\langle\mathbb{E}_k[g^k] - \nabla f(x_g^k), \eta x_g^k + (1 - \eta)x_f^k - x^*\rangle + \gamma^2\eta^2\|g^k\|^2 \\
&- 2\gamma\eta\langle g^k - \mathbb{E}_k[g^k], \eta x_g^k + (1 - \eta)x_f^k - x^*\rangle \\
\leq &(1 + \alpha\eta\gamma)\|\eta x_g^k + (1 - \eta)x_f^k - x^*\|^2 + \frac{\gamma\eta}{\alpha}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2.
\end{aligned}$$

$$- 2\gamma\eta\langle\nabla f(x_g^k), \eta x_g^k + (1-\eta)x_f^k - x^*\rangle + \gamma^2\eta^2\|g^k\|^2$$
$$- 2\gamma\eta\langle g^k - \mathbb{E}_k[g^k], \eta x_g^k + (1-\eta)x_f^k - x^*\rangle$$

In the last step we also applied Cauchy-Schwartz inequality in the form (1) with $c > 0$. Taking the conditional expectation, we get

$$\mathbb{E}_k[\|x^{k+1} - x^*\|^2] \leq (1 + \alpha\eta\gamma)\|\eta x_g^k + (1-\eta)x_f^k - x^*\|^2$$
$$- 2\gamma\eta\langle\nabla f(x_g^k), \eta x_g^k + (1-\eta)x_f^k - x^*\rangle$$
$$+ \gamma^2\eta^2\mathbb{E}_k[\|g^k\|^2] + \frac{\gamma\eta}{\alpha}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2. \qquad (11)$$

Now let us handle expression $\|\eta x_g^k + (1-\eta)x_f^k - x^*\|^2$ for a while. Taking into account line 4 and the choice of θ such that $\theta = (1-\eta)/(\beta - \eta)$ (in particular, $\beta = \eta + (1-\eta)/\theta$ and $(1-\eta)(\theta-1)/\theta = 1-\beta$), we get

$$\eta x_g^k + (1-\eta)x_f^k = \eta x_g^k + \frac{(1-\eta)}{\theta}x_g^k - \frac{(1-\eta)(1-\theta)}{\theta}x^k = \beta x_g^k + (1-\beta)x^k$$

Substituting into $\|\eta x_g^k + (1-\eta)x_f^k - x^*\|^2$, we get

$$\|\eta x_g^k + (1-\eta)x_f^k - x^*\|^2 = \|\beta x_g^k + (1-\beta)x^k - x^*\|^2$$
$$= \|x^k - x^* + \beta(x_g^k - x^k)\|^2$$
$$= \|x^k - x^*\|^2 + 2\beta\langle x^k - x^*, x_g^k - x^k\rangle + \beta^2\|x^k - x_g^k\|^2$$
$$= \|x^k - x^*\|^2 + \beta\left(\|x_g^k - x^*\|^2 - \|x^k - x^*\|^2 - \|x_g^k - x^k\|^2\right) + \beta^2\|x^k - x_g^k\|^2$$
$$= (1-\beta)\|x^k - x^*\|^2 + \beta\|x_g^k - x^*\|^2 + (\beta^2 - \beta)\|x^k - x_g^k\|^2. \qquad (12)$$

Combining (12) with (11), we finish the proof. $\qquad\square$

Lemma 5. *Let Assumptions 1 and 2 be satisfied. Let problem (3) be solved by Algorithm 1. Then for any $u \in \mathbb{R}^d$, we get*

$$\mathbb{E}_k[f(x_f^{k+1})] \leq f(u) - \langle\nabla f(x_g^k), u - x_g^k\rangle - \frac{\mu}{2}\|u - x_g^k\|^2 - \frac{\gamma}{2}\|\nabla f(x_g^k)\|^2$$
$$+ \frac{\gamma}{2}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}_k[\|g^k\|^2].$$

