First Order Methods with Markovian Noise: from Acceleration to Variational Inequalities

Aleksandr Beznosikov
Innopolis University, Skoltech, MIPT, Yandex

Sergey Samsonov
HSE University

Marina Sheshukova
HSE University

Alexander Gasnikov
MIPT, Skoltech, IITP RAS

Alexey Naumov
HSE University

Eric Moulines
Ecole polytechnique

Abstract

This paper delves into stochastic optimization problems that involve Markovian noise. We present a unified approach for the theoretical analysis of first-order gradient methods for stochastic optimization and variational inequalities. Our approach covers scenarios for both non-convex and strongly convex minimization problems. To achieve an optimal (linear) dependence on the mixing time of the underlying noise sequence, we use the randomized batching scheme, which is based on the multilevel Monte Carlo method. Moreover, our technique allows us to eliminate the limiting assumptions of previous research on Markov noise, such as the need for a bounded domain and uniformly bounded stochastic gradients. Our extension to variational inequalities under Markovian noise is original. Additionally, we provide lower bounds that match the oracle complexity of our method in the case of strongly convex optimization problems.

1 Introduction

Stochastic gradient methods are an essential ingredient for solving various optimization problems, with a wide range of applications in various fields such as machine learning [36, 37], empirical risk minimization problems [96], and reinforcement learning [93, 85, 69]. Various stochastic gradient descent methods (SGD) and their accelerated versions [75, 31] have been extensively studied under different statistical frameworks [17, 97]. The standard assumption for stochastic optimization algorithms is to consider independent and identically distributed noise variables. However, the growing usage of stochastic optimization methods in reinforcement learning [10, 87, 25] and distributed optimization [63, 18, 65] has led to increased interest in problems with Markovian noise. Despite this, existing theoretical works that consider Markov noise have significant limitations, and their analysis often results in suboptimal finite-time error bounds.

Our research aims to fill the gap in the existing literature on the first-order Markovian setting. By focusing on uniformly geometrically ergodic Markov chains, we obtain finite-time complexity bounds for achieving ε-accurate solutions that scale linearly with the mixing time of the underlying Markov chain. Our approach is based on careful applications of randomized batch size schemes and provides a unified view on both non-convex and strongly convex minimization problems, as well as variational inequalities.

Our contributions. Our main contributions are the following:

$\diamond$ **Accelerated SGD.** We provide the first analysis of SGD, including the Nesterov accelerated SGD method, with Markov noise without the assumption of bounded domain and uniformly bounded stochastic gradient estimates. Our results are summarised in Table 1 and Section 2.1 and cover both strongly convex and non-convex scenarios. Our findings for non-convex minimization problems complement the results obtained in [21].

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

$\diamond$ **Lower bounds.** In Section 2.2 we give the lower bounds showing that the presence of mixing time in the upper complexity bounds is not an artefact of the proof. This is consistent with the results reported in [71].

$\diamond$ **Extensions.** In Section 2.4 we provide, as far as we know, the first analysis for variational inequalities with general stochastic Markov oracle, arbitrary optimization set, and arbitrary composite term. Our finite-time performance analysis provides complexity bounds in terms of oracle calls that scale linearly with the mixing time of the underlying chain, which is an improvement over the bounds obtained in [99] for the Markov setting.

Related works. Next, we briefly summarize the related works.

$\diamond$ **Stochastic gradient methods.** Numerous research papers have reported significant improvements achieved by accelerated methods for stochastic optimization with stochastic gradient oracles involving independent and identically distributed (i.i.d.) noise. These methods have been extensively studied in theory [44, 14, 16, 58, 61, 26, 30, 97, 94, 3, 39, 102] and have shown practical success [55, 91]. The finite-time analysis of first-order methods in i.i.d. noise settings has been extensively studied by many authors, as discussed in [59] and references therein. In Table 1 we include only some important results because i.i.d. setting is not in the interest of this paper.

While the literature on i.i.d. noise is extensive, existing research on the first-order Markovian setting is relatively sparse. In this study, we focus on Markov chains that are uniformly geometrically ergodic, and we refer the reader to Section 2 for detailed definitions. We note that the complexity bounds which scale linearly with the mixing time of the underlying general Markov chain are currently available only for general convex and non-convex minimization problems. Namely, [23] has investigated a version of the ergodic mirror descent algorithm that yields optimal convergence rates for Lipschitz, general convex and non-convex problems. Recently, [21] proposed a random batch size algorithm that adapts to the mixing time of the underlying chain for non-convex optimization with a compact domain. In particular, [21, Theorem 4.3] yields optimal complexity rates in terms of the number of oracle calls required for non-convex problems, which is consistent with the results obtained in [23]. Unlike previous studies, this method is insensitive to the mixing time of the noise sequence.

For the general case of Markovian noise the finite-time analysis of non-accelerated SGD-type algorithms was carried out in [90] and [19]. However, [90] heavily relies on the bounded domain assumption and uniformly bounded stochastic gradient oracles, while its bound in [90, Theorem 5] has a suboptimal dependence on the mixing time of the underlying chain, see Table 1. Additionally, [90] does not cover the strongly convex setting. On the other hand, [19] covers both non-convex and strongly convex settings, but the bounds of [19, Theorem 1] has terms that are *exponential* in the mixing time, and a careful examination reveals suboptimal dependence on the initial condition for strongly convex problems when SGD is applied.

In the study of Nesterov-accelerated SGD with Markovian noise, the authors of [20] considered the use of a batch size of 1 and achieved a rate of forgetting the initial condition that matches that of the i.i.d. noise setting. However, their result is suboptimal in terms of the variance terms in both non-convex and strongly convex settings, as detailed in Table 1. We emphasize that the case of unbounded gradient oracles with Markov noise is not treated in contrast to the i.i.d. setup [97, 62].

The above papers deal with general Markovian noise optimization. But there are also results that deal with Markovian stochasticity with a finite state space. Here we can highlight the work [28], where the author gives quite extensive results and achieves linear scaling by mixing time in the non-convex as well as strongly convex cases. Recently, numerous papers have appeared dealing with the special scenario of distributed optimization [89]. [99] investigates the generalization and stability of Markov SGD with special attention to the excess variance guarantees. We note that first, these algorithms only need to deal with a very special case of Markov gradients, and second, the corresponding dependence on the mixing time of the Markov chain is again quadratic. At the same time, there exist particular results, e.g. [71], which provide a lower bound for the particular finite sum problems in the Markovian setting.

$\diamond$ **Variational inequalities.** Variational inequalities [29] have been an active area of research in applied mathematics for more than half a century [78, 41, 86]. VI cover important special cases, e.g., minimization over a convex domain, saddle point or min-max and fixed point problems. computational game theory [29], robust [7] and nonsmooth [73, 72] optimization, supervised [51, 4] and unsupervised [103, 5] learning, image denoising [27, 11]. In the last 5 years, variational inequalities and their special cases have attracted much interest in the machine learning community due to new connections to reinforcement learning [79, 50], adversarial training [64], and GANs [15, 33, 66, 12, 60, 82].

Variational inequalities (VI) and saddle point problems have their own well-established theory and methods. Unlike minimization problems, solving variational inequalities doesn't rely on (accelerated) gradient descent. Instead, the extragradient method [57], various modified versions [72, 42], or similar techniques [95] are recommended as the basic and theoretically optimal methods. While deterministic methods have long been used for solving variational inequalities, stochastic methods have gained importance only in the last 15 years, following pioneering works by [49, 52]. We summarise the results on methods for stochastic variational inequalities with the Lipschitz operator and smooth stochastic saddle point problems in Table 2. The number of papers dealing with stochastic VIs and saddle point problems is small compared to those dealing with stochastic optimization, we include in Table 2 papers with the i.i.d. noise (which we do not do for stochastic optimization). The only competing work dealing with Markovian noise in saddle point problems consider the finite sum problem and thus the finite Markov chain [99], therefore we do not include it in Table 2. Moreover, the results from [99] has much worse oracle complexity guarantees $\mathcal{O}(\tau^2/\varepsilon^2)$ in terms of τ. There are more papers dealing with stochastic finite-sum variational inequalities or saddle point problems, but in the i.i.d. setting [12, 80, 104, 2, 8]. We also do not consider in Table 2 because of the difference in the stochastic oracle structure. It is important to note that, unlike most previous works, we consider the most general formulation of VI itself for an arbitrary optimization set and composite term.

Notations and definitions. Let $(\mathsf{Z}, \mathsf{d}_{\mathsf{Z}})$ 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 $\mathbb{P}_\xi(Z_k \in A | Z_{k-1}) = \mathrm{Q}(Z_{k-1}, A)$, $\mathbb{P}_\xi$-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 $x^1, \ldots, x^k$ being the iterates of any stochastic first-order method, we denote $\mathcal{F}_k = \sigma(x^j, j \le k)$ and write $\mathbb{E}_k$ as an alias for $\mathbb{E}[\cdot|\mathcal{F}_k]$. We also write $\mathbb{N}^* := \mathbb{N} \setminus \{0\}$. For the sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ we write $a_n \lesssim b_n$ if there exists a constant c such that that $a_n \le c b_n$ for all $n \in \mathbb{N}$.

Table 1: This table summarizes our results on first-order method with Markovian noise. The columns of the table indicate whether the authors consider optimization over bounded domain, potentially unbounded gradients, and whether or not they assume additional restrictions on the Markovian noise (finite state space or reversibility). For ease of comparison we provide the respective results on SGD and ASGD (accelerated SGD) in the i.i.d. setting.

		Unbounded					
	Method	**Domain**	**Gradient noise**	**General MC**	**Acceleration**	**Oracle complexity (Smooth and non-convex)**	**Oracle complexity (Smooth and strongly convex)**
i.i.d.	SGD [84, 70, 32]	✓	✗	N/A	✗	$\tilde{\mathcal{O}}\left(L(f(x^0)-f(x^*))\left[\frac{1}{\varepsilon^2}+\frac{\sigma^2}{\varepsilon^4}\right]\right)$	$\tilde{\mathcal{O}}\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right)$
i.i.d.	ASGD [97, 13] [(1)]	✓	✓	N/A	✓	$\tilde{\mathcal{O}}\left(L(f(x^0)-f(x^*))\left[\frac{1+\delta^2}{\varepsilon^2}+\frac{\sigma^2}{\varepsilon^4}\right]\right)$	$\tilde{\mathcal{O}}\left((1+\delta^2)\sqrt{\frac{L}{\mu}}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right)$
Markovian	EMD [23][(2)]	✗	✗	✓	✗	$\tilde{\mathcal{O}}\left(\frac{\tau G^2 D^2}{\varepsilon^4}\right)$	✗
Markovian	MC SGD [90][(3)]	✓	✗	✗	✗	$\tilde{\mathcal{O}}\left(h(G,L)\left(\frac{\tau}{\varepsilon^2}\right)^{1/(1-q)}\right)$	✗
Markovian	MC SGD [19][(4)]	✓	✓	✓	✗	$\tilde{\mathcal{O}}\left(\frac{\tau L^2(1+\|x^*\|^2+\|x^0-x^*\|^2)}{\varepsilon^4}\right)$	$\tilde{\mathcal{O}}\left(e^{\tau(L/\mu)^2}\left[h(\frac{L}{\mu})\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\tau^2L^2(1+\|x^*\|^2)}{\mu^2\varepsilon}\right]\right)$
Markovian	ASGD[20][(5)]	✗	✗	✗	✓	$\tilde{\mathcal{O}}\left(\frac{1}{\varepsilon^4}\left[B^2+G^6(L^2\tau^2+1)\right]\right)$	$\tilde{\mathcal{O}}\left(\sqrt{\frac{L}{\mu}}\frac{\|x^0-x^*\|^2}{\varepsilon^{1/2}}+\frac{\tau^2(G^2+\mu GD+\mu LD^2)}{\mu^2\varepsilon}\right)$
Markovian	MAG [21][(6)]	✓	✗	✓	✗	$\tilde{\mathcal{O}}\left(\frac{\tau(G+L+B)^2G^2}{\varepsilon^4}\right)$	✗
Markovian	MC SGD [28] (Sec. 5.1) [(7)]	✓	✗	✗	✗	$\mathcal{O}\left(\frac{\tau(L(f(x^0)-f(x^*))+\sigma^2)}{\varepsilon^2}+\frac{\tau(L(f(x^0)-f(x^*))+\sigma^2)\sigma^2}{\varepsilon^4}\right)$	$\mathcal{O}\left(\frac{\tau L}{\mu}\log\frac{(f(x^0)-f(x^*))/\mu+\sigma^2/(\mu L)}{\varepsilon}+\frac{\tau\sigma^2}{\mu^2\varepsilon}\right)$
Markovian	MC SGD [28] (Sec. 5.2) [(8)]	✓	✗	✗	✗	✗	$O\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{L\tau\sigma_*^2}{\mu^3\varepsilon}\right)$
Markovian	RASGD (ours)	✓	✓	✓	✓	$\tilde{\mathcal{O}}\left(\tau L(f(x^0)-f(x^*))\left[\frac{1+\delta^2}{\varepsilon^2}+\frac{\sigma^2}{\varepsilon^4}\right]\right)$	$\tilde{\mathcal{O}}\left(\tau\left[(1+\delta^2)\sqrt{\frac{L}{\mu}}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right]\right)$

notation: μ and L are as in A 1 and A 2,$G = \sup_{x,z}\|\nabla F(x,z)\|$. Note that $G \ge L$ and $G^2 \ge \sigma^2$ under A 4. We also set $B = \sup_x |f(x)|$; x^0 - starting point, x^* - solution, $\mathcal{D}$ - optimisation domain; $D = \sup_{x\in\mathcal{D}}\|x - x^*\|$, σ and δ - stochasticity parameters (see A 4); σ_* - stochasticity parameter in x^* ; τ - mixing time of the chain (see A 3), ε - accuracy of the solution, measured as $\mathbb{E}[\|\nabla f(x)\|^2] \lesssim \varepsilon^2$ for non-convex problems and $\mathbb{E}[\|x - x^*\|^2] \lesssim \varepsilon$ for the strongly convex ones. Functions $h(L/\mu)$ and $h(G,L)$ stands for an implicit dependence of the respective parameters.

[(1)] gives results with stepsize as a parameter, we choose it the close way as in our Corollary 1. [(2)] covers more general noise setting, then just Markovian. [(3)] for general state-space Markov noise the analysis of [90] requires reversibility. Parameter $q \in (1/2; 1)$ refers to the step size $\sim 1/k^q$. [(4)] The fluctuation terms in [19, Theorem 1,3] contain hidden dependence on the initial error and $\|x^*\|$ in the fluctuation terms, making the result comparison complicated. They also contain hidden factors, which are exponential in $C = \tau/\log 4$ in the notations of our paper. Moreover, the analysis of [19] requires that $F(x,z)$ is Lipschitz w.r.t. x for any $z \in \mathsf{Z}$. [(5)] considers Markovian noise with finite state space and a specifically decreasing step size. Moreover, in the proof of [20, Theorem 3] (equations $(64)-(66)$) the authors lost the factor C^2, with $C = \tau/\log 4$. The result in the table accounts for this lost factor. [(6)] considers the adaptive tuning of batch size, which is oblivious to τ. [(7)] considers Markov noise with finite state space and additionally assumes that all stochastic realization $F(\cdot, Z)$ are L-smooth. [(8)] considers Markovian noise with finite state space, σ_* bounds noise only in x^*, but additionally assumes that all stochastic realization $F(\cdot, Z)$ are L-smooth and μ-strongly convex.

2 Main results

Assumptions. In this paper we study the minimization problem

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{Z\sim\pi}[F(x,Z)]\,, \tag{1}$$

where the access to the function f and its gradient is available only through the (unbiased) noisy oracle $F(x,Z)$ and $\nabla F(x,Z)$, respectively. In the following presentation we impose at least one of the following regularity constraint on the underlying function f itself:

Table 2: This table summarizes the findings on methods for solving stochastic (strongly) monotone variational inequalities with a Lipschitz operator and (un)bounded stochasticity. The columns of the table indicate whether the authors consider variational inequalities or only certain saddle point problems, the arbitrariness of the sets, and the use of additional composite terms. The columns on stochasticity provide information on the assumptions made with respect to the stochastic operator, such as bounded variance and the Markovian noise setting. Note that with the exception of our work, all other studies assume the independent noise.

		Statement			Stochasticity		
	Method	**VI?**	**Any set?**	**Composite?**	**Unbounded?**	**Markovian?**	**Oracle complexity**
Strongly monotone	SPEG [33, 42]	✓	✓	✗	✗	✗	$\tilde{\mathcal{O}}\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right)$ (1)
	SEG [53]	✓	✓	✗	✗	✗	$\tilde{\mathcal{O}}\left(\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{B^2+\sigma^2+(B+\sigma)(1+LD)}{\sigma^2\varepsilon}\right)$
	SS-SEG [68, 38]	✓	✓	✓	✓	✗	$\tilde{\mathcal{O}}\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma_*^2}{\mu^2\varepsilon}\right)$
	SEG [9]	✗	✓	✗	✗	✗	$\tilde{\mathcal{O}}\left(\frac{L}{\mu}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right)$
	DSEG [43]	✓	✗	✗	✓	✗	$\mathcal{O}\left(\left[\frac{L^2\sigma^2}{\mu^4\varepsilon}\right]^3\right)$ (2)
	UEG [38]	✓	✗	✗	✓	✗	$\mathcal{O}\left(\left(\frac{L+\Delta}{\mu}+\frac{\Delta^2}{\mu^2}\right)\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right)$
	SGDA [8]	✓	✓	✓	✗	✗	$\mathcal{O}\left(\frac{L^2}{\mu^2}\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right)$ (3)
	REG (ours)	✓	✓	✓	✓	✓	$\tilde{\mathcal{O}}\left(\tau\cdot\left[\left(\frac{L+\Delta}{\mu}+\frac{\Delta^2}{\mu^2}\right)\log\frac{\|x^0-x^*\|^2}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\right]\right)$
Monotone	SMP [52]	✓	✓	✗	✗	✗	$\mathcal{O}\left(\frac{LD^2}{\varepsilon}+\frac{\sigma^2\Delta^2}{\varepsilon^2}\right)$
	VR-SEG [45]	✓	✓	✗	✓	✗	$\mathcal{O}\left(\frac{(\sigma+\Delta)^8D^4}{\varepsilon^2}+\frac{D^4}{\varepsilon^2}\right)$
	IPM [46]	✓	✓	✓	✓	✗	$\mathcal{O}\left(\tilde{\mathcal{O}}\left(\frac{L^4D^4}{\varepsilon^2}+\frac{\sigma_*^2D^4}{\varepsilon^2}\right)\right)$
	SS-SEG [68]	✓	✓	✓	✓	✗	$\tilde{\mathcal{O}}\left(\frac{L^2D^4}{\varepsilon}+\frac{\sigma_*^4}{L^2\varepsilon^2}\right)$
	SEG [9]	✗	✓	✗	✗	✗	$\mathcal{O}\left(\frac{LD^2}{\varepsilon}+\frac{\sigma^2\Delta^2}{\varepsilon^2}\right)$
	REG (ours)	✓	✓	✓	✓	✓	$\tilde{\mathcal{O}}\left(\tau\cdot\left[\frac{LD^2}{\varepsilon}+\frac{\sigma^2D^2}{\varepsilon^2}+\frac{\Delta^2D^4}{\varepsilon^2}\right]\right)$

notation: μ = constant of strong monotonicity of operator F, L = Lipschitz constant of F, B = uniform bound of F, D = uniform bound of iterations x^k, x^0 = starting point, x^* = solution, Δ and σ = stochasticity parameters (see A 7, [52, 33, 42, 53, 9, 8] take $\Delta = 0$), σ_* = stochasticity parameter in x^* (see [68]), τ = mixing time of the chain (see A 3), ε = accuracy of the solution. (1) give results with stepsize as a parameter, we choose it according to Section 3 from [88]. (2) consider A 7, but do not provide explicit rates if $\Delta \neq 0$ (see also [38, Table 1]). (3) consider the cocoercive case, for which in general $\ell = L^2/\mu$.

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

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

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

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

Next we specify our assumptions on the sequence of noise variables $\{Z_i\}_{i=0}^{\infty}$. We consider here the general setting of $\{Z_i\}_{i=0}^{\infty}$ being a time-homogeneous Markov chain. Such problems naturally arise in stochastic optimization. In the empirical risk minimization problems it naturally appears in the context of non-random minibatch choice. Indeed, a random choice of a batch number may lose to a non-random one, see [67, 56]. A wide range of problems dealing with Markovian noise is spawned by the reinforcement learning methods. The usual MDP setting falls naturally inside this paradigm, moreover, the analysis of non-tabular RL problems requires to deal with the general state-space Markov noise. Here the potential range of applications include the policy evaluation methods, such as the temporal difference methods [92], and policy optimization algorithms, such as policy gradient family, e.g. the celebrated REINFORCE algorithm [100].

We denote by Q the Markov kernel corresponding to the sequence $\{Z_i\}_{i=0}^{\infty}$ and impose the following assumption on the mixing properties of Q:

A 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 $k \in \mathbb{N}$,*

$$\Delta(\mathrm{Q}^k) = \sup_{z,z' \in \mathsf{Z}} (1/2)\|\mathrm{Q}^k(z,\cdot) - \mathrm{Q}^k(z',\cdot)\|_{\mathsf{TV}} \leq (1/4)^{\lfloor k/\tau \rfloor} . \tag{3}$$

The assumption A 3 is classical in the literature on optimization methods with Markovian noise and has been considered in particular in recent works [90, 21, 20]. In particular, this assumption covers finite state-space Markov chains with irreducible and aperiodic transition matrix considered in

[28]. Yet our definition of the mixing time τ is more classical in the probability literature [81], and is slightly different from the one considered e.g. in [28, 65]. Next we specify our assumptions on stochastic gradient:

A 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*

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

The assumption A 4 resembles the strong growth condition [97], which is classical for the over-parametrized learning setup [97, 98]. The main difference is that A 4 concerns the almost sure bound in (4), which is unavoidable when dealing with uniformly geometrically ergodic Markovian noise A 3. Note that it is possible that the quantity δ^2 in (4) is not instance-independent and scales with the ratio L/μ from A 1-A 2 in the particular problems. With the assumptions A 3 and A 4 we can prove the result on the mean squred error of the stochastic gradient estimate computed over batch size n under arbitrary initial distribution. This result is summarized below in Lemma 1:

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

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

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

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

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

Proof. We first prove (5). Note that due to [81, Proposition 3.4] the Markov kernel Q under A 3 admits a positive pseudospectral gap $\gamma_{ps} > 0$ such that $1/\gamma_{ps} \leq 2\tau$. Thus, applying the statement of [81, Theorem 3.2], we get under A 4 that

$$\mathbb{E}_\pi[\|n^{-1} \textstyle\sum_{i=1}^n \nabla F(x, Z_i) - \nabla f(x)\|^2] \leq \frac{4\mathbb{E}_\pi[\|\nabla F(x, Z_1) - \nabla f(x)\|^2]}{n\gamma_{ps}} \leq \frac{8\tau}{n} \left(\sigma^2 + \delta^2 \|\nabla f(x)\|^2\right) .$$

To prove the second part we use the maximal exact coupling construction and follow, e.g., [24, Theorem 1]. The complete proof is given in Appendix B.1. $\square$

The proof of Lemma 1 simplifies the arguments in [21, Lemma 4] and allows us to obtain tighter values for the constants when dealing with the randomized batch size. Note that it is especially important to have the result for MSE under arbitrary initial distribution ξ, since in the proofs of our main results we will inevitably deal with the conditional expectations w.r.t. the previous iterate. We provide more details on the bias and variance of the Markov SGD gradients in the next section.

2.1 Accelerated method

We begin with a version of Nesterov accelerated SGD with randomized batch size, described in Algorithm 1. Due to the unboundedness of the stochastic gradient variance (see A 4), using of the classical Nesterov accelerated method [76, Section 2.2.] does not give the desired result, it is necessary to introduce an additional momentum [74, 97]. We use our own version, but partially similar to [74, 97]. The main feature of Algorithm 1 is that the number of samples used during the k-th gradient computation scales as 2^{J_k}, where J_k comes from a truncated geometric distribution. The truncation parameter needs to be adopted (see Theorem 1) in order to control the computational complexity of the algorithm.

Randomized batch size allows for efficient *bias* reduction in the stochastic gradient estimates and can be seen as a particular case of the so called multilevel MCMC [35, 34]. In the optimization context this approach was successfully used by [21] for the non-convex problems. Indeed, this bias naturally appears under the Markovian stochastic gradients oracles. It is easy to see that, with the counter T^k defined in Line 9, we have

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

Below we show how the bias of the gradient estimate scales with the truncation parameter M. The statement of Lemma 2 yields that the gradient estimates g_k introduced above have the bias, which decreases *quadratically* with M.

Lemma 2. *Assume A 3 and A 4. Then for the gradient estimates g^k from Algorithm 1 it holds that $\mathbb{E}_k[g^k] = \mathbb{E}_k[g^k_{\lfloor \log_2 M \rfloor}]$. Moreover,*

$$\mathbb{E}_k[\|\nabla f(x^k) - g^k\|^2] \lesssim \left(\tau B^{-1} \log_2 M + \tau^2 B^{-2}\right)(\sigma^2 + \delta^2 \|\nabla f(x^k)\|^2) \,,$$
$$\|\nabla f(x^k) - \mathbb{E}_k[g^k]\|^2 \lesssim \tau^2 M^{-2} B^{-2} (\sigma^2 + \delta^2 \|\nabla f(x^k)\|^2) \,.$$

Algorithm 1 `Randomized Accelerated GD`

Parameters: stepsize $\gamma > 0$, momentums θ, η, β, p, number of iterations N, batchsize limit M
Initialization: choose $x^0 = x_f^0$
for $k = 0, 1, 2, \ldots, N-1$ **do**
  $x_g^k = \theta x_f^k + (1-\theta)x^k$
  Sample $J_k \sim \text{Geom}\,(1/2)$
  $g^k = g_0^k + \begin{cases} 2^{J_k}\left(g_{J_k}^k - g_{J_k-1}^k\right), & \text{if } 2^{J_k} \leq M \\ 0, & \text{otherwise} \end{cases}$ with $g_j^k = 2^{-j}B^{-1}\sum_{i=1}^{2^j B} \nabla f(x_g^k, Z_{T^k+i})$
  $x_f^{k+1} = x_g^k - p\gamma g^k$
  $x^{k+1} = \eta x_f^{k+1} + (p-\eta)x_f^k + (1-p)(1-\beta)x^k + (1-p)\beta x_g^k$
  $T^{k+1} = T^k + 2^{J_k}B$
end for

The proof and the statement with explicit constants are given in Appendix B.2. Note that the Lemma 2 is a natural counterpart of the deterministic bound Lemma 1. Moreover, it gives the idea of the trade-off between the parameters B and M. Namely, the expected number of oracle calls to compute g_k is $\mathcal{O}(B\log_2(M))$ with the bias scaling as M^{-2}. Thus the increase of M drastically reduced the bias with only a logarithmic payment in variance. At the same time, gradient variance scales as $(\tau/B)^2$, but the increase of B is much more expensive for the computational cost of the whole procedure. Taking into account the considerations above, we can prove the following result:

Theorem 1. *Assume A 1 – A 4. Let problem* (1) *be solved by Algorithm 1. Then for any $b \in \mathbb{N}^*$, $\gamma \in (0; \frac{3}{4L}]$, and $\beta, \theta, \eta, p, M, B$ satisfying*

$$p \simeq (1 + (1+\gamma L)[\delta^2\tau b^{-1} + \delta^2\tau^2 b^{-2}])^{-1}, \quad \beta \simeq \sqrt{p^2\mu\gamma}, \quad \eta \simeq \sqrt{\tfrac{1}{\mu\gamma}},$$
$$\theta \simeq \tfrac{p\eta^{-1}-1}{\beta p\eta^{-1}-1}, \quad M \simeq \max\{2; \sqrt{p^{-1}(1+p/\beta)}\}, \quad B = \lceil b\log_2 M\rceil\,,$$

it holds that

$$\mathbb{E}\left[\|x^N - x^*\|^2 + \tfrac{6}{\mu}(f(x_f^N) - f(x^*))\right] \lesssim \exp\bigl(-N\sqrt{\tfrac{p^2\mu\gamma}{3}}\bigr)\left[\|x^0 - x^*\|^2 + \tfrac{6}{\mu}(f(x^0) - f(x^*))\right] + \tfrac{p\sqrt{\gamma}}{\mu^{3/2}}\left(\sigma^2\tau b^{-1} + \sigma^2\tau^2 b^{-2}\right). \quad (7)$$

The proof is provided in Appendix B.3. The result of Theorem 1 can be rewritten as an upper complexity bound under an appropriate choice of the remaining free parameter b:

Corollary 1. *Under the conditions of Theorem 1, choosing $b = \tau$ and γ as*

$$\gamma \simeq \min\left\{\tfrac{1}{L}; \tfrac{1}{p^2\mu N^2}\ln\left(\max\left\{2; \tfrac{\mu^2 N[\|x^0-x^*\|^2 + 6\mu^{-1}(f(x_f^0)-f(x^*))]}{\sigma^2}\right\}\right)\right\},$$

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

$$\tilde{\mathcal{O}}\left(\tau\left[(1+\delta^2)\sqrt{\tfrac{L}{\mu}}\log\tfrac{1}{\varepsilon} + \tfrac{\sigma^2}{\mu^2\varepsilon}\right]\right) \quad \textit{oracle calls}\,. \quad (8)$$

The results of Corollary 1 are obtained with fixed parameters of the method. In Corollary 1 these parameters are selected a bit artificially, e.g., the stepsize γ depends on the iteration horizon N. In Appendix B.4 we show how one can similar results, but with a decreasing stepsize.

Comparison. Running the procedure above requires to know the mixing time τ. Estimating the mixing time from a single trajectory of the running Markov chain is known to be computationally hard problem, see e.g. [101] and references therein. At the same time, methods, which share the same (optimal) linear scaling of the sample complexity w.r.t. the mixing time also share the same drawback as our method. In particular, it holds true for the EMD algorithm [23], SGD-DD algorithm [71], and usual SGD with Markovian data [28]. At the same time, in the non-convex scenario the paper [21] is truly oblivious to mixing time, allowing to obtain sample complexity rates for non–convex problems, which are homogeneous w.r.t. τ with AdaGrad-type learning rate. An interesting direction for the future work to suggest a procedure that would allow to generalize the results of [21] to accelerated SGD setting.

It is possible that the sample complexity bound (8) is worse than the respective bounds for non-accelerated SGD with Markov data, provided that δ^2 grows quickly with L/μ. At the same time, this drawback is shared by the classical results on learning under the strong growth condition, see e.g. [97]. As it is shown in [62], the respective rates can be worse than the ones obtained by usual SGD even under the i.i.d. noise setting, see Appendix F.3 in [62]. Making the analysis of accelerated SGD 'backward compatible' w.r.t. the rates of usual SGD requires to perform analysis in terms of additional problem-specific quantities, see [47, 62].

The closest equivalent of the result Corollary 1 is given by [20, Theorem 3]. However, the corresponding bound of [20, Theorem 3] is incomplete, since the factor τ^2 is lost in the proof (see equations $(64-66)$). With this completion, the bound of [20, Theorem 3] yields a variance term of order $\tilde{\mathcal{O}}\left(\frac{\sigma^2\tau^2}{\mu^2\varepsilon}\right)$, which is suboptimal with respect to τ. Moreover, the corresponding analysis relies heavily on the assumption of a bounded domain. In [28], the author considers Markovian noise with a finite number of states and manages to obtain a rather interesting result of the form $O\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}+\frac{L\tau\sigma_*^2}{\mu^3\varepsilon}\right)$. Here the first term does not depend on τ, and the second consists only σ^* (stochasticity in x^*), but the price for this is an additional factor L/μ in the second term and more strict assumption that all realizations $F(\cdot,z)$ are smooth and strongly convex. In the context of overparameterized learning, our results are almost consistent with the bound of [97, Theorem 1] under i.i.d. sampling. The difference is that the term δ^2 in A 4 can be more pessimistic than the expectation bound in [97].