Proof. Using Assumption 1 and line 7 of Algorithm 1, we get

$$f(x_f^{k+1}) \leq f(x_g^k) + \langle\nabla f(x_g^k), x_f^{k+1} - x_g^k\rangle + \frac{L}{2}\|x_f^{k+1} - x_g^k\|^2$$

$$= f(x_g^k) - \gamma\langle\nabla f(x_g^k), g^k\rangle + \frac{L\gamma^2}{2}\|g^k\|^2$$
$$= f(x_g^k) - \gamma\langle\nabla f(x_g^k), \nabla f(x_g^k)\rangle - \gamma\langle\nabla f(x_g^k), \mathbb{E}_k[g^k] - \nabla f(x_g^k)\rangle$$
$$- \gamma\langle\nabla f(x_g^k), g^k - \mathbb{E}_k[g^k]\rangle + \frac{L\gamma^2}{2}\|g^k\|^2$$
$$\leq f(x_g^k) - \gamma\|\nabla f(x_g^k)\|^2 + \frac{\gamma}{2}\|\nabla f(x_g^k)\|^2 + \frac{\gamma}{2}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2$$

$$- \gamma \langle \nabla f(x_g^k), g^k - \mathbb{E}_k[g^k] \rangle + \frac{L\gamma^2}{2} \|g^k\|^2.$$

Here we also used Cauchy-Schwartz inequality (1) with $a = \nabla f(x_g^k)$, $b = \nabla f(x_g^k) - \mathbb{E}_k[g^k]$ and $c = 1$. Taking the conditional expectation, we get

$$\mathbb{E}_k[f(x_f^{k+1})] \le f(x_g^k) - \frac{\gamma}{2} \|\nabla f(x_g^k)\|^2 + \frac{\gamma}{2} \|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + \frac{L\gamma^2}{2} \mathbb{E}_k[\|g^k\|^2].$$

Using Assumption 2 with $x = u$ and $y = x_g^k$, one can conclude that for any $u \in \mathbb{R}^d$ it holds

$$\mathbb{E}_k[f(x_f^{k+1})] \le f(u) - \langle \nabla f(x_g^k), u - x_g^k \rangle - \frac{\mu}{2} \|u - x_g^k\|^2 - \frac{\gamma}{2} \|\nabla f(x_g^k)\|^2$$

$$+ \frac{\gamma}{2} \|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + \frac{L\gamma^2}{2} \mathbb{E}_k[\|g^k\|^2]. \qquad \square$$

Taking into account all of the considerations above, we can prove the following result:

Theorem 1. *Consider Assumptions 1 – 4. Let the problem (3) be solved by Algorithm 1. Then for $\beta, \theta, \eta, \gamma, M$ satisfying*

$$M = (1 + 2/\beta), \quad \beta = \sqrt{\frac{4\mu\gamma}{9}}, \quad \eta = \frac{9\beta}{2\mu\gamma} = \sqrt{\frac{9}{\mu\gamma}},$$

$$\gamma \lesssim \min \left\{ \frac{\mu^3}{\delta^4 \tau^2}; \frac{1}{L} \right\}, \quad \theta = \frac{1 - \eta}{\beta - \eta},$$

it holds that

$$\mathbb{E}\left[\|x^N - x^*\|^2 + \frac{18}{\mu}(f(x_f^N) - f(x^*)) \right]$$

$$\lesssim \exp\left(-N\sqrt{\frac{\mu\gamma}{9}} \right) \left[\|x^0 - x^*\|^2 + \frac{18}{\mu}(f(x^0) - f(x^*)) \right] + \frac{\sqrt{\gamma}}{\mu^{3/2}} C_1 \tau \log_2 M \sigma^2.$$

Proof. Using Lemma 5 with $u = x^*$ and $u = x_f^k$, we get

$$\mathbb{E}_k[f(x_f^{k+1})] \le f(x^*) - \langle \nabla f(x_g^k), x^* - x_g^k \rangle - \frac{\mu}{2} \|x^* - x_g^k\|^2 - \frac{\gamma}{2} \|\nabla f(x_g^k)\|^2$$

$$+ \frac{\gamma}{2} \|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + \frac{L\gamma^2}{2} \mathbb{E}_k[\|g^k\|^2],$$

$$\mathbb{E}_k[f(x_f^{k+1})] \le f(x_f^k) - \langle \nabla f(x_g^k), x_f^k - x_g^k \rangle - \frac{\mu}{2} \|x_f^k - x_g^k\|^2 - \frac{\gamma}{2} \|\nabla f(x_g^k)\|^2$$

$$+ \frac{\gamma}{2} \|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + \frac{L\gamma^2}{2} \mathbb{E}_k[\|g^k\|^2].$$

Summing the first inequality with coefficient $2\gamma\eta$, the second with coefficient $2\gamma\eta(\eta - 1)$ and (10), we obtain

$$
\begin{aligned}
&\mathbb{E}_k[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2 f(x_f^{k+1})] \\
&\leq (1 + \alpha\gamma\eta)(1 - \beta)\|x^k - x^*\|^2 + (1 + \alpha\gamma\eta)\beta\|x_g^k - x^*\|^2 \\
&\quad + (1 + \alpha\gamma\eta)(\beta^2 - \beta)\|x^k - x_g^k\|^2 - 2\eta\gamma\langle\nabla f(x_g^k), \eta x_g^k + (1 - \eta)x_f^k - x^*\rangle \\
&\quad + \eta^2\gamma^2\mathbb{E}_k[\|g^k\|^2] + \frac{\eta\gamma}{\alpha}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 \\
&\quad + 2\gamma\eta\Big(f(x^*) - \langle\nabla f(x_g^k), x^* - x_g^k\rangle - \frac{\mu}{2}\|x^* - x_g^k\|^2 - \frac{\gamma}{2}\|\nabla f(x_g^k)\|^2 \\
&\quad\quad + \frac{\gamma}{2}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}_k[\|g^k\|^2]\Big) \\
&\quad + 2\gamma\eta(\eta - 1)\Big(f(x_f^k) - \langle\nabla f(x_g^k), x_f^k - x_g^k\rangle - \frac{\mu}{2}\|x_f^k - x_g^k\|^2 - \frac{\gamma}{2}\|\nabla f(x_g^k)\|^2 \\
&\quad\quad + \frac{\gamma}{2}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}_k[\|g^k\|^2]\Big) \\
&= (1 + \alpha\gamma\eta)(1 - \beta)\|x^k - x^*\|^2 + 2\gamma\eta(\eta - 1)(f(x_f^k) - 2\gamma\eta f(x^*)) \\
&\quad + ((1 + \alpha\gamma\eta)\beta - \gamma\eta\mu)\|x_g^k - x^*\|^2 \\
&\quad + (1 + \alpha\gamma\eta)(\beta^2 - \beta)\|x^k - x_g^k\|^2 - \gamma^2\eta^2\|\nabla f(x_g^k)\|^2 \\
&\quad + \left(\frac{\eta\gamma}{\alpha} + \gamma^2\eta^2\right)\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + \left(\eta^2\gamma^2 + \gamma^3\eta^2 L\right)\mathbb{E}_k[\|g^k\|^2] \\
&\leq (1 + \alpha\gamma\eta)(1 - \beta)\|x^k - x^*\|^2 + 2\gamma\eta(\eta - 1)(f(x_f^k) - 2\gamma\eta f(x^*)) \\
&\quad + ((1 + \alpha\gamma\eta)\beta - \gamma\eta\mu)\|x_g^k - x^*\|^2 \\
&\quad + (1 + \alpha\gamma\eta)(\beta^2 - \beta)\|x^k - x_g^k\|^2 - \gamma^2\eta^2\|\nabla f(x_g^k)\|^2 \\
&\quad + \eta\gamma\left(\frac{1}{\alpha} + \gamma\eta\right)\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + 8\eta^2\gamma^2(1 + \gamma L)\mathbb{E}_k[\|g^k - \nabla f(x_g^k)\|^2] \\
&\quad + \frac{1}{2}\eta^2\gamma^2(1 + \gamma L)\mathbb{E}_k[\|\nabla f(x_g^k)\|^2].
\end{aligned}
$$