2.2 Lower bounds

We start with a lower bound for the complexity of Markovian stochastic optimization under the assumptions A 1 –A 4. Below we provide a result that highlights that the bound of Theorem 1 is tight provided that δ does not scale with the instance-dependent quantities, e.g., condition number L/μ.

Theorem 2. *There exists an instance of the optimization problem satisfying assumptions A 1 –A 4 with $\delta=1$ and arbitrary $\sigma\geq 0, L,\mu>0,\tau\in\mathbb{N}^*$, such that for any first-order gradient method it takes at least*

$$N=\Omega\left(\tau\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}+\frac{\tau\sigma^2}{\mu^2\varepsilon}\right)$$

oracle calls in order to achieve $\mathbb{E}[\|x^N-x^*\|^2]\leq\varepsilon$.

The proof is provided in Appendix B.5. The idea of the constructed lower for deterministic part bound $\Omega\big(\tau\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\big)$ goes back to [76, Theorem 2.1.13]. The stochastic part lower bound goes back to the classical statistical reasoning, and is well explained for i.i.d. noise in [59, Chapter 4.1]. Our adaptation for Markovian setting is based on Le Cam's theory, see [1, Theorem 8], and also [105]. For the case of Markov noise this lower bound is, to the best of our knowledge, original. The closest result to ours is the stochastic term lower bound in [28, Proposition 1], but it is valid only for the vanilla stochastic gradient methods. Below we provide another lower bound showing that the dependence of the sample complexity Corollary 1 on δ is not an artefact of the proof.

Proposition 1. *There exists an instance of the optimization problem satisfying assumptions A 1 –A 4 with arbitrary $L,\mu>0,\tau\in\mathbb{N}^*$, $\delta=\frac{L}{\mu}$, and $\sigma=0$, such that for any first-order gradient method it takes at least*

$$N=\Omega\left(\tau\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$$

gradient calls in order to achieve $\mathbb{E}[\|x^N-x^*\|^2]\leq\varepsilon$.

This lower bound is adapted from the information-theoretic lower bound [71]. The detailed proof can be found in Appendix B.5. Recent studies [54, 71, 13] have revealed the impossibility of accelerating stochastic gradient descent (SGD) for online linear regression problems with specific noise structures. To address this issue, researchers have proposed various solutions, such as the MaSS algorithm [62] and the approach presented in [48]. However, these methods rely heavily on the particular structure of the online regression setup. Another question that naturally arises is whether one can get rid of the dependence on τ in the deterministic part of (8) if $\delta=0$. The following counterexample shows that this is not the case in general.

Proposition 2. *There exists an instance of the optimisation problem satisfying assumptions A 1 –A 4 with with arbitrary $L,\mu>0,\tau\in\mathbb{N}^*$, $\sigma=1,\delta=0$, such that for any first-order gradient method it takes at least*

$$N=\Omega\left(\left(\tau+\sqrt{\frac{L}{\mu}}\right)\log\frac{1}{\varepsilon}\right)$$

oracle calls in order to achieve $\mathbb{E}[\|x^N - x^*\|^2] \leq \varepsilon$.

The proof is provided in Appendix B.5.

2.3 Non-convex problems

Now we proceed with a randomized batch size version of the simple SGD algorithm. It is summarized in Algorithm 2 and can be shown to achieve optimal rates of convergence for smooth non-convex problems. For the case of non-convex problems with Markov noise similar analysis appeared in [21, Theorem 4].

Algorithm 2 `Randomized GD`

Parameters: stepsize $\gamma > 0$, number of iterations K, bound on batchsize B, mixing time τ;
Initialization: choose $x^0 \in \mathcal{X}$
for $k = 0, 1, 2, \ldots, N-1$ **do**
  Sample $J_k \sim$ Geom $\left(\frac{1}{2}\right)$
  $g^k = g_0^k + \begin{cases} 2^{J_k}\left(g_{J_k}^k - g_{J_k-1}^k\right), & \text{if } 2^{J_k} \leq M \\ 0, & \text{otherwise} \end{cases}$ with $g_j^k = 2^{-j}B^{-1}\sum_{i=1}^{2^j B} \nabla f(x^k, Z_{T^k+i})$
  $x^{k+1} = x^k - \gamma g^k$
  $T^{k+1} = T^k + 2^{J_k} B$
end for

By balancing the values of B and M with Lemma 2, we establish the following result:

Theorem 3. *Assume A 1, A 3, A 4. Let problem* (1) *be solved by Algorithm 2. Let f^* be a global (maybe not unique) minimum of f. Then for any $b \in \mathbb{N}^*$, and γ, M satisfying*

$$\gamma \lesssim (L[1 + \delta^2\tau b^{-1} + \delta^2\tau^2 b^{-2}])^{-1}, \quad M \simeq \max\{2; \sqrt{\gamma^{-1}L^{-1}}\}, \quad B = \lceil b \log_2 M \rceil,$$

it holds that

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=0}^{N-1}\|\nabla f(x^k)\|^2\right] \lesssim \frac{f(x^0) - f^*}{\gamma N} + L\gamma \cdot \left[\sigma^2\tau b^{-1} + \sigma^2\tau^2 b^{-2}\right].$$

The proof is provided in Appendix B.6. The next corollary immediately follows from the theorem.

Corollary 2. *Under the conditions of Theorem 3, if we choose $b = \tau$ and γ given by*

$$\gamma \simeq \min\left\{\frac{1}{L(1+\delta^2)}; \sqrt{\frac{f(x^0)-f^*}{LN\sigma^2}}\right\},$$

then to achieve ε-solution (in terms of $\mathbb{E}[\|\nabla f(x)\|^2] \lesssim \varepsilon^2$) we need

$$\tilde{\mathcal{O}}\left(\tau \cdot \left[\frac{(1+\delta^2)L(f(x^0)-f^*)}{\varepsilon^2} + \frac{L(f(x^0)-f^*)\sigma^2}{\varepsilon^4}\right]\right) \quad \textit{oracle calls.}$$

Comparison. The respective bound for the non-convex setting provided in [20, Theorem 1] yields the sample complexity of order $\tilde{\mathcal{O}}\left(\frac{\tau^2 L(f(x^0)-f(x^*))\sigma^2}{\varepsilon^4}\right)$. Also we can note the results of [28, Theorem 2] with the following estimate $O\left(\frac{\tau(L(f(x^0)-f(x^*))+\sigma^2)}{\varepsilon^2} + \frac{\tau(L(f(x^0)-f(x^*))+\sigma^2)\sigma^2}{\varepsilon^4}\right)$.

To achieve linear convergence rates in the non-convex setting we can use the Polyak-Lojasiewicz (PL) condition [83]. The respective result is provided in Appendix B.7.

2.4 Variational inequalities

In this section, we are interested in the following problem:

$$\text{Find } x^* \in \mathcal{X} \text{ such that } \langle F(x^*), x - x^*\rangle + r(x) - r(x^*) \geq 0 \text{ for all } x \in \mathcal{X}. \tag{9}$$

Here $F: \mathbb{R}^d \to \mathbb{R}^d$ an operator, $\mathcal{X}$ a convex set, and $r: \mathbb{R}^d \to \mathbb{R}$ is a regularization term (a suitable lower semicontinuous convex function) which is assumed to have a simple structure. As mentioned earlier, this problem is quite general and covers a wide range of possible problem formulations. For example, if the operator F is the gradient of a convex function f, then the problem (9) is equivalent to the composite minimization problem [6], i.e., minimization of $f(x) + r(x)$. In the meantime, (9) is also a reformulation of the min-max problem

$$\min_{x_1 \in \mathcal{X}_1} \max_{x_2 \in \mathcal{X}_2} r_1(x_1) + g(x_1, x_2) - r_2(x_2), \tag{10}$$

with convex-concave continuously differentiable g, convex sets $\mathcal{X}_1$, $\mathcal{X}_2$ and convex functions r_1, r_2. Using the first-order optimality conditions, it is easy to verify that (10) is equivalent to (9) with $x = (x_1^T, x_2^T)^T$, $F(x) = (\nabla_{x_1} f(x_1, x_2)^T, -\nabla_{x_2} f(x_1, x_2)^T)^T$, and $r(x) = r_1(x_1) + r_2(x_2)$.

A 5. *The operator F is L-Lipschitz continuous on $\mathcal{X}$ with $L > 0$, i.e., the following inequality holds for all $x, y \in \mathcal{X}$:*

$$\|F(x) - F(y)\| \leq L\|x - y\|, \qquad \forall x, y \in \mathcal{X}.$$

A 6. *The operator F is μ_F-strongly monotone on $\mathcal{X}$, i.e., the following inequality holds for all $x, y \in \mathcal{X}$:*

$$\langle F(x) - F(y), x - y\rangle \geq \mu_F \|x - y\|^2. \tag{11}$$

The function r is μ_r-strongly convex on $\mathcal{X}$, i.e. for all $x, y \in \mathcal{X}$ and any $r'(x) \in \partial r(x)$ we have

$$r(y) \geq r(x) + \langle r'(x), y - x\rangle + (\mu_r/2)\|x - y\|^2. \tag{12}$$

These two assumptions are more than standard for the study of variational inequalities and are found in all the papers from Table 2. We consider two cases: strongly monotone/convex with $\mu_F + \mu_r > 0$ and monotone/convex with $\mu_F + \mu_r = 0$.

A 7. *For all $x \in \mathbb{R}^d$ it holds that $\mathbb{E}_\pi[F(x, Z)] = F(x)$. Moreover, for all $z \in \mathsf{Z}$ and $x \in \mathcal{X}$ it holds that*

$$\|F(x, z) - F(x)\|^2 \leq \sigma^2 + \Delta^2 \|x - x^*\|^2\,, \tag{13}$$

where x^* is some point from the solution set.

A 7 is found in the literature on variational inequalities [43, 45, 38] and is considered to be analog to A 4 on overparametrized learning.

Just as the Nesterov accelerated method is optimal for smooth convex minimization problems, the ExtraGradient method [57, 72, 52] is optimal for monotone variational inequalities. Therefore, we take it as a base. On the extrapolation step (Line 4) of Algorithm 3, we simply collect a batch of size B, but on the main step (Line 8) we use the randomization as in Algorithm 1. The next theorem gives the convergence of our method.

Theorem 4. *Assume A 5, A 6 with $\mu_F + \mu_r > 0$, A 3, A 7. Let problem* (9) *be solved by Algorithm 3. Then for any $b \in \mathbb{N}^*$, and γ, M satisfying*

$$\gamma \lesssim \min\left\{(\mu_F + \mu_r)^{-1}; L^{-1}; (\mu_F + \mu_r)(\Delta^2 \tau b^{-1} + \Delta^2 \tau^2 b^{-2})^{-1}; \sqrt{\Delta^{-2}\tau^{-1} b}\right\},$$
$$M \simeq \max\{2; \sqrt{\gamma^{-1}(\mu_F + \mu_r)^{-1}}\}, \quad B = \lceil b \log_2 M \rceil\,,$$

it holds that

$$\mathbb{E}\left[\|x^N - x^*\|^2\right] \lesssim \exp\left(-\tfrac{N(\mu_F + \mu_r)\gamma}{2}\right)\|x^0 - x^*\|^2 + \tfrac{\gamma}{\mu}(\sigma^2 \tau b^{-1} + \sigma^2 \tau^2 b^{-2})\,.$$

The proof is postponed to Appendix B.8. One can get an estimate on oracle complexity.

Algorithm 3 `Randomized ExtraGradient`

1: **Parameters:** stepsize $\gamma > 0$, number of iterations N
2: **Initialization:** choose $x^0 \in \mathcal{X}$
3: **for** $k = 0, 1, 2, \ldots, N-1$ **do**
4: $\quad x^{k+1/2} = \text{prox}_{\gamma r}\big(x^k - \gamma B^{-1} \sum_{i=1}^{B} F(x^k, Z_{T^k+i})\big)$
5: $\quad T^{k+1/2} = T^k + B$
6: $\quad$ Sample $J_k \sim \text{Geom}\big(\frac{1}{2}\big)$
7: $\quad g^k = g_0^k + \begin{cases} 2^{J_k}\left(g_{J_k}^k - g_{J_k-1}^k\right), & \text{if } 2^{J_k} \leq M \\ 0, & \text{otherwise} \end{cases}$ with $g_j^k = 2^{-j} B^{-1} \sum_{i=1}^{2^j \cdot B} F(x^{k+1/2}, Z_{T^{k+1/2}+i})$
8: $\quad x^{k+1} = \text{prox}_{\gamma r}\big(x^k - \gamma g^k\big)$
9: $\quad T^{k+1} = T^{k+1/2} + 2^{J_k} B$
10: **end for**

Corollary 3. *Under the conditions of Theorem 4, if we choose $b = \tau$ and γ as follows*

$$\gamma \simeq \min\left\{\tfrac{1}{\mu_F + \mu_r}; \tfrac{1}{L}; \tfrac{\mu_F + \mu_r}{\Delta^2}; \tfrac{1}{\Delta}; \tfrac{1}{N(\mu_F + \mu_r)} \ln\left(\max\left\{2; \tfrac{\mu N \|x^0 - x^*\|^2}{\sigma^2}\right\}\right)\right\},$$

then to achieve ε-solution (in terms of $\mathbb{E}[\|x - x^*\|^2] \lesssim \varepsilon$) we need

$$\tilde{\mathcal{O}}\left(\tau \cdot \left[\left(1 + \tfrac{L}{\mu_F + \mu_r} + \tfrac{\Delta}{\mu_F + \mu_r} + \tfrac{\Delta^2}{(\mu_F + \mu_r)^2}\right)\log\tfrac{1}{\varepsilon} + \tfrac{\sigma^2}{\mu^2 \varepsilon}\right]\right) \quad \textit{oracle calls.}$$

Note that one provide an (almost) matching lower complexity bounds for variational inequalities via lower bounds for saddle point problems, which are a special case of variational inequalities. The method for obtaining lower bounds for saddle point problems is reduced to obtaining estimates for the strongly convex minimization problem (see [106, 40] for respective deterministic lower bounds), which we provide in Section 2.2. Similarly, the question of constructing a lower bound which is tight w.r.t. Δ remains open.

For the monotone case, we use the *gap function* as a convergence criterion:

$$\mathrm{Gap}(x) = \sup_{y \in \mathcal{X}} \left[\langle F(y), x - y \rangle + r(x) - r(y) \right] . \tag{14}$$

Such a criterion is standard and classical for monotone variational inequalities [72, 52]. An important assumption for the gap function is the boundedness of the set $\mathcal{X}$.

A 8. *The set $\mathcal{X}$ is bounded and has a diameter D, i.e., for all $x, y \in \mathcal{X}$: $\|x - y\|^2 \leq D^2$.*

A 8 can be slightly relaxed. We need to use a simple trick from [77]. In particular, we need to consider $\mathcal{C}$ – a compact subset of $\mathcal{X}$ and change $\mathcal{X}$ to $\mathcal{C}$ in (14). But such a technique is rather technical and does not change the essence. Finally, the following result holds.

Theorem 5. *Assume A 5, A 6 with $\mu_F + \mu_r = 0$, A 8, A 3, A 7. Let problem* (9) *be solved by Algorithm 3. Then for any $B \in \mathbb{N}^*$, and γ, M satisfying $\gamma \lesssim L^{-1}$, $M = \sqrt{N}$, it holds that*

$$\mathbb{E}\left[\mathit{Gap}(\bar{x}^N)\right] \lesssim \tfrac{D^2}{\gamma N} + \gamma(\tau B^{-1} \log_2 N + \tau^2 B^{-2})(\sigma^2 + \Delta^2 D^2) \;\; \textit{where} \;\; \bar{x}^N = \tfrac{1}{N} \textstyle\sum_{k=0}^{N-1} x^{k+1/2} \, .$$

The proof is postponed to Appendix B.9. The following corollary holds.