In the last step, we also used (2) with $c = 4$. Since $\gamma \leq \frac{9}{16L}$, the choice of $\alpha = \frac{\beta}{2\eta\gamma}$, $\beta = \sqrt{16\mu\gamma/9}$ gives

$$
\beta = \sqrt{16\mu\gamma/9} \leq \sqrt{\mu/L} \leq 1,
$$
$$
(1 + \alpha\eta\gamma)(1 - \beta) = \left(1 + \frac{\beta}{2}\right)(1 - \beta) \leq \left(1 - \frac{\beta}{2}\right),
$$

and, therefore,

$$
\begin{aligned}
&\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2 f(x_f^{k+1})\right] \\
&\quad\quad \leq (1 - \beta/2)\|x^k - x^*\|^2 + 2\gamma\eta(\eta - 1)(f(x_f^k) - 2\gamma\eta f(x^*)) \\
&\quad\quad\quad + \eta^2\gamma^2(1 + 2/\beta)\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 \\
&\quad\quad\quad + 8\eta^2\gamma^2(1 + \gamma L)\mathbb{E}_k[\|g^k - \nabla f(x_g^k)\|^2]
\end{aligned}
$$

$$+ ((1 + \alpha\gamma\eta)\beta - \gamma\eta\mu) \|x_g^k - x^*\|^2$$

Subtracting $2\gamma\eta^2 f(x^*)$ from both sides, we get

$$\mathbb{E}_k \left[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{k+1}) - f(x^*)) \right]$$
$$\leq (1 - \beta/2) \|x^k - x^*\|^2 + (1 - 1/\eta) \cdot 2\gamma\eta^2(f(x_f^k) - f(x^*))$$
$$+ \eta^2\gamma^2 (1 + 2/\beta) \|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2$$
$$+ 8\eta^2\gamma^2 (1 + \gamma L) \mathbb{E}_k[\|g^k - \nabla f(x_g^k)\|^2]$$
$$+ ((1 + \alpha\gamma\eta)\beta - \gamma\eta\mu) \|x_g^k - x^*\|^2$$

Applying Lemma 3 and $\gamma L \leq 1$, one can obtain

$$\mathbb{E}_k \left[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{k+1}) - f(x^*)) \right]$$
$$\leq (1 - \beta/2) \|x^k - x^*\|^2 + (1 - 1/\eta) \cdot 2\gamma\eta^2(f(x_f^k) - f(x^*))$$
$$+ \eta^2\gamma^2 (1 + 2/\beta) \cdot 2C_1\tau M^{-1}(\sigma^2 + \delta^2\|x_g^k - x^*\|^2)$$
$$+ 16\eta^2\gamma^2 \cdot 13C_1\tau \log_2 M(\sigma^2 + \delta^2\|x_g^k - x^*\|^2)$$
$$+ ((1 + \alpha\gamma\eta)\beta - \gamma\eta\mu) \|x_g^k - x^*\|^2$$

With $M \geq (1 + 2/\beta)$, $\sqrt{\gamma} \leq \frac{\mu^{\frac{3}{2}}}{1872C_1\tau\delta^2 \log_2 M}$, $\alpha = \frac{\beta}{2\gamma\eta}$, $\beta = \frac{2}{3}\sqrt{\mu\gamma}$ and $\eta = \sqrt{\frac{9}{\mu\gamma}}$, we have:

$$(1 + \alpha\gamma\eta)\beta - \gamma\eta\mu + \eta^2\gamma^2\delta^2\big((1 + 2/\beta) \cdot 2C_1\tau M^{-1} + 208C_1\tau \log_2 M\big)$$
$$\leq (1 + \alpha\gamma\eta)\beta - 3\sqrt{\mu\gamma} + \frac{C_1\tau\delta^2\gamma}{\mu}\big(18 + 1872 \log_2 M\big)$$
$$\leq \sqrt{\mu\gamma} - 3\sqrt{\mu\gamma} + 2\sqrt{\mu\gamma} \leq 0,$$

and then,

$$\mathbb{E}_k \left[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{k+1}) - f(x^*)) \right]$$
$$\leq \big(1 - \beta/2\big)\|x^k - x^*\|^2 + (1 - 1/\eta) \cdot 2\gamma\eta^2(f(x_f^k) - f(x^*))$$
$$+ \Big(\eta^2\gamma^2 (1 + 2/\beta) \cdot 2C_1\tau M^{-1} + 16\eta^2\gamma^2 \cdot 13C_1\tau \log_2 M\Big)\sigma^2$$
$$\leq \max\left\{(1 - \beta/2), (1 - 1/\eta)\right\} \left[\|x^k - x^*\|^2 + 2\gamma\eta^2(f(x_f^k) - f(x^*)) \right]$$
$$+ \Big(\eta^2\gamma^2 (1 + 2/\beta) \cdot 2C_1\tau M^{-1} + 16\eta^2\gamma^2 \cdot 13C_1\tau \log_2 M\Big)\sigma^2.$$