Corollary 4. *Under the conditions of Theorem 5, if we choose $B = \tau$ and γ as follows*

$$\gamma \simeq \min\left\{ \tfrac{1}{L}; \sqrt{\tfrac{D^2}{(\sigma^2 + \Delta^2 D^2) N}} \right\},$$

then to achieve ε-solution (in terms of $\mathbb{E}[\mathit{Gap}(x)] \lesssim \varepsilon$) we need

$$\tilde{\mathcal{O}}\left(\tau \left[\tfrac{L D^2}{\varepsilon} + \tfrac{\sigma^2 D^2 + \Delta^2 D^4}{\varepsilon^2} \right] \right) \quad \textit{oracle calls.}$$

Comparison. These results is the first for variational inequalities with Markovian stochasticity, either in the strongly monotone or monotone cases. The only close work is [99]. The authors work with convex-concave saddle point problems and provide the following estimate on the oracle complexity $\mathcal{O}\left(\tau^2 \cdot \frac{G^4}{\varepsilon^2} + \frac{D^2}{\varepsilon^2}\right)$ (with G – the uniform bound of the operator), which is worse than ours at least in terms of τ. Moreover, the authors consider the case of a finite Markov chain, which is a special case of our setup.

3 Conclusion

In this paper, we present a unified random batch size framework that achieves optimal finite-time performance for non-convex and strongly convex optimization problems with Markov noise, as well as for variational inequalities. Unlike existing methods, our framework relaxes the assumptions typically imposed on the domain and stochastic gradient oracle. We also provide a variety of lower bounds, which are to the best of our knowledge original in the Markov setting.

Acknowledgments

This research of A. Beznosikov has been supported by The Analytical Center for the Government of the Russian Federation (Agreement No. 70-2021-00143 dd. 01.11.2021, IGK 000000D730321P5Q0002). E. Moulines received support from the grant ANR-19-CHIA-002 SCAI and parts of his work has been done under the auspices of Lagrange Center for maths and computing.

References

[1] Eddie Aamari and Clément Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. *The Annals of Statistics*, 47(1):177 – 204, 2019.

[2] Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pages 778–816. PMLR, 2022.

[3] Necdet Serhat Aybat, Alireza Fallah, Mert Gurbuzbalaban, and Asuman Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. *Advances in neural information processing systems*, 32, 2019.

[4] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *arXiv preprint arXiv:1108.0775*, 2011.

[5] Francis Bach, Julien Mairal, and Jean Ponce. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*, 2008.

[6] Amir Beck. *First-order methods in optimization*. Society for Industrial and Applied Mathematics (SIAM), 2017.

[7] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

[8] Aleksandr Beznosikov, Eduard Gorbunov, Hugo Berard, and Nicolas Loizou. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 172–235. PMLR, 2023.

[9] Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov. Distributed saddle-point problems: Lower bounds, optimal and robust algorithms. *arXiv preprint arXiv:2010.13112*, 2020.

[10] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.

[11] Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.

[12] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] You-Lin Chen, Sen Na, and Mladen Kolar. Convergence analysis of accelerated stochastic gradient descent under the growth condition. *arXiv preprint arXiv:2006.06782*, 2020.

[14] Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. Better mini-batch algorithms via accelerated gradient methods. *Advances in neural information processing systems*, 24, 2011.

[15] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[16] Olivier Devolder et al. Stochastic first order methods in smooth convex optimization. Technical report, CORE, 2011.

[17] Aymeric Dieuleveut, Nicolas Flammarion, and Francis Bach. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research*, 18(1):3520–3570, 2017.

[18] Alexandros G. Dimakis, Soummya Kar, José M. F. Moura, Michael G. Rabbat, and Anna Scaglione. Gossip algorithms for distributed signal processing. *Proceedings of the IEEE*, 98(11):1847–1864, 2010.

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

[20] Thinh T Doan, Lam M Nguyen, Nhan H Pham, and Justin Romberg. Convergence rates of accelerated markov gradient descent with applications in reinforcement learning. *arXiv preprint arXiv:2002.02873*, 2020.

[21] Ron Dorfman and Kfir Yehuda Levy. Adapting to mixing time in stochastic optimization with markovian data. In *International Conference on Machine Learning*, pages 5429–5446. PMLR, 2022.

[22] R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*. Springer Series in Operations Research and Financial Engineering. Springer, 2018.

[23] John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

[24] Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Marina Sheshukova. Rosenthal-type inequalities for linear statistics of markov chains. *arXiv preprint arXiv:2303.05838*, 2023.

[25] Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the stability of random matrix product with markovian noise: Application to linear stochastic approximation and td learning. In *Conference on Learning Theory*, pages 1711–1752. PMLR, 2021.

[26] Pavel Dvurechensky and Alexander Gasnikov. Stochastic intermediate gradient method for convex problems with stochastic inexact oracle. *Journal of Optimization Theory and Applications*, 171:121–145, 2016.

[27] Ernie Esser, Xiaoqun Zhang, and Tony F Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

[28] Mathieu Even. Stochastic gradient descent under Markovian sampling schemes. *arXiv preprint arXiv:2302.14428*, 2023.

[29] F. Facchinei and J.S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2007.

[30] Alexander Vladimirovich Gasnikov and Yu E Nesterov. Universal method for stochastic composite optimization problems. *Computational Mathematics and Mathematical Physics*, 58:48–64, 2018.

[31] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[32] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

[33] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

[34] Michael B. Giles. Multilevel monte carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

[35] Peter W Glynn and Chang-han Rhee. Exact estimation for markov chain equilibrium expectations. *Journal of Applied Probability*, 51(A):377–389, 2014.

[36] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[37] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. `http://www.deeplearningbook.org`.

[38] Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR, 2022.

[39] Eduard Gorbunov, Marina Danilova, Innokentiy Shibaev, Pavel Dvurechensky, and Alexander Gasnikov. Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. *arXiv preprint arXiv:2106.05958*, 2021.

[40] Yuze Han, Guangzeng Xie, and Zhihua Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *arXiv preprint arXiv:2103.08280*, 2021.

[41] P. T. Harker and J.-S. Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 1990.

[42] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[43] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. *Advances in Neural Information Processing Systems*, 33:16223–16234, 2020.

[44] Chonghai Hu, Weike Pan, and James Kwok. Accelerated gradient methods for stochastic optimization and online learning. *Advances in Neural Information Processing Systems*, 22, 2009.

[45] 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.

[46] Alfredo N Iusem, Alejandro Jofré, and Philip Thompson. Incremental constraint projection methods for monotone stochastic variational inequalities. *Mathematics of Operations Research*, 44(1):236–263, 2019.

[47] Prateek Jain, Sham M Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604. PMLR, 2018.

[48] Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018.

[49] Houyuan Jiang and Huifu Xu. Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Transactions on Automatic Control*, 53(6):1462–1475, 2008.

[50] Yujia Jin and Aaron Sidford. Efficiently solving MDPs with stochastic mirror descent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 4890–4900. PMLR, 2020.

[51] Thorsten Joachims. A support vector method for multivariate performance measures. pages 377–384, 01 2005.

[52] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[53] Aswin Kannan and Uday V Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *Computational Optimization and Applications*, 74(3):779–820, 2019.

[54] Rahul Kidambi, Praneeth Netrapalli, Prateek Jain, and Sham Kakade. On the insufficiency of existing momentum schemes for stochastic optimization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[55] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[56] Anastasia Koloskova, Nikita Doikov, Sebastian U Stich, and Martin Jaggi. Shuffle SGD is always better than SGD: Improved analysis of SGD with arbitrary data orders. *arXiv preprint arXiv:2305.19259*, 2023.

[57] G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:35–49, 1977.

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

[59] Guanghui Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. 01 2020.

[60] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 907–915. PMLR, 16–18 Apr 2019.

[61] Qihang Lin, Xi Chen, and Javier Pena. A smoothing stochastic gradient method for composite optimization. *Optimization Methods and Software*, 29(6):1281–1301, 2014.

[62] Chaoyue Liu and Mikhail Belkin. Accelerating sgd with momentum for over-parameterized learning. *arXiv preprint arXiv:1810.13395*, 2018.

[63] Cassio G. Lopes and Ali H. Sayed. Incremental adaptive strategies over distributed networks. *IEEE Transactions on Signal Processing*, 55(8):4064–4077, 2007.

[64] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[65] Xianghui Mao, Kun Yuan, Yubin Hu, Yuantao Gu, Ali H. Sayed, and Wotao Yin. Walkman: A communication-efficient random-walk algorithm for decentralized optimization. *IEEE Transactions on Signal Processing*, 68:2513–2528, 2020.

[66] Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

[67] Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33:17309–17320, 2020.

[68] Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

[69] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[70] Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[71] Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676, 2020.

[72] Arkadi Nemirovski. Prox-method with rate of convergence O(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[73] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

[74] Yu. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[75] Yu. 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.

[76] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[77] Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

[78] J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1944.

[79] Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P. How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pages 2681–2690. PMLR, 2017.

[80] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems*, 29, 2016.

[81] Daniel Paulin. Concentration inequalities for Markov chains by Marton couplings and spectral methods. *Electronic Journal of Probability*, 20(none):1 – 32, 2015.

[82] Wei Peng, Yu-Hong Dai, Hui Zhang, and Lizhi Cheng. Training gans with centripetal acceleration. *Optimization Methods and Software*, 35(5):955–973, 2020.

[83] B.T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[84] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[85] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[86] Gesualdo Scutari, Daniel Palomar, Francisco Facchinei, and Jong-shi Pang. Convex optimization, game theory, and variational inequality theory. *Signal Processing Magazine, IEEE*, 27:35 – 49, 06 2010.

[87] Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

[88] Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

[89] Tao Sun, Dongsheng Li, and Bao Wang. Adaptive Random Walk Gradient Descent for Decentralized Optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan 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.

[90] Tao Sun, Yuejiao Sun, and Wotao Yin. On Markov chain gradient descent. *Advances in neural information processing systems*, 31, 2018.

[91] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David 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.

[92] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988.

[93] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018.

[94] Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In *Conference on Learning Theory*, pages 2934–2992. PMLR, 2019.

[95] Paul Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

[96] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[97] Sharan Vaswani, Francis Bach, and Mark 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, 2019.

[98] Sharan Vaswani, Aaron Mishkin, Issam H. Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *CoRR*, abs/1905.09997, 2019.

[99] Puyu Wang, Yunwen Lei, Yiming Ying, and Ding-Xuan Zhou. Stability and generalization for markov chain stochastic gradient methods. *arXiv preprint arXiv:2209.08005*, 2022.

[100] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

[101] Geoffrey Wolfer and Aryeh Kontorovich. Estimating the mixing time of ergodic markov chains. In *Conference on Learning Theory*, pages 3120–3159. PMLR, 2019.

[102] Blake E Woodworth and Nathan Srebro. An even more optimal stochastic optimization algorithm: minibatching and interpolation learning. *Advances in Neural Information Processing Systems*, 34:7333–7345, 2021.

[103] Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2005.

[104] Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1153–1165. Curran Associates, Inc., 2020.

[105] Bin Yu. *Assouad, Fano, and Le Cam*, pages 423–435. Springer New York, New York, NY, 1997.

[106] Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the saddle point problems. *arXiv preprint arXiv:1912.07481*, 2019.

A Notations and definitions.

Let $(\mathsf{Z}, \mathsf{d}_\mathsf{Z})$ 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 $\mathbb{P}_\xi(Z_k \in A | Z_{k-1}) = \mathrm{Q}(Z_{k-1}, A)$, $\mathbb{P}_\xi$-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. We denote $\mathcal{F}_k = \sigma(x^j, j \leq k)$ and write $\mathbb{E}_k$ as an alias for $\mathbb{E}[\cdot|\mathcal{F}_k]$. For each function $f : \mathsf{Z} \mapsto \mathbb{R}$ with $\pi(f) < \infty$ we write $\bar{f}(z) = f(z) - \pi(f)$.

B Proofs of Section 2.1, Section 2.3

B.1 Proof of Lemma 1

Lemma 3 (Lemma 1). *Assume A 3 and A 4. Then, for any $n \geq 1$ and $x \in \mathbb{R}^d$, it holds that*

$$\mathbb{E}_\pi[\|n^{-1} \textstyle\sum_{i=1}^n \nabla F(x, Z_i) - \nabla f(x)\|^2] \leq \frac{8\tau}{n} \left(\sigma^2 + \delta^2 \|\nabla f(x)\|^2\right) . \tag{15}$$

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

$$\mathbb{E}_\xi[\|n^{-1} \textstyle\sum_{i=1}^n \nabla F(x, Z_i) - \nabla f(x)\|^2] \leq \frac{C_1\tau}{n} \left(\sigma^2 + \delta^2 \|\nabla f(x)\|^2\right) , \tag{16}$$

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

By [22, 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 $\mathbb{P}^{\mathrm{Q}}_\xi$ and $\mathbb{P}^{\mathrm{Q}}_{\xi'}$, that is,

$$\|\xi \mathrm{Q}^n - \xi' \mathrm{Q}^n\|_{\mathrm{TV}} = 2\tilde{\mathbb{P}}_{\xi,\xi'}(T > n) . \tag{17}$$

We write $\tilde{\mathbb{E}}_{\xi,\xi'}$ for the expectation with respect to $\tilde{\mathbb{P}}_{\xi,\xi'}$. Using the coupling construction (17),

$$\begin{aligned}\mathbb{E}_\xi^{1/2}[\| \textstyle\sum_{i=1}^n \{\nabla f(x, Z_i) - \nabla f(x)\}\|^2] \leq \mathbb{E}_\pi^{1/2}[\| \textstyle\sum_{i=0}^{n-1} \nabla f(x, Z_i) - \nabla f(x)\|^2] + \\ \tilde{\mathbb{E}}_{\xi,\pi}^{1/2}[\| \textstyle\sum_{i=0}^{n-1} \{\nabla f(x, Z_i) - \nabla f(x, Z'_i)\}\|^2] .\end{aligned}$$

The first term is bounded with (15). Moreover, with (17) and A 4, we get

$$\begin{aligned}\| \sum_{i=0}^{n-1} \{\nabla f(x, Z_i) - \nabla f(x, Z'_i)\}\|^2 &\leq 8 \left(\sigma^2 + \delta^2 \|\nabla f(x)\|^2\right) \Big(\sum_{i=0}^{n-1} \mathbb{1}_{\{Z_i \neq Z'_i\}}\Big)^2 \\ &= 8 \left(\sigma^2 + \delta^2 \|\nabla f(x)\|^2\right) \Big(\sum_{i=0}^{n-1} \mathbb{1}_{\{T > i\}}\Big)^2 \\ &\leq 16 \left(\sigma^2 + \delta^2 \|\nabla f(x)\|^2\right) \sum_{i=1}^{\infty} i\, \mathbb{1}_{\{T > i\}} .\end{aligned}$$

Thus, using the assumption A 3, we bound

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

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

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

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

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

and (16) holds. If $n > c_1\tau$, it holds that

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

and we also get (16).

B.2 Proof of Lemma 2

Before we proceed to the proof, we give a statement of Lemma 2 with exact constants.