Using that $\eta\gamma = 9\beta/(2\mu)$, $\beta/2 = 1/\eta$ and $\gamma \leq L^{-1}$, we have

$$\mathbb{E}_k \left[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{k+1}) - f(x^*)) \right]$$
$$\leq (1 - \beta/2) \left[\|x^k - x^*\|^2 + 2\gamma\eta^2(f(x_f^k) - f(x^*)) \right]$$
$$+ \frac{81}{4}\beta^2\mu^{-2}\Big((1 + 2/\beta) \cdot 2C_1\tau M^{-1} + 208C_1\tau \log_2 M\Big)\sigma^2$$

Finally, we perform the recursion and substitute $\beta = \sqrt{4\mu\gamma/9}$:

$$\mathbb{E}\big[\|x^N - x^*\|^2 + 2\gamma\eta^2(f(x_f^N) - f(x^*))\big]$$

$$\leq \left(1 - \sqrt{\frac{\mu\gamma}{9}}\right)^N [\|x^0 - x^*\|^2 + 2\gamma\eta^2(f(x_f^0) - f(x^*))]$$

$$+ \frac{81}{2}\beta\mu^{-2}\Big((1 + 2/\beta) \cdot 2C_1\tau M^{-1} + 208C_1\tau\log_2 M\Big)\sigma^2$$

$$\leq \exp\left(-\sqrt{\frac{\mu\gamma N^2}{9}}\right)[\|x^0 - x^*\|^2 + 2\gamma\eta^2(f(x_f^0) - f(x^*))]$$

$$+ \frac{81\sqrt{\gamma}}{\mu^{3/2}}C_1\tau\Big(1 + 104\log_2 M\Big)\sigma^2 \,.$$

Substituting of $\eta = \sqrt{\frac{9}{\mu\gamma}}$ concludes the proof. $\square$

Corollary 1 (Step tuning for Theorem 1). *Under the conditions of Theorem 1, choosing γ as*

$$\gamma \lesssim \min\left\{\frac{\mu^3}{\delta^4\tau^2}; \frac{1}{L}; \frac{1}{\mu N^2}\ln^2\left(\frac{\mu^2 N[\|x^0 - x^*\|^2 + 18\mu^{-1}(f(x_f^0) - f(x^*))]}{\tau\sigma^2}\right)\right\},$$

in order to achieve ϵ-approximate solution (in terms of $\mathbb{E}\big[\|x^N - x^*\|^2\big] \lesssim \epsilon$) it takes

$$\tilde{\mathcal{O}}\left(\left(\sqrt{\frac{L}{\mu}} + \frac{\tau\delta^2}{\mu^2}\right)\log\left(\frac{1}{\epsilon}\right) + \frac{\tau\sigma^2}{\mu^2\epsilon}\right) \quad \textit{oracle calls.}$$

4 Numerical experiments

In this section, we present numerical experiments that compare the proposed method and the existing approaches for the problem of finding consensus in distributed network.

4.1 Problem formulation

Let us consider the next problem. Assume that we have $\{x_i\}_{i=1}^d$, where $x_i \in \mathbb{R}$. Also we get a communication network, where i^{th} agent stores x_i. Moreover, the communication graph can be described as $G_k = (V, E_k)$, where the set of edges depends on the k – the current moment. The task is formulated as a consensus search, i.e., to find $\overline{x} = \frac{1}{d}\sum_{i=1}^d x_i$ – the average value of the agents.
To formalize our problem, we introduce the Laplacian matrix of the graph G_k: $W_k = D_k - A_k$ (here D_k is the diagonal matrix with degrees of nodes, A_k – adjacency matrix) and its properties:

1. $[W_k]_{i,j} \neq 0$ if and only if $(i, j) \in E_k$ or $i = j$,
2. $\ker W_k \supset \big\{(x_1, \ldots, x_d) \in \mathbb{R}^d : x_1 = \ldots = x_d\big\}$,

3. range $W_k \subset \left\{ (x_1, \ldots, x_d) \in \mathbb{R}^d : \sum_{i=1}^d x_i = 0 \right\}$.

If we consider $x = (x_1, \ldots, x_d)^\top$, then, because of second property, one can obtain

$$x_1 = \ldots = x_n \Leftrightarrow W_k x = 0.$$

Moreover, it is known that

$$W_k x = 0 \Leftrightarrow \sqrt{W_k} x = 0.$$

Hence, the problem of finding the consensus on the moment k can be reformulated as

$$\min_{x \in \mathbb{R}} \left[f(x) := \frac{1}{2} x^\top W_k x \right]. \tag{13}$$

It is important that the problem formulations (13) for each k have the same optimal point x^*, which is equal to consensus.

The classic approaches to find a consensus is a gossip protocol [36]. In terms of problem (13) the method can be formulated as a gradient descent:

$$x^{k+1} = x^k - \gamma W_k x^k = (1 - \gamma W_k) x^k.$$

This iteration sequence gives the consensus since the third property is fulfilled – it allows to keep the sum of coordinates of x^k the same, preventing the departure from the desired optimal point.

As mentioned above, the problem changes over time as the set of edges specifying the communication system changes. This situation occurs quite often in practice – when additional resources are available to improve the network, edges may be added to speed up processes, and in some system failures, communications between agents may be disconnected due to crashes and overloads. Therefore, it is natural to assume that the changes in the graphs G_k occur according to the Markovian law, since the changes are confined only to the current state of the communication system.

Since for the problem (13) the gradient is equal to $W_k x$, we have

$$\|W_k x - \mathbb{E}(W_k) x\|^2 = \|W_k x - W_k x^* - \mathbb{E}(W_k) x + \mathbb{E}(W_k) x^*\|^2$$
$$\leq \lambda_{max}^2 (W_k - \mathbb{E}(W_k)) \|x - x^*\|^2,$$

where $\mathbb{E}(W_k)$ is an expectation of Laplacian matrix of a graph G_k taking into account the stochasticity responsible for the changes in the graph (more detailed description see later). Consequently, the considered problem satisfies Assumption 4, what means that the theoretical analysis of our paper is applicable to (13).

4.2   Setup

In numerical experiments, we consider the problem described above on different topologies with certain Markovian stochastisity.

Brief description. We design the experiments in the following way. Suppose we have some starting, or base topology. Then we modify it according to some Markovian law, during which we cannot affect the base graph (i.e., discard edges from it). Based on these changes, we compare two methods: proposed and classic one.

Topologies. As a base topologies we consider two types of graphs – cycle-graph and star-graph. For each starting network we conducted numerical experiments for problems with different dimensions: 10, 100, 1000.

Markovian stochasticity. The network changes in time in the certain way. On each moment k with probability $\frac{1}{2}$ the random edge can be added to the topology, but if it already exists in the graph, then nothing happens. At the same time, with the same probability the random edge can be removed from the network. Nevertheless, if this edge is in the base topology or communication topology does not contain this edge, we keep the graph in the same condition.

4.3   Results

We performed numerical experiments with different base topologies (see Figures 1 and 2) with $d = 10$ (see Figures 1a, 2a), 100 (see Figures 1b, 2b) and 1000 (see Figures 1c, 2c). As a result, the proposed method outperform the classic approach [36] showing a faster rate of convergence, especially for the high-dimension problem.
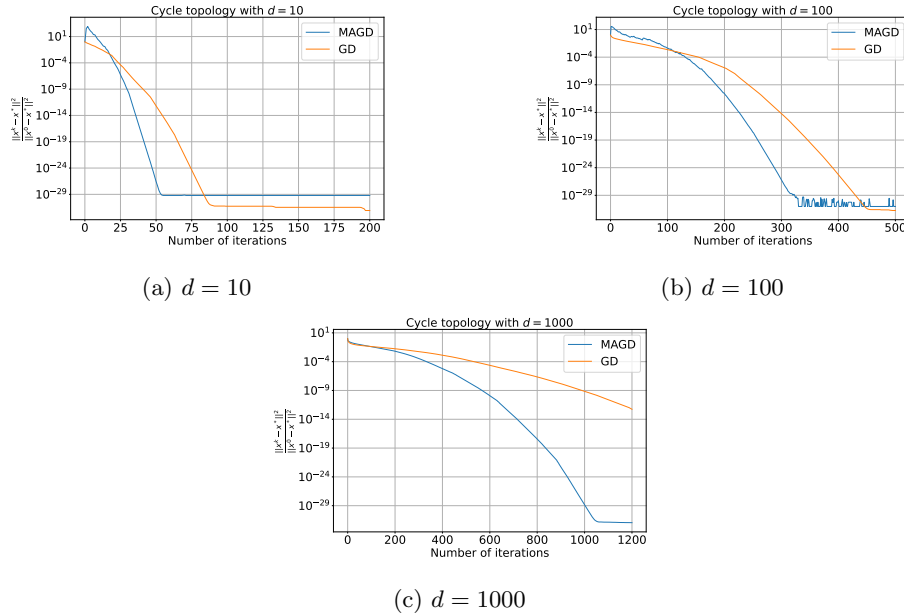