Lemma 4 (Lemma 2)**.** *Assume A 3 and A 4. Then for the gradient estimates g^k from Algorithm 1 it holds that $\mathbb{E}_k[g^k] = \mathbb{E}_k[g^k_{\lfloor \log_2 M \rfloor}]$. Moreover,*

$$\begin{aligned}
&\mathbb{E}_k[\|\nabla f(x_g^k) - g^k\|^2] \leq \left(4C_1\tau B^{-1}\log_2 M + (4C_1+2)\tau^2 B^{-2}\right)(\sigma^2 + \delta^2\|\nabla f(x_g^k)\|^2) \,, \qquad (18)\\
&\|\nabla f(x_g^k) - \mathbb{E}_k[g^k]\|^2 \leq C_2\tau^2 M^{-2}B^{-2}(\sigma^2 + \delta^2\|\nabla f(x_g^k)\|^2) \,,
\end{aligned}$$

where C_1 is defined in (16) *and $C_2 = 256/3$.*

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:

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

We start with the proof of the first statement of (18) by taking the conditional expectation for J_k:

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

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 1 and get

$$\begin{aligned}
\mathbb{E}_k[\|\nabla f(x_g^k) - g^k\|^2] &\leq 2\sigma^2 + 4\sum_{i=1}^{\lfloor \log_2 M \rfloor} 2^i \left(\frac{C_1\tau}{2^i B}(\sigma^2 + \delta^2\|\nabla f(x_g^k)\|^2) + \frac{C_1\tau^2}{2^{2i}B^2}(\sigma^2 + \delta^2\|\nabla f(x_g^k)\|^2)\right) \\
&\leq \frac{4C_1(\sigma^2 + \delta^2\|\nabla f(x_g^k)\|^2)\tau\log_2 M}{B} + \frac{(4C_1+2)(\sigma^2 + \delta^2\|\nabla f(x_g^k)\|^2)\tau^2}{B^2} \,.
\end{aligned}$$

To show the second part of the statement, we use Lemma 2 and get

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

The remaining proof once again uses Lemma 1 and is omitted. To conclude we use that $2^{\lfloor \log_2 M \rfloor} \geq M/2$. ∎

B.3 Proof of Theorem 1.

We preface the proof by two technical Lemmas.

Lemma 5. *Assume A 1 and A 2. Then for the iterates of Algorithm 1 with* $\theta = (p\eta^{-1}-1)/(\beta p\eta^{-1}-1)$*,* $\theta > 0$*,* $\eta \geq 1$*,* $p > 0$*, it holds that*

$$\begin{aligned}\mathbb{E}_k[\|x^{k+1} - x^*\|^2] \leq &(1 + \alpha p\gamma\eta)(1-\beta)\|x^k - x^*\|^2 + (1 + \alpha p\gamma\eta)\beta\|x_g^k - x^*\|^2 \\ &+ (1 + \alpha p\gamma\eta)(\beta^2 - \beta)\|x^k - x_g^k\|^2 + p^2\eta^2\gamma^2\mathbb{E}_k[\|g^k\|^2] \\ &- 2\eta^2\gamma\langle \nabla f(x_g^k), x_g^k + (p\eta^{-1} - 1)x_f^k - \eta^{-1}px^*\rangle \\ &+ \frac{p\eta\gamma}{\alpha}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 \,, \end{aligned} \tag{19}$$

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

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

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

Using straightforward algebra, we get

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

In the last step we also applied Cauchy-Schwartz inequality in the form (43) with $\alpha > 0$. Taking the conditional expectation, we get

$$\begin{aligned}\mathbb{E}_k[\|x^{k+1} - x^*\|^2] \leq&(1 + \alpha p\eta\gamma)\|\eta x_g^k + (p-\eta)x_f^k + (1-p)(1-\beta)x^k + (1-p)\beta x_g^k - x^*\|^2 \\ &- 2p\gamma\eta\langle \nabla f(x_g^k), \eta x_g^k + (p-\eta)x_f^k + (1-p)(1-\beta)x^k + (1-p)\beta x_g^k - x^*\rangle \\ &+ p^2\gamma^2\eta^2\mathbb{E}_k[\|g^k\|^2] + \frac{p\gamma\eta}{\alpha}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 \,. \end{aligned} \tag{20}$$

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

$$\begin{aligned}&\eta x_g^k + (p-\eta)x_f^k + (1-p)(1-\beta)x^k + (1-p)\beta x_g^k \\ &\quad= (\eta + (1-p)\beta)x_g^k + (p-\eta)x_f^k + (1-p)(1-\beta)x^k \\ &\quad= (\eta + (1-p)\beta)x_g^k + \eta(p\eta^{-1}-1)x_f^k + (1-p)(1-\beta)x^k \\ &\quad= (\eta + (1-p)\beta)x_g^k + \eta(\beta p\eta^{-1}-1)\theta x_f^k + (1-p)(1-\beta)x^k \\ &\quad= (\eta + (1-p)\beta)x_g^k + \eta(\beta p\eta^{-1}-1)(x_g^k - (1-\theta)x^k) + (1-p)(1-\beta)x^k \\ &\quad= \beta x_g^k - \eta(\beta p\eta^{-1}-1)(1-\theta)x^k + (1-p)(1-\beta)x^k \\ &\quad= \beta x_g^k + p(1-\beta)x^k + (1-p)(1-\beta)x^k \end{aligned}$$

$$= \beta x_g^k + (1-\beta)x^k \, .$$

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

$$\begin{aligned}
&\|\eta x_g^k + (p-\eta)x_f^k + (1-p)(1-\beta)x^k + (1-p)\beta x_g^k - x^*\|^2 \\
&\qquad = \|\beta x_g^k + (1-\beta)x^k - x^*\|^2 \\
&\qquad = \|x^k - x^* + \beta(x_g^k - x^k)\|^2 \\
&\qquad = \|x^k - x^*\|^2 + 2\beta\langle x^k - x^*, x_g^k - x^k\rangle + \beta^2\|x^k - x_g^k\|^2 \\
&\qquad = \|x^k - x^*\|^2 + \beta\left(\|x_g^k - x^*\|^2 - \|x^k - x^*\|^2 - \|x_g^k - x^k\|^2\right) + \beta^2\|x^k - x_g^k\|^2 \\
&\qquad = (1-\beta)\|x^k - x^*\|^2 + \beta\|x_g^k - x^*\|^2 + (\beta^2 - \beta)\|x^k - x_g^k\|^2. \qquad (21)
\end{aligned}$$

Again with line 4 and the choice of θ such that $\theta = (p\eta^{-1} - 1)/(\beta p\eta^{-1} - 1)$ (in particular, $\eta^{-1}p(1-\beta) = (1-\beta p\eta^{-1})(1-\theta)$ and $(\beta p\eta^{-1} - 1)\theta = (p\eta^{-1} - 1)$), one can also note

$$\begin{aligned}
&\eta x_g^k + (p-\eta)x_f^k + (1-p)(1-\beta)x^k + (1-p)\beta x_g^k - x^* \\
&\quad = (\eta + (1-p)\beta)x_g^k + (p-\eta)x_f^k + (1-p)(1-\beta)x^k - x^* \\
&\quad = \eta p^{-1}\left((p + (1-p)\eta^{-1}p\beta)x_g^k + (p\eta^{-1} - 1)px_f^k + (1-p)(1-\beta)p\eta^{-1}x^k - \eta^{-1}px^*\right) \\
&\quad = \eta p^{-1}\left((p + (1-p)\eta^{-1}p\beta)x_g^k + (p\eta^{-1} - 1)px_f^k + (1-p)(1-\beta p\eta^{-1})(1-\theta)x^k - \eta^{-1}px^*\right) \\
&\quad = \eta p^{-1}\left((p + (1-p)\eta^{-1}p\beta)x_g^k + (p\eta^{-1} - 1)px_f^k + (1-p)(1-\beta p\eta^{-1})(x_g^k - \theta x_f^k) - \eta^{-1}px^*\right) \\
&\quad = \eta p^{-1}\left(x_g^k + (p\eta^{-1} - 1)px_f^k - (1-p)(1-\beta p\eta^{-1})\theta x_f^k - \eta^{-1}px^*\right) \\
&\quad = \eta p^{-1}\left(x_g^k + (p\eta^{-1} - 1)px_f^k + (1-p)(p\eta^{-1} - 1)x_f^k - \eta^{-1}px^*\right) \\
&\quad = \eta p^{-1}\left(x_g^k + (p\eta^{-1} - 1)x_f^k - \eta^{-1}px^*\right). \qquad (22)
\end{aligned}$$

Combining (21) and (22) with (20), we finish the proof. □

Lemma 6. *Assume A 1-A 2. Let problem* (1) *be solved by Algorithm 1. Then for any $u \in \mathbb{R}^d$, we get*

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

Proof. Using A 1 in the form (42) with $x = x_f^{k+1}$, $y = x_g^k$ and line 7 of Algorithm 1, we get

$$\begin{aligned}
f(x_f^{k+1}) \le\, & f(x_g^k) + \langle\nabla f(x_g^k), x_f^{k+1} - x_g^k\rangle + \frac{L}{2}\|x_f^{k+1} - x_g^k\|^2 \\
=\, & f(x_g^k) - p\gamma\langle\nabla f(x_g^k), g^k\rangle + \frac{Lp^2\gamma^2}{2}\|g^k\|^2 \\
=\, & f(x_g^k) - p\gamma\langle\nabla f(x_g^k), \nabla f(x_g^k)\rangle - p\gamma\langle\nabla f(x_g^k), \mathbb{E}_k[g^k] - \nabla f(x_g^k)\rangle \\
& - p\gamma\langle\nabla f(x_g^k), g^k - \mathbb{E}_k[g^k]\rangle + \frac{Lp^2\gamma^2}{2}\|g^k\|^2 \\
\le\, & f(x_g^k) - p\gamma\|\nabla f(x_g^k)\|^2 + \frac{p\gamma}{2}\|\nabla f(x_g^k)\|^2 + \frac{p\gamma}{2}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 \\
& - p\gamma\langle\nabla f(x_g^k), g^k - \mathbb{E}_k[g^k]\rangle + \frac{Lp^2\gamma^2}{2}\|g^k\|^2.
\end{aligned}$$

Here we also used Cauchy Schwartz inequality (43) 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

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

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

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

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

□

Theorem 6 (Theorem 1). *Assume A 1 – A 4. Let problem* (1) *be solved by Algorithm 1. Then for any* $b\in\mathbb{N}^*$, $\gamma\in(0;\frac{3}{4L}]$, *and* β,θ,η,p,M,B *satisfying*

$$p=\left[1+2\left(1+\gamma L\right)\left(1+4\left[C_1\tau b^{-1}+(C_1+1)\tau^2b^{-2}\right]\delta^2\right)\right]^{-1},$$
$$\beta=\sqrt{\tfrac{4p^2\mu\gamma}{3}},\quad \eta=\tfrac{3\beta}{2p\mu\gamma}=\sqrt{\tfrac{3}{\mu\gamma}},\quad \theta=\tfrac{p\eta^{-1}-1}{\beta p\eta^{-1}-1},$$
$$M=\max\{2;\sqrt{C_2p^{-1}(1+2p/\beta)}\},\quad B=\lceil b\log_2 M\rceil.$$

it holds that

$$\begin{aligned}
&\mathbb{E}\left[\|x^N-x^*\|^2+\frac{6}{\mu}(f(x_f^N)-f(x^*))\right]\\
&\quad\lesssim\exp\left(-N\sqrt{\frac{p^2\mu\gamma}{3}}\right)\left[\|x^0-x^*\|^2+\frac{6}{\mu}(f(x^0)-f(x^*))\right]+\frac{p\sqrt{\gamma}}{\mu^{3/2}}\left(\sigma^2\tau b^{-1}+\sigma^2\tau^2b^{-2}\right).
\end{aligned}$$

Proof. Using Lemma 6 with $u=x^*$ and $u=x_f^k$, we get

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

$$\begin{aligned}
\mathbb{E}_k[f(x_f^{k+1})]\le& f(x_f^k)-\langle\nabla f(x_g^k),x_f^k-x_g^k\rangle-\frac{\mu}{2}\|x_f^k-x_g^k\|^2-\frac{p\gamma}{2}\|\nabla f(x_g^k)\|^2\\
&+\frac{p\gamma}{2}\|\mathbb{E}_k[g^k]-\nabla f(x_g^k)\|^2+\frac{Lp^2\gamma^2}{2}\mathbb{E}_k[\|g^k\|^2].
\end{aligned}$$

Summing the first inequality with coefficient $2p\gamma\eta$, the second with coefficient $2\gamma\eta(\eta-p)$ and (19), we obtain

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

$$
\begin{aligned}
&+ (1+\alpha p\gamma\eta)(\beta^2-\beta)\|x^k - x_g^k\|^2 - p\gamma^2\eta^2\|\nabla f(x_g^k)\|^2 \\
&+ p\eta\gamma\left(\frac{1}{\alpha}+\gamma\eta\right)\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + 2p^2\eta^2\gamma^2\left(1+\gamma L\right)\mathbb{E}_k[\|g^k - \nabla f(x_g^k)\|^2] \\
&+ 2p^2\eta^2\gamma^2\left(1+\gamma L\right)\mathbb{E}_k[\|\nabla f(x_g^k)\|^2].
\end{aligned}
$$

In the last step we also used (44) with $c = 1$. Since $\gamma \le \frac{3}{4L}$, the choice of $\alpha = \frac{\beta}{2p\eta\gamma}$, $\beta = \sqrt{4p^2\mu\gamma/3}$, and $p\mu\gamma\eta = 3\beta/2$ gives

$$
\begin{aligned}
&\beta = \sqrt{4p^2\mu\gamma/3} \le \sqrt{p^2\mu/L} \le 1, \\
&(1+\alpha p\eta\gamma)(1-\beta) = \left(1+\frac{\beta}{2}\right)(1-\beta) \le \left(1-\frac{\beta}{2}\right), \\
&((1+\alpha p\eta\gamma)\beta - p\mu\gamma\eta) = \left(\beta + \frac{\beta^2}{2} - p\mu\gamma\eta\right) \le \left(\frac{3\beta}{2} - p\mu\gamma\eta\right) \le 0,
\end{aligned}
$$

and, therefore,

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

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

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

Applying Lemma 4, one can obtain

$$
\begin{aligned}
\mathbb{E}_k&\left[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{k+1}) - f(x^*))\right] \\
&\le \left(1-\beta/2\right)\|x^k - x^*\|^2 + \left(1 - p/\eta\right)\cdot 2\gamma\eta^2(f(x_f^k) - f(x^*)) \\
&\quad + p\eta^2\gamma^2\left(1+2p/\beta\right)\cdot C_2\tau^2M^{-2}B^{-2}(\sigma^2 + \delta^2\|\nabla f(x_g^k)\|^2) \\
&\quad + 2p^2\eta^2\gamma^2\left(1+\gamma L\right)\cdot\left(4C_1\tau B^{-1}\log_2 M + (4C_1+2)\tau^2B^{-2}\right)(\sigma^2 + \delta^2\|\nabla f(x_g^k)\|^2) \\
&\quad - p\gamma^2\eta^2(1-2p(1+\gamma L))\|\nabla f(x_g^k)\|^2.
\end{aligned}
$$

With $M \ge \sqrt{C_2 p^{-1}(1+2p/\beta)}$, we have

$$
\begin{aligned}
\mathbb{E}_k&\left[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2(f(x_f^{k+1}) - f(x^*))\right] \\
&\le \left(1-\beta/2\right)\|x^k - x^*\|^2 + \left(1 - p/\eta\right)\cdot 2\gamma\eta^2(f(x_f^k) - f(x^*)) \\
&\quad + p^2\eta^2\gamma^2\tau^2B^{-2}(\sigma^2 + \delta^2\|\nabla f(x_g^k)\|^2) \\
&\quad + 2p^2\eta^2\gamma^2\left(1+\gamma L\right)\cdot\left(4C_1\tau B^{-1}\log_2 M + (4C_1+2)\tau^2B^{-2}\right)(\sigma^2 + \delta^2\|\nabla f(x_g^k)\|^2) \\
&\quad - p\gamma^2\eta^2(1-2p(1+\gamma L))\|\nabla f(x_g^k)\|^2 \\
&\le \left(1-\beta/2\right)\|x^k - x^*\|^2 + \left(1 - p/\eta\right)\cdot 2\gamma\eta^2(f(x_f^k) - f(x^*)) \\
&\quad + 8p^2\eta^2\gamma^2\left(1+\gamma L\right)\cdot\left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2B^{-2}\right)\sigma^2 \\
&\quad - p\gamma^2\eta^2\left[1 - 2p\left(1+\gamma L\right)\left(1 + 4\left[C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2B^{-2}\right]\delta^2\right)\right]\|\nabla f(x_g^k)\|^2.
\end{aligned}
$$