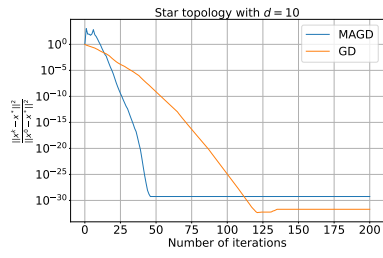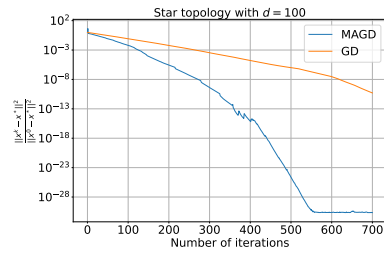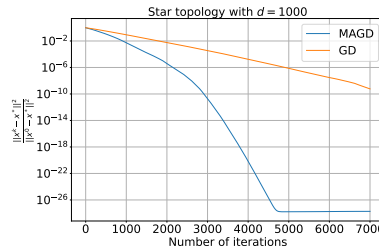(a) $d = 10$

(b) $d = 100$

(c) $d = 1000$

Fig. 1: Comparison of MAGD and GD for the consensus problem (13) on the cycle topology with different dimensions.

(a) $d = 10$

(b) $d = 100$

(c) $d = 1000$

Fig. 2: Comparison of MAGD and GD for the consensus problem (13) on the star topology with different dimensions.

References

1. A. Beznosikov, S. Samsonov, M. Sheshukova, A. Gasnikov, A. Naumov, E. Moulines. First order methods with markovian noise: from acceleration to variational inequalities. *Advances in Neural Information Processing Systems*, 36, 2022. https://doi.org/10.48550/arXiv.2305.15938

2. C. Hu, W. Pan, and J. Kwok. Accelerated gradient methods for stochastic optimization and online learning. *Advances in Neural Information Processing Systems*, 22, 2009.

3. A. Cotter, O. Shamir, N. Srebro, and K. Sridharan. Better mini-batch algorithms via accelerated gradient methods. *Advances in neural information processing systems*, 24, 2011.

4. O. Devolder et al. Stochastic first order methods in smooth convex optimization. Technical report, CORE, 2011.

5. G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

6. Q. Lin, X. Chen, and J. Pena. A smoothing stochastic gradient method for composite optimization. *Optimization Methods and Software*, 29(6):1281–1301, 2014.

7. P. Dvurechensky and A. Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171:121–145, 2016.

8. A. V. Gasnikov and Y. E. Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58: 48–64, 2018.

9. S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pages 1195–1204. PMLR, 2019a.

10. A. Taylor and F. Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Conference on Learning Theory*, pages 2934–2992. PMLR, 2019.

11. N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. *Advances in neural information processing systems*, 32, 2019.

12. E. Gorbunov, M. Danilova, I. Shibaev, P. Dvurechensky, and A. Gasnikov. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*, 2021.

13. B. E. Woodworth and N. Srebro. An even more optimal stochastic optimization algorithm: minibatching and interpolation learning. *Advances in Neural Information Processing Systems*, 34:7333–7345, 2021.