Since $p = \left[1 + 2\left(1+\gamma L\right)\left(1 + 4\left[C_1 \tau b^{-1} + (C_1+1)\tau^2 b^{-2}\right]\delta^2\right)\right]^{-1}$, $B = \lceil b \log_2 M \rceil$ and $M \geq 2$, we obtain

$$
\begin{aligned}
p &= \left[1 + 2\left(1+\gamma L\right)\left(1 + 4\left[C_1 \tau b^{-1} + (C_1+1)\tau^2 b^{-2}\right]\delta^2\right)\right]^{-1} \\
&\leq \left[1 + 2\left(1+\gamma L\right)\left(1 + 4\left[C_1 \tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right]\delta^2\right)\right]^{-1},
\end{aligned}
$$

and then,

$$
\begin{aligned}
&\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2 (f(x_f^{k+1}) - f(x^*))\right] \\
&\qquad \leq (1-\beta/2)\|x^k - x^*\|^2 + (1 - p/\eta)\cdot 2\gamma\eta^2(f(x_f^k) - f(x^*)) \\
&\qquad\quad + 8p^2\eta^2\gamma^2\left(1+\gamma L\right)\cdot\left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2 \\
&\qquad \leq \max\left\{(1-\beta/2), (1-p/\eta)\right\}\left[\|x^k - x^*\|^2 + 2\gamma\eta^2(f(x_f^k) - f(x^*))\right] \\
&\qquad\quad + 8p^2\eta^2\gamma^2\left(1+\gamma L\right)\cdot\left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2.
\end{aligned}
$$

Using that $p\eta\gamma = 3\beta/(2\mu)$, $\beta/2 = p/\eta$, $B = \lceil b\log_2 M\rceil$ and $\gamma \leq L^{-1}$, we have

$$
\begin{aligned}
&\mathbb{E}_k\left[\|x^{k+1} - x^*\|^2 + 2\gamma\eta^2 (f(x_f^{k+1}) - f(x^*))\right] \\
&\qquad \leq (1-\beta/2)\left[\|x^k - x^*\|^2 + 2\gamma\eta^2(f(x_f^k) - f(x^*))\right] \\
&\qquad\quad + 36\beta^2\mu^{-2}\left(C_1\tau b^{-1} + (C_1+1)\tau^2 b^{-2}\right)\sigma^2. \qquad (23)
\end{aligned}
$$

Here we also took into account that $M \geq 2$. Finally, we perform the recursion and substitute $\beta = \sqrt{4p^2\mu\gamma/3}$

$$
\begin{aligned}
&\mathbb{E}\left[\|x^N - x^*\|^2 + 2\gamma\eta^2(f(x_f^N) - f(x^*))\right] \\
&\qquad \leq \left(1 - \sqrt{\frac{p^2\mu\gamma}{3}}\right)^N \left[\|x^0 - x^*\|^2 + 2\gamma\eta^2(f(x_f^0) - f(x^*))\right] \\
&\qquad\quad + 72\beta\mu^{-2}\left(C_1\tau b^{-1} + (C_1+1)\tau^2 b^{-2}\right)\sigma^2 \\
&\qquad \leq \exp\left(-\sqrt{\frac{p^2\mu\gamma N^2}{3}}\right)\left[\|x^0 - x^*\|^2 + 2\gamma\eta^2(f(x_f^0) - f(x^*))\right] \\
&\qquad\quad + \frac{144p\sqrt{\gamma}}{\sqrt{3}\mu^{3/2}}\left(C_1\sigma^2\tau b^{-1} + (C_1+1)\sigma^2\tau^2 b^{-2}\right).
\end{aligned}
$$

Substituting of $\eta = \sqrt{\frac{3}{\mu\gamma}}$ concludes the proof. □

B.4 Results of Section 2.1 with decreasing stepsize

The first thing we need to change is to make the parameters of Algorithm 1 depend on the iteration number k: $\gamma, p, \beta, \eta, M, B \to \gamma_k, p_k, \beta_k, \eta_k, M_k, B_k$. For this new version of Algorithm 1 one can reprove Theorem 1.

Theorem 7. *Assume A 1 – A 4. Let problem* (1) *be solved by Algorithm 1. Then for any $b \in \mathbb{N}^*$, $\gamma_k \in (0; \frac{3}{4L}]$, and $\beta_k, \theta_k, \eta_k, p_k, M_k, B_k$ satisfying*

$$
p_k \simeq (1 + (1+\gamma_k L)[\delta^2\tau b^{-1} + \delta^2\tau^2 b^{-2}])^{-1}, \quad \beta_k \simeq \sqrt{p_k^2\mu\gamma_k}, \quad \eta_k \simeq \sqrt{\tfrac{1}{\mu\gamma_k}},
$$

$$
\theta_k \simeq \tfrac{p_k\eta_k^{-1} - 1}{\beta_k p_k \eta^{-1} - 1}, \quad M_k \simeq \max\{2; \sqrt{p_k^{-1}(1 + p_k/\beta_k)}\}, \quad B_k = \lceil b\log_2 M_k\rceil,
$$

it holds that

$$
\begin{aligned}
&\mathbb{E}\left[\|x^{k+1} - x^*\|^2 + \frac{6}{\mu}(f(x_f^{k+1}) - f(x^*))\right] \\
&\qquad \leq \left(1 - \sqrt{\frac{p_k^2\mu\gamma_k}{3}}\right)\left[\|x^k - x^*\|^2 + \frac{6}{\mu}(f(x_f^k) - f(x^*))\right] \\
&\qquad\quad + \frac{48p_k^2\gamma_k}{\mu}\left(C_1\tau b^{-1} + (C_1+1)\tau^2 b^{-2}\right)\sigma^2.
\end{aligned}
$$

Proof. All steps of the proof remain the same with of Theorem 1 and we get (23):

$$\mathbb{E}\big[\|x^{k+1}-x^*\|^2+2\gamma_k\eta_k^2(f(x_f^{k+1})-f(x^*))\big]$$
$$\le(1-\beta_k/2)\left[\|x^k-x^*\|^2+2\gamma_k\eta_k^2(f(x_f^k)-f(x^*))\right]$$
$$+36\beta_k^2\mu^{-2}\left(C_1\tau b^{-1}+(C_1+1)\tau^2 b^{-2}\right)\sigma^2.$$

By substituting $\beta_k=\sqrt{4p_k^2\mu\gamma_k/3}$ and $\eta_k=\sqrt{\frac{3}{\mu\gamma_k}}$, we finishes the proof. □

Since $p_k=\left[1+2\left(1+\gamma_k L\right)\left(1+4\left[C_1\tau b^{-1}+(C_1+1)\tau^2 b^{-2}\right]\delta^2\right)\right]^{-1}$ and $\gamma_k\in(0;\frac{3}{4L})$, then $p_k\in[p_l;p_u]$, where $p_l,p_u\sim(1+(1+\tau b^{-1}+\tau b^{-2})\delta^2)^{-1}$. It means that we can rewrite the results of the theorem as follows:

$$\mathbb{E}\left[\|x^{k+1}-x^*\|^2+\frac{6}{\mu}(f(x_f^{k+1})-f(x^*))\right]$$
$$\le\left(1-\sqrt{\frac{p_l^2\mu\gamma_k}{3}}\right)\left[\|x^k-x^*\|^2+\frac{6}{\mu}(f(x_f^k)-f(x^*))\right]$$
$$+\frac{48p_u^2\gamma_k}{\mu}\left(C_1\tau b^{-1}+(C_1+1)\tau^2 b^{-2}\right)\sigma^2.$$

With notation $r_k=\mathbb{E}\left[\|x^k-x^*\|^2+\frac{6}{\mu}(f(x_f^k)-f(x^*))\right]$, $a=\sqrt{p_l^2\mu/3}$, $\omega_k=\sqrt{\gamma_k}$ and $C=\frac{48p_u^2}{\mu}\left(C_1\tau b^{-1}+(C_1+1)\tau^2 b^{-2}\right)\sigma^2$, one can rewrite the previous estimate:

$$r_{k+1}\le\left(1-a\omega_k\right)r_k+\omega_k^2 C,$$

where $0<\omega_k\le d=\sqrt{3/(4L)}$. For this kind of recursion, we can use the results of Lemma 3 of [88]. In particular, we can choose γ_k as follows

$$\begin{aligned}
&\text{if } N\le\frac{d}{a}, && \gamma_k=\frac{1}{d},\\
&\text{if } N>\frac{d}{a}\text{ and } k<\left\lceil\frac{N}{2}\right\rceil, && \gamma_k=\frac{1}{d},\\
&\text{if } N>\frac{d}{a}\text{ and } k\ge\left\lceil\frac{N}{2}\right\rceil, && \gamma_k=\frac{2}{a(k+\frac{2d}{a}+\lceil\frac{N}{2}\rceil)},
\end{aligned}$$

and get

$$r_N=\mathcal{O}\left(\frac{dr_0}{a}\exp\left(-\frac{aN}{2d}\right)+\frac{C}{a^2N}\right).$$

But the stepsize still depends on the horizon of iterations N. To fix it, we can apply the following restart procedure. We construct a sequence of the iteration number $N_t=2^t$ for $t\ge0$. For each restart t we set the stepsize $\gamma(N_t)$ according to Lemma 3 of [88], run the algorithm for N_t basic iterations. If we do not achieve the unknown horizon of the total iteration number N, then we use the obtained point as a warm-start for the next restart. For simplicity, we can also use the same starting x^0 point for all the restarts. Let us now assume that the algorithm made N iterations. This means that it made at least $T=\lfloor\log_2(N+1)\rfloor$ finished restarts. Since at the end of the last restart it made N_T basic iterations with the stepsize $\gamma(N_T)$, we can guarantee that

$$r_{N_T}=\mathcal{O}\left(\frac{dr_0}{a}\exp\left(-\frac{aN_T}{2d}\right)+\frac{C}{a^2N_T}\right).$$

One can note that $N_T\sim N$, then

$$r_{N_T}=\mathcal{O}\left(\frac{dr_0}{a}\exp\left(-\frac{aN}{2d}\right)+\frac{C}{a^2N}\right).$$

This algorithm does not require to fix the number of basic steps N in advance, but if we want to have ε-solution in terms of r_N, then we have the following estimate on the number of iterations:

$$N=\mathcal{O}\left(\frac{d}{a}\log\frac{1}{\varepsilon}+\frac{C}{a^2\varepsilon}\right)=\mathcal{O}\left(\left[1+(1+\tau b^{-1}+\tau^2 b^{-2})\delta^2\right]\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}+\frac{\sigma^2}{\mu^2\varepsilon}\left(\tau b^{-1}+\tau^2 b^{-2}\right)\right).$$

To get the close to Corollary 1 results on the oracle complexity one need to take $b = \tau$ and note that now $B_k = b\log_2 M_k = b\log_2 M_k \sim b\log_2 N \sim b\log_2 \varepsilon^{-1}$. Finally, it gives additional logarithmic factor in the estimate for the oracle complexity. But this factor does not really change the bound and it means that we obtain the result of Corollary 1.

B.5 Lower bounds proofs

Proof of Theorem 2.

We begin the proof with two lemmas, showing the lower bounds for deterministic and stochastic components of the error separately. Then we combine the two in Theorem 8 and complete the proof of Theorem 2. First, we consider the lower bound for the deterministic part of the error, and construct a problem with $\delta = 1$ and $\sigma = 0$.

Lemma 7. *There exists an instance of the optimization problem satisfying assumptions A 1 –A 4 with $\delta = 1$ and $\sigma = 0$, such that for any first-order gradient method it takes at least*

$$N = \Omega\big(\tau\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\big)$$

oracle calls in order to achieve $\mathbb{E}[\|x^N - x^*\|^2] \leq \varepsilon$.

Proof. Consider the optimization problem

$$f_1(x) = \frac{\mu(Q-1)}{4}\big(\frac{x^\top Ax}{2} - e_1^\top x\big) + \frac{\mu}{2}\|x\|^2 \to \min_{x\in\mathbb{R}^d}\,, \tag{24}$$

where $x \in \mathbb{R}^d$, $\mu > 0, Q > 1$, dimension d is even, $d = 2u$, $e_1 = (1, 0, \ldots, 0) \in \mathbb{R}^d$ is the first coordinate vector, and $A \in \mathbb{R}^{d\times d}$ is a symmetric nonnegative-definite matrix given by

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & \ldots & 0 & 0 \\ -1 & 2 & -1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 2 & -1 & \ldots & 0 & 0 \\ & & & & \ldots & & \\ 0 & 0 & 0 & 0 & \ldots & 2 & -1 \\ 0 & 0 & 0 & 0 & \ldots & -1 & \alpha\,, \end{pmatrix} \tag{25}$$

where $\alpha = \frac{\sqrt{Q}+3}{\sqrt{Q}+1}$. Straightforward calculations (see e.g. [59, Chapter 5.1.4] for more details) yield

$$0 \preceq A \preceq 4I\,, \nabla f_1(x) = \frac{\mu(Q-1)}{4}Ax - e_1 + \mu x\,.$$

Thus the problem (24) is $L-$smooth with $L = \mu Q$ and μ-strongly convex, i.e., the assumptions A 1 and A 2 are satisfied, and the corresponding condition number is equal to $L/\mu = Q$. For $\epsilon \in (0; 1/2)$ we now consider the two-state Markov transition matrix (or kernel)

$$\mathrm{P}_1 = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix} \tag{26}$$

and denote by $(Z_i)_{i=1}^{\infty}$ the corresponding Markov Chain, $Z_i \in \{-1, 1\}$. It is easy to see that the Markov kernel P is uniformly geometrically ergodic and satisfies A 4 with $\tau \leq \epsilon^{-1}\log 4$. It is easy to check that the corresponding invariant distribution is $\pi = (1/2, 1/2)$. For $Z \in \{-1, 1\}$ we now consider the noise matrix

$$W(Z) = 2\,\mathrm{diag}\big\{\mathbb{1}_{\{Z=1\}}, \mathbb{1}_{\{Z=-1\}}, \mathbb{1}_{\{Z=1\}}, \ldots, \mathbb{1}_{\{Z=-1\}}\big\} \in \mathbb{R}^{d\times d}\,.$$

Now for $x \in \mathbb{R}^d$ and $Z \in \{-1, 1\}$ we define the stochastic gradient oracle as

$$\nabla F_1(x, Z) = W(Z)\nabla f(x)\,. \tag{27}$$

It is easy to check that $\mathbb{E}_\pi[W(Z)] = \mathrm{I}$, and the direct calculations imply $\|\nabla F_1(x, Z) - \nabla f_1(x)\| \leq \|\nabla f_1(x)\|$, that is, the assumption A 4 holds with $\delta = 1$ and $\sigma = 0$. Following [76], [59, Chapter 5.1.4], the solution to the minimization problem (24) is given by

$$x^* = (q^1, \ldots, q^d) \in \mathbb{R}^d\,, q = \frac{\sqrt{Q}-1}{\sqrt{Q}+1}\,. \tag{28}$$

Suppose that we start from $Z_1 = 1$ and initial point $x_0 = 0 \in \mathbb{R}^d$. Then after 1 oracle call we observe the 1-st coordinate of x. At the same time, the second component can not be computed until the time moment $T_2 = \inf\{i \in \mathbb{N} : Z_i = -1\}$. Similarly, the next computation of the 3-rd component of the solution requires the chain to go back to state 1 and can not happen earlier then $T_3 = \inf\{i \geq \tau_2 : Z_i = 1\}$. Thus, after k iterations of any first-order method, the respective MSE is lower bounded by

$$\mathbb{E}_\pi[\|x^k - x^*\|^2] \geq \mathbb{E}_\pi \left[\mathbb{1}_{\{Z_1=1\}} \sum_{i=N_k}^{d} q^{2i} \right] = \frac{1}{2}\mathbb{E}_{\delta_1}\left[\frac{q^{2N_k} - q^{2d}}{1-q^2}\right] .$$

In the formula above we denoted by N_k the number of state changes in the sequence $(Z_i)_{i=1}^k$. Using Jensen's inequality and the explicit construction of the Markov kernel P_1 in (26), we deduce that

$$\begin{aligned}
\mathbb{E}_\pi[\|x^k - x^*\|^2] \geq \frac{1}{2}\frac{q^{2\mathbb{E}_{\delta_1}[N_k]} - q^{2d}}{1-q^2} = \frac{1}{2}\frac{q^{2(k-1)\epsilon} - q^{2d}}{1-q^2} \geq (1/2)(1-q^2)^{-1} q^{(2/\log 4)k/\tau} = \\
= (1/2)(1-q^2)^{-1}\left(1 - \frac{2}{\sqrt{Q}+1}\right)^{(2/\log 4)k/\tau} \\
\geq (1/2)(1-q^2)^{-1}\exp\left(-\frac{8k}{(\sqrt{Q}+1)\tau\log 4}\right) ,
\end{aligned}$$

provided that d is large enough. In the last inequality we also used that $1 - x \geq e^{-2x}$ for $x \in [0; 1/2]$. Hence, taking into account that Q is the condition number of the problem (24), we get the desired lower bound. ☐

Now we consider an instance of the problem with $\delta = 0$, arbitrary $\sigma \geq 0$, and construct the respective lower bound for the stochastic part of the error.

Lemma 8. *There exists an instance of the optimization problem satisfying assumptions A 1 –A 4 with $\delta = 0$ and arbitrary $\sigma \geq 0$, such that for any first-order gradient method it takes at least*

$$N = \Omega\left(\frac{\tau\sigma^2}{\mu^2\varepsilon}\right)$$

oracle calls in order to achieve $\mathbb{E}[\|x^N - x^*\|^2] \leq \varepsilon$.

Proof. Our proof is based on a simple 1-dimensional optimization problem and Le Cam's lemma [1, Theorem 8], see also [105]. Consider the following minimization problem

$$f_2(x) = \frac{\mu}{2}(x - x^*)^2 \mapsto \min_{x\in\mathbb{R}} . \tag{29}$$

Obviously this problem satisfies A 2 with strong convexity constant μ and A 1 with $L = \mu$. Consider the noisy gradient oracle

$$\nabla F_2(x, Y) = \mu(x - x^*) + \frac{\sigma}{2}Y , \tag{30}$$

where Y is a noise variable taking values $Y \in \{-1, 1\}$. For now we do not specify the distribution of Y, yet we easily note that for any distribution π on $\{-1, 1\}$ we have

$$\|\nabla F_2(x, Y) - \mathbb{E}_\pi \nabla F_2(x, Y)\|^2 \leq \sigma^2 .$$

Consider the sequence of noise variables $(Y_i)_{i=1}^n$ with the joint distribution to the specified later, and any sequence of design points $(x_i)_{i=1}^n$, where the resulting gradients are evaluated. At this point the statistician observes the gradients

$$(\mu(x_i - x^*) + \frac{\sigma}{2}Y_i) \, , i = 1 \ldots, n \, ,$$

and, since x_i and μ are known, this is equivalent to observing

$$x^* - \frac{\sigma}{2\mu}Y_i \, , i = 1 \ldots, n.$$

Now we aim to construct to "almost indistinguishable" models for the noise variables Y_i. Namely, we consider the parametric family of Markov kernels

$$
\mathrm{P}_\varphi = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon+\varphi & 1-\epsilon-\varphi \end{pmatrix}\,, \tag{31}
$$

where the parameters $\varphi, \epsilon \in (0; 1/4)$, and $\varphi \in [0; \alpha]$, and the parameter α will be set depending on ϵ and n later. It is easy to check that the invariant distribution of the Markov kernel P_φ is given by

$$
\pi^\varphi = \big(\frac{\epsilon+\varphi}{2\epsilon+\varphi}, \frac{\epsilon}{2\epsilon+\varphi}\big)\,.
$$

Now we consider the setting of Le Cam's lemma [1, Theorem 8]. Namely, for a fixed sample size n we consider the family of Markov kernels $(\mathrm{P}_\varphi)_{\varphi\in[0;\alpha]}$, and family of corresponding joint $n-$step distributions under stationarity, that is, $\pi^\varphi \mathrm{P}_\varphi^{\otimes n}$. The reader not familiar with the respective notation could find it, in particular, in [22, Chapter 1]. As a parameter of interest we consider the expectation

$$
\theta(\varphi) := \theta(\pi^\varphi \mathrm{P}_\varphi^{\otimes n}) := \mathbb{E}_{\pi^\varphi}[x^* - \frac{\sigma}{2\mu} Z_i] = x^* - \frac{\sigma\varphi}{2\mu(2\epsilon+\varphi)}\,. \tag{32}
$$

Now we consider the 2 representatives of the above class, that is, the n-step distributions corresponding the parameters $\varphi = 0$ and $\varphi = \alpha$. Then the direct application of Le Cam's lemma [1, Theorem 8] yields

$$
\inf_{\widehat{\theta}} \sup_{\varphi\in[0;\alpha]} \mathbb{E}^{1/2}_{\pi^\varphi \mathrm{P}_\varphi^{\otimes n}}[|\widehat{\theta} - \theta(\varphi)|^2] \geq \frac{1}{2}|\theta(0) - \theta(\alpha)|(1 - \|\pi^0 \mathrm{P}_0^{\otimes n} - \pi^\alpha \mathrm{P}_\alpha^{\otimes n}\|_{\mathrm{TV}})\,, \tag{33}
$$

where $\widehat{\theta} = \widehat{\theta}(Y_1, \ldots, Y_n)$ is any measurable function. Thus, taking square and using the definition of $\theta(\varphi)$ in (32), we obtain that

$$
\inf_{\widehat{\theta}} \sup_{\varphi\in[0;\alpha]} \mathbb{E}_\pi[|\theta - \theta(\varphi)|^2] \geq \frac{\sigma^2\alpha^2}{16\mu^2(2\epsilon+\alpha)^2}(1 - \|\pi^0 \mathrm{P}_0^{\otimes n} - \pi^\alpha \mathrm{P}_\alpha^{\otimes n}\|_{\mathrm{TV}})\,. \tag{34}
$$

Now we set $\alpha = \sqrt{\frac{\epsilon}{n}}$ and apply the statement of Lemma 9 with this choice of α. Note that we impose at this point the regularity condition $n \geq \epsilon^{-1}$ in order to have $\alpha \leq \epsilon$. Thus we get

$$
\inf_{\widehat{\theta}} \sup_{\varphi\in[0;\sqrt{\frac{\epsilon}{n}}]} \mathbb{E}_\pi[|\widehat{\theta} - \theta(\varphi)|^2] \geq \frac{\sigma^2\epsilon}{32\mu^2 n(2\epsilon + \sqrt{\frac{\epsilon}{n}})^2} \geq \frac{\sigma^2}{288\mu^2 n\epsilon}\,,
$$

and the statement follows by noticing that the corresponding mixing time $\tau \leq c\epsilon^{-1}$ for some $c > 0$ (see e.g. [71, Proposition 1]). □

Lemma 9. *Consider the family of Markov kernels*

$$
\mathrm{P}_\varphi = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon+\varphi & 1-\epsilon-\varphi \end{pmatrix}
$$

and the corresponding invariant distributions $\pi^\varphi = \big(\frac{\epsilon+\varphi}{2\epsilon+\varphi}, \frac{\epsilon}{2\epsilon+\varphi}\big)$ *for* $\varphi \in \{0, \alpha\}$. *Then it holds that*

$$
\|\pi^0 \mathrm{P}_0^{\otimes n} - \pi^\alpha \mathrm{P}_\alpha^{\otimes n}\|_{\mathrm{TV}} \leq \frac{1}{2}\sqrt{\frac{n\alpha^2}{\epsilon}}\,.
$$

Proof. Note first that an application of Pinsker's inequality yields

$$
\|\pi^0 \mathrm{P}_0^{\otimes n} - \pi^\alpha \mathrm{P}_\alpha^{\otimes n}\|_{\mathrm{TV}} \leq \sqrt{(1/2)\,\mathrm{KL}(\pi^0 \mathrm{P}_0^{\otimes n} || \pi^\alpha \mathrm{P}_\alpha^{\otimes n})}\,.
$$

Using the chain rule for KL-divergence, we get

$$
\mathrm{KL}(\pi^0 \mathrm{P}_0^{\otimes n} || \pi^\alpha \mathrm{P}_\alpha^{\otimes n}) = \mathrm{KL}(\pi^0 || \pi^\alpha) + \sum_{i=1}^{n-1} \sum_{y\in\{-1,1\}} \mathrm{P}_{\pi^0 \mathrm{P}_0^{\otimes n}}(Y_i = y)\,\mathrm{KL}(\mathrm{P}_0(\cdot|y) || \mathrm{P}_\alpha(\cdot|y))\,. \tag{35}
$$

In the notation above for $y \in \{-1, 1\}$ we have set $\mathrm{KL}(\mathrm{P}_0(\cdot|y)||\mathrm{P}_\alpha(\cdot|y))$ for the $1-$step conditional distribution

$$\mathrm{KL}(\mathrm{P}_0(\cdot|y)||\mathrm{P}_\alpha(\cdot|y)) = \sum_{x \in \{-1,1\}} \mathrm{P}_0(x|y) \log \frac{\mathrm{P}_0(x|y)}{\mathrm{P}_\alpha(x|y)} .$$

Now an application of reversed Pinsker's inequality together with $\alpha \leq \epsilon$ yields that

$$\mathrm{KL}(\mathrm{P}_0(\cdot|y)||\mathrm{P}_\alpha(\cdot|y)) \leq \frac{\alpha^2}{2\epsilon} ,$$

and the bound (35) implies that

$$\mathrm{KL}(\pi^0 \mathrm{P}_0^{\otimes n} || \pi^\alpha \mathrm{P}_\alpha^{\otimes n}) \leq \frac{n\alpha^2}{2\epsilon} .$$

Combining the bounds above yields the statement. □

Now we are ready to combine the bounds above and prove Theorem 2.

Theorem 8 (Theorem 2)**.** *There exists an instance of the optimization problem satisfying assumptions A 1 –A 4 with $\delta = 1$ and arbitrary $\sigma \geq 0, L, \mu > 0, \tau \in \mathbb{N}^*$, such that for any first-order gradient method it takes at least*

$$N = \Omega\big(\tau \sqrt{\tfrac{L}{\mu}} \log \tfrac{1}{\varepsilon} + \tfrac{\tau\sigma^2}{\mu^2\varepsilon}\big)$$

oracle calls in order to achieve $\mathbb{E}[\|x^N - x^*\|^2] \leq \varepsilon$.

Proof. We split the original problem into two parts. Indeed, for any $d \in \mathbb{N}^*$ we consider $x = (x_{det}, x_{stoch}) \in \mathbb{R}^{d+1}$, where $x_{det} \in \mathbb{R}^d$ and $x_{stoch} \in \mathbb{R}$. Now we consider the minimization problem

$$f(x) = f(x_{det}, x_{stoch}) = f_1(x_{det}) + f_2(x_{stoch}) \to \min_{x \in \mathbb{R}^{d+1}} , \tag{36}$$

where the functions $f_1 : \mathbb{R}^d \to \mathbb{R}$ and $f_2 : \mathbb{R} \to \mathbb{R}$ are defined in (24) and (29), respectively. We fix the respective parameters μ, Q, and σ. Applying Lemma 7 and Lemma 8, we get that the respective problem (36) is L-smooth and μ-strongly convex with $L = \mu Q$ and parameter $Q > 1$ defined in (24). For $Z, Y \in \{-1, 1\}$ we define the stochastic gradient oracle as

$$\nabla F(x, Z, Y) = (\nabla F_1(x_{det}, Z), \nabla F_2(x_{stoch}, Y)) \in \mathbb{R}^{d+1} .$$

The oracles $\nabla F_1(x_{det}, Z)$ and $\nabla F_2(x_{stoch}, Y)$ are defined in (27) and (30), respectively. Lemma 7 and Lemma 8 imply that A 4 holds with $\delta = 1$ and $\sigma > 0$ defined in (30). Consider now the Markov chains $(Z_i)_{i=1}^\infty$ with the transition kernel P_1 defined in (26) and $(Y_i)_{i=1}^\infty$ with the transition kernel P_φ of the form (31). As in the proof of Lemma 8, we take $\varphi \in [0; \sqrt{\epsilon/n}]$ and assume that $n \geq \epsilon^{-1}$. Consider the joint process $(X_i, Y_i)_{i=1}^\infty$ of independently evolving Markov chains $(Z_i)_{i=1}^\infty$ and $(Y_i)_{i=1}^\infty$. It is easy to see that such a process is a Markov chain on $\{-1, 1\}^2$ with the transition kernel

$$\mathrm{P} = \mathrm{P}_1 \otimes \mathrm{P}_\varphi ,$$

where $\otimes$ stands for the Kronecker's product. In is clear that P is irreducible and aperiodic, hence the assumption A 3 holds. Note that both P_1 and P_φ are reversible (see e.g. [81][Section 3.1] for the respective definitions). Thus their Kronecker's product is also reversible, with the spectrum given by the pairwise products of eigenvalues of P_1 and P_φ. Hence, with the direct calculations, we compute the eigenvalues of P: $\{1, 1 - 2\epsilon - \varphi, 1 - 2\epsilon, (1 - 2\epsilon)(1 - 2\epsilon - \varphi)\}$. Thus the corresponding spectral gap $\gamma = 2\epsilon$, and the mixing time τ of P is bounded by

$$\frac{1}{2\epsilon(1 + 1/\log 2)} \leq \tau \leq \frac{2\log 2 + \log 6}{4\epsilon} ,$$

see [81][Proposition 3.3]. Hence, the mixing time of the corresponding joint chain scales as ϵ^{-1}, as for $(Z_i)_{i=1}^\infty$ and $(Y_i)_{i=1}^\infty$ separately. On the k-th step of the stochastic gradient computations we rely on the stochastic gradient

$$\nabla F(x_k, Z_k, Y_k) ,$$

computed using the pair (Z_k, Y_k). To complete the proof it remains to apply the complexity results of Lemma 7 and Lemma 8 to the parts x_{det} and x_{stoch}, respectively. □

Proposition 3 (Proposition 1)**.** *There exists an instance of the optimization problem satisfying assumptions A 1 –A 4 with arbitrary $L, \mu > 0, \tau \in \mathbb{N}^*$, $\delta = \frac{L}{\mu}$, and $\sigma = 0$, such that for any first-order gradient method it takes at least*

$$N = \Omega\big(\tau \tfrac{L}{\mu} \log \tfrac{1}{\varepsilon}\big)$$

gradient calls in order to achieve $\mathbb{E}[\|x^N - x^*\|^2] \leq \varepsilon$.