14. J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

15. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

16. I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

17. G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. 01 2020. ISBN 978-3-030-39567-4. https://doi.org/10.1007/978-3-030-39568-1.

18. R. Dorfman and K. Y. Levy. Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR, 2022.

19. T. Sun, Y. Sun, and W. Yin. On Markov chain gradient descent. *Advances in neural information processing systems*, 31, 2018.

20. A. Beznosikov, V. Samokhin, and A. Gasnikov. Distributed saddle-point problems: Lower bounds, optimal and robust algorithms. *arXiv preprint arXiv:2010.13112*, 2020.

21. T. T. Doan. Finite-time analysis of markov gradient descent. *IEEE Transactions on Automatic Control*, 68(4):2140–2153, 2023. https://doi.org/10.1109/TAC.2022.3172593.

22. T. T. Doan, L. M. Nguyen, N. H. Pham, and J. Romberg. Convergence rates of accelerated markov gradient descent with applications in reinforcement learning. *arXiv preprint arXiv:2002.02873*, 2020.

23. C. Liu and M. Belkin. Accelerating sgd with momentum for over-parameterized learning. *arXiv preprint arXiv:1810.13395*, 2018.

24. T. Sun, D. Li, and B. Wang. Adaptive Random Walk Gradient Descent for Decentralized Optimization. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20790–20809. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/sun22b.html`.

25. M. Even. Stochastic gradient descent under Markovian sampling schemes. *arXiv preprint arXiv:2302.14428*, 2023.

26. P. Wang, Y. Lei, Y. Ying, and D.-X. Zhou. Stability and generalization for markov chain stochastic gradient methods. *arXiv preprint arXiv:2209.08005*, 2022.

27. D. Nagaraj, X. Wu, G. Bresler, P. Jain, and P. Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676, 2020.

28. Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Advances in Neural Information Processing Systems*, 33:16223–16234, 2020.

29. E. Gorbunov, H. Berard, G. Gidel, and N. Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR, 2022.

30. A. N. Iusem, A. Jofré, R. I. Oliveira, and P. Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017. https://doi.org/10.1137/15M1031953. URL `https://doi.org/10.1137/15M1031953`.

31. R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, 2018. ISBN 978-3-319-97703-4.

32. E. Gorbunov, M. Danilova, A. Gasnikov Stochastic Optimization with Heavy-Tailed Noise via Accelerated Gradient Clipping *Advances in Neural Information Processing Systems*, 34, 2020. https://doi.org/10.48550/arXiv.2005.10785

33. Herbert Robbins. Sutton Monro. A Stochastic Approximation Method. Ann. Math. Statist. 22 (3) 400 - 407, September, 1951. URL https://doi.org/10.1214/aoms/1177729586

34. Arkadi S Nemirovski and David Berkovich Yudin. Cesari convergence of the gradient method of approximating saddle points of convex-concave functions. Doklady Akademii Nauk, volume 239, pages 1056–1059. Russian Academy of Sciences, 1978.
35. Arkadi Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.
36. Bertsekas, Dimitri and Tsitsiklis, John. Parallel and distributed computation: numerical methods. 2015.
37. Y. E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR*, 269(3):543–547, 1983. ISSN 0002-3264.
38. Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012. https://doi.org/10.1137/100802001.
39. Lan, Guanghui. Efficient Methods for Stochastic Composite Optimization `https://api.semanticscholar.org/CorpusID:15780105`
40. M. Schmidt and N. Le Roux. Fast Convergence of Stochastic Gradient Descent under a Strong Growth Condition, 2013, `https://arxiv.org/pdf/1308.6370.pdf`
41. H. Wang, M. Gürbüzbalaban, L. Zhu, U. Şimşekli and M. A. Erdogdu. Convergence Rates of Stochastic Gradient Descent under Infinite Noise Variance, *Advances in Neural Information Processing Systems*, 34, 2021. `https://doi.org/10.48550/arXiv.2102.10346`
42. Qu, Guannan and Li, Na Accelerated Distributed Nesterov Gradient Descent, IEEE Transactions on Automatic Control, 2020: 2566–2581. `https://doi.org/10.1109/TAC.2019.2937496`
43. Devolder, Olivier. Stochastic first order methods in smooth convex optimization. CORE Discussion Paper ; 2011/70 (2011)