Proof. In this part we closely follow the setting of [71]. We consider the setting of linear regression:

$$f(x) = \frac{1}{2}\mathbb{E}_{(\varphi,Y)\sim\mathcal{D}}\big[|Y - \varphi^\top x|^2\big] \to \min_x \,, \tag{37}$$

where $\varphi \in \mathbb{R}^d$ is a (random) feature vector, $Y \in \mathbb{R}$ is a (random) regressor, with the joint distribution $(\varphi, Y) \sim \mathcal{D}$, and $x \in \mathbb{R}^d$ is the optimized parameter. We consider the so-called realizable case, that is, we assume that

$$Y = \varphi^\top x^*$$

for some vector $x^* \in \mathbb{R}^d$. In this scenario the problem (37) reduces to

$$f(x) = \frac{1}{2}(x - x^*)^\top \Sigma^2 (x - x^*) \to \min_x \,,$$

where we have denoted $\Sigma^2 = \mathbb{E}_\pi[\varphi\varphi^\top]$. This means that the exact gradient is given by $\nabla f(x) = \Sigma^2(x - x^*)$. Now we consider the stochastic setting of the online regression with sequentially observed data points $(\varphi_i, Y_i)_{i=1}^N$ with $Y_i = \varphi_i^\top x^*$. In this case the i-th realization of stochastic gradient at point $x \in \mathbb{R}^d$ is given by

$$\nabla F(x, \varphi_i, Y_i) = \varphi_i(\varphi_i^\top x - Y_i) = \varphi_i\varphi_i^\top(x - x^*) \,.$$

Hence, with a simple algebra we get

$$\|\nabla F(x, \varphi_i, Y_i) - \nabla f(x)\| = \|(\Sigma^2 - \varphi_i\varphi_i^\top)(x - x^*)\| = \|(\mathrm{I} - \varphi_i\varphi_i^\top\Sigma^{-2})\nabla f(x)\| \,,$$

where we have used the fact that $x - x^* = (\Sigma^2)^{-1}\nabla f(x)$ and used additional notation $\Sigma^{-2} := (\Sigma^2)^{-1}$. Fix now the condition number $Q > 1$, parameter $\epsilon \in (0; 1/4)$ and consider the Markov kernel

$$\mathrm{P} = \begin{pmatrix} 1 - \frac{\epsilon}{Q-1} & \frac{\epsilon}{Q-1} \\ \epsilon & 1 - \epsilon \end{pmatrix}$$

and the corresponding canonical chain $(Z_i)_{i=1}^N$. The invariant distribution of P is given by $\pi = (1 - 1/Q, 1/Q)$, and the corresponding mixing time τ is bounded by

$$\tau \leq \frac{(Q-1)\log 4}{Q\epsilon} \,,$$

see e.g. [71, Proposition 1]. We let $\varphi = \varphi(Z)$, and w.l.o.g. we can assume that $Z \in \{-1, 1\}$. Consider

$$\varphi(1) = (1, 0) \,, \quad \varphi(-1) = (0, 1) \,.$$

The design matrix Σ^2 is given by

$$\Sigma^2 = \mathbb{E}_\pi[\varphi(Z_i)\varphi(Z_i)^\top] = \begin{pmatrix} 1 - 1/Q & 0 \\ 0 & 1/Q \end{pmatrix} \,,$$

which implies that A 1 and A 2 are satisfied with $\mu = 1/Q$ and $L = 1 - 1/Q$. Then the direct calculations yield

$$\|\nabla F(x, \varphi(Z_i), Y_i) - \nabla f(x)\| \leq (Q - 1)\|\nabla f(x)\| \,,$$

and the assumption A 4 is satisfied with $\delta = Q - 1$. Then the direct application of lower bound [71] implies the lower bound

$$\mathbb{E}_\pi[\|x^k - x^*\|^2] \geq \exp\left(-\frac{ck}{Q\tau}\right)$$

after k iterations of any first-order method with Markovian sampling oracle defined above. Here $c > 0$ is some absolute positive constant not dependeing upon τ and Q. This means that the instance-dependent increase of δ yields to inevitably slower convergence rates. □

Proposition 4 (Proposition 2)**.** *There exists an instance of the optimization problem satisfying assumptions A 1 –A 4 with with arbitrary $L, \mu > 0, \tau \in \mathbb{N}^*$, $\sigma = 1, \delta = 0$, such that for any first-order gradient method it takes at least*

$$N = \Omega\left(\left(\tau + \sqrt{\tfrac{L}{\mu}}\right)\log\{\tfrac{1}{\varepsilon}\}\right)$$

oracle calls in order to achieve $\mathbb{E}[\|x^N - x^*\|^2] \le \varepsilon$.

Proof. Let us consider the same minimization problem (24) as in the proof of Theorem 2. Recall that the true gradient in this setting is given by

$$\nabla f(x) = \frac{\mu(Q-1)}{4}Ax - e_1 + \mu x\,.$$

Hence the problem (24) is $L-$smooth with $L = \mu Q$ and μ-strongly convex, that is, assumptions A 1 and A 2 are satisfied, and the corresponding condition number equals $L/\mu = Q$. Now for $\epsilon \in (0; 1/2)$ we consider the discrete-state space Markov kernel

$$\mathrm{P} = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix} \tag{38}$$

and the corresponding Markov Chain $(Z_i)_{i=1}^{\infty}$. It is easy to see that the Markov kernel P is uniformly geometrically ergodic and satisfies A 4 with $\tau \le \epsilon^{-1}\log 4$. Each Z_i takes 2 different values, and w.l.o.g. we can assume that $Z_i \in \{-1, 1\}$. It is easy to check that the corresponding invariant distribution is $\pi = (1/2, 1/2)$. For $Z \in \{-1, 1\}$ we now consider the noisy oracle

$$\nabla F(x, Z) = \frac{\mu(Q-1)}{4}Ax - (1 + \mathbb{1}_{\{Z=-1\}} - \mathbb{1}_{\{Z=1\}})e_1 + \mu x\,.$$

It is easy to check that $\mathbb{E}_\pi[\nabla F(x, Z)] = \nabla f(x)$, and the direct calculations imply $\|\nabla F(x, Z) - \nabla f(x)\| \le 1$, that is, the assumption A 4 holds with $\delta = 0$ and $\sigma = 1$. Suppose that we start from $Z_1 = 1$ and initial point $x_0 = 0 \in \mathbb{R}^d$. Then we observe $\nabla F(x, Z) = 0 \in \mathbb{R}^d$ unless the time moment $T_2 = \inf\{i \in \mathbb{N} : Z_i = -1\}$. Thus, after k iterations of any first-order method, the respective MSE is lower bounded by

$$\begin{aligned}
\mathbb{E}_\pi[\|x^k - x^*\|^2] &\ge \mathbb{E}_\pi\left[\mathbb{1}_{\{Z_1=1\}}\sum_{i=k}^{d} q^{2i}\right] + \mathbb{E}_\pi\left[\mathbb{1}_{\{Z_1=1, T_2 \ge k\}}(1 - q^{2d})\right] \\
&\ge \frac{1}{2}(1-q^2)^{-1}(q^{2k} - q^{2d}) + \frac{1}{2}(1-q^2)^{-1}\mathbb{P}_{\delta_1}(T_2 \ge k) \\
&= \frac{1}{2}(1-q^2)^{-1}(q^{2k} - q^{2d}) + \frac{1}{2}(1-q^2)^{-1}(1-\epsilon)^{k-1}\,.
\end{aligned}$$

Hence, with the defition of q in (28), we get from the previous bound that

$$\mathbb{E}_\pi[\|x^k - x^*\|^2] \ge \frac{1}{2}(1-q^2)^{-1}\left[\exp\left(-\frac{4k}{(\sqrt{Q}+1)}\right) - q^{2d}\right] + \frac{1}{2}(1-q^2)^{-1}\exp\left(-\frac{2k}{\tau\log 4}\right)\,,$$

where in the last inequality we also used that $1 - x \ge e^{-2x}$ for $x \in [0; 1/2]$. Now the statement follows from the definition of $Q = L/\mu$. □

B.6 Proof of Theorem 3

Theorem 9 (Theorem 3)**.** *Assume A 1, A 3, A 4. Let problem* (1) *be solved by Algorithm 2. Let f^* be a global (maybe not unique) minimum of f. Then for any $b \in \mathbb{N}^*$, and γ, M satisfying*

$$\gamma \le \left[4L\left(1 + 4\left[C_1\tau b^{-1} + (C_1+1)\tau^2 b^{-2}\right]\delta^2\right)\right]^{-1}\,,$$
$$M = \max\{2; \sqrt{C_2\gamma^{-1}L^{-1}}\}, \quad B = \lceil b\log_2 M\rceil,$$

it holds that

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=0}^{N-1}\|\nabla f(x^k)\|^2\right] \lesssim \frac{f(x^0) - f^*}{\gamma N} + L\gamma\cdot\left[\sigma^2\tau b^{-1} + \sigma^2\tau^2 b^{-2}\right]\,.$$

Proof. We start from A 1 (in the form (42) with $x = x^{k+1}$ and $y = x^k$) and line 6 of Algorithm 2:

$$
\begin{aligned}
f(x^{k+1}) &\le f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2}\|x^{k+1} - x^k\|^2 \\
&\le f(x^k) - \gamma\langle \nabla f(x^k), g^k \rangle + \frac{\gamma^2 L}{2}\|g^k\|^2 \\
&= f(x^k) - \gamma\langle \nabla f(x^k), \nabla f(x^k) \rangle - \gamma\langle \nabla f(x^k), \mathbb{E}_k[g^k] - \nabla f(x^k) \rangle \\
&- \gamma\langle \nabla f(x^k), g^k - \mathbb{E}_k[g^k] \rangle + \frac{L\gamma^2}{2}\mathbb{E}_k[\|g^k\|^2]
\end{aligned}
$$

Subtracting f^* from both sides, using Cauchy Schwartz inequality (43) and taking the conditional expectation, we get

$$
\begin{aligned}
\mathbb{E}_k[f(x^{k+1}) - f^*] \le & f(x^k) - f^* - \gamma\|\nabla f(x^k)\|^2 + \frac{\gamma}{2}\|\nabla f(x^k)\|^2 \\
& + \frac{\gamma}{2}\|\mathbb{E}_k[g^k] - \nabla f(x_g^k)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}_k[\|g^k\|^2] \\
= & f(x^k) - f^* - \frac{\gamma}{2}\|\nabla f(x^k)\|^2 + \frac{\gamma}{2}\|\mathbb{E}_k[g^k] - \nabla f(x^k)\|^2 + \frac{L\gamma^2}{2}\mathbb{E}_k[\|g^k\|^2].
\end{aligned}
$$

Reapplying Cauchy Schwartz inequality (44) one more time, we have

$$
\begin{aligned}
\mathbb{E}_k[f(x^{k+1}) - f^*] \le & f(x^k) - f^* - \frac{\gamma}{2}(1 - 2\gamma L)\|\nabla f(x^k)\|^2 \\
& + \frac{\gamma}{2}\|\mathbb{E}_k[g^k] - \nabla f(x^k)\|^2 + L\gamma^2\mathbb{E}_k[\|g^k - \nabla f(x^k)\|^2]\,.
\end{aligned}
$$

Lemma 4 with x_g^k replaced by x^k gives

$$
\begin{aligned}
\mathbb{E}_k[f(x^{k+1}) - f^*] \le & f(x^k) - f^* - \frac{\gamma}{2}(1 - 2\gamma L)\|\nabla f(x^k)\|^2 \\
& + \frac{\gamma}{2} \cdot C_2\tau^2 M^{-2} B^{-2}(\sigma^2 + \delta^2\|\nabla f(x^k)\|^2) \\
& + L\gamma^2 \cdot \left(4C_1\tau B^{-1}\log_2 M + (4C_1 + 2)\tau^2 B^{-2}\right)\left(\sigma^2 + \delta^2\|\nabla f(x^k)\|^2\right).
\end{aligned}
$$

With $M \ge \sqrt{C_2\gamma^{-1}L^{-1}}$, we have

$$
\begin{aligned}
\mathbb{E}_k[f(x^{k+1}) - f^*] \le & f(x^k) - f^* - \frac{\gamma}{2}(1 - 2\gamma L)\|\nabla f(x^k)\|^2 \\
& + \frac{L\gamma^2}{2} \cdot \tau^2 B^{-2}(\sigma^2 + \delta^2\|\nabla f(x^k)\|^2) \\
& + L\gamma^2 \cdot \left(4C_1\tau B^{-1}\log_2 M + (4C_1 + 2)\tau^2 B^{-2}\right)\left(\sigma^2 + \delta^2\|\nabla f(x^k)\|^2\right) \\
\le & f(x^k) - f^* \\
& - \frac{\gamma}{2}\left[1 - 2\gamma L\left(1 + 4\left[C_1\tau B^{-1}\log_2 M + (C_1 + 1)\tau^2 B^{-2}\right]\delta^2\right)\right]\|\nabla f(x^k)\|^2 \\
& + 4L\gamma^2 \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1 + 1)\tau^2 B^{-2}\right)\sigma^2\,.
\end{aligned}
$$

Since $\gamma \le \left[4L\left(1 + 4\left[C_1\tau b^{-1} + (C_1 + 1)\tau^2 b^{-2}\right]\delta^2\right)\right]^{-1}$, $B = \lceil b\log_2 M\rceil$ and $M \ge 2$, one can obtain

$$
\begin{aligned}
\gamma &\le \left[4L\left(1 + 4\left[C_1\tau b^{-1} + (C_1 + 1)\tau^2 b^{-2}\right]\delta^2\right)\right]^{-1} \\
&\le \left[4L\left(1 + 4\left[C_1\tau B^{-1}\log_2 M + (C_1 + 1)\tau^2 B^{-2}\right]\delta^2\right)\right]^{-1},
\end{aligned}
$$

and then,

$$
\begin{aligned}
\mathbb{E}_k[f(x^{k+1}) - f^*] \le & f(x^k) - f^* - \frac{\gamma}{4}\|\nabla f(x^k)\|^2 \\
& + 4L\gamma^2 \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1 + 1)\tau^2 B^{-2}\right)\sigma^2\,.
\end{aligned} \tag{39}
$$

By doing a small rearrangements, summing over all k from 0 to $N - 1$, averaging over N iterations, taking the full expectation of both sides, we get

$$
\mathbb{E}\left[\frac{1}{N}\sum_{k=0}^{N-1}\|\nabla f(x^k)\|^2\right] \le \frac{4(f(x^0) - f^*)}{\gamma N} + 16L\gamma \cdot \left[C_1\sigma^2\tau B^{-1}\log_2 M + (C_2 + 1)\sigma^2\tau^2 B^{-2}\right].
$$

Substituting $B = \lceil b\log_2 M\rceil$ and using $M \ge 2$ finish the proof. $\square$

B.7 Result for Polyak-Loiasyewitch condition

A 9. *The function f satisfies PL condition on $\mathbb{R}^d$ with $\mu > 0$, i.e. the following inequality holds for all $x \in \mathbb{R}^d$:*

$$\|\nabla f(x)\| \geq 2\mu(f(x) - f^*),$$

where f^* is a global (potentially not unique) minimum of f.

Corollary 5. *Under the conditions of Theorem 3 and A 9, if we choose $b = \tau$ and γ given by*

$$\gamma \simeq \min\left\{\frac{1}{(1+\delta^2)L}; \frac{1}{\mu N}\ln\left(\max\left\{2; \frac{\mu^2 N(f(x^0) - f^*)}{L\sigma^2}\right\}\right)\right\}, \tag{40}$$

then to achieve ε-solution (in terms of $\mathbb{E}[f(x) - f^*] \lesssim \varepsilon$) we need

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

Proof. We start from (39) and apply A 9.

$$\begin{aligned}\mathbb{E}_k[f(x^{k+1}) - f^*] \leq &(1 - \mu\gamma/2)(f(x^k) - f^*) \\ &+ 4L\gamma^2 \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2 .\end{aligned}$$

Next, we perform the recursion

$$\begin{aligned}\mathbb{E}[f(x^N) - f^*] \leq &(1 - \mu\gamma/2)^N(f(x^k) - f^*) \\ &+ 8L\mu^{-1}\gamma \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2 \\ \leq &\exp(-\mu\gamma N/2)(f(x^0) - f^*) \\ &+ 8L\mu^{-1}\gamma \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2 .\end{aligned}$$

It remains to substitute γ from (40), $B = \lceil b\log_2 M\rceil$ and $b = \tau$. □

B.8 Proof of Theorem 4

We preface the proof by technical Lemma.

Lemma 10. *Let r be μ_r-strongly convex and $x^+ = prox_{\gamma r}(x)$. Then for all $u \in \mathcal{X}$ the following iniqulity hold:*

$$\langle x^+ - x, u - x^+\rangle \geq \gamma\left(r(x^+) - r(u) + \frac{\mu_r}{2}\|x^+ - u\|^2\right).$$

Proof. The optimality condition for $x^+ = \mathrm{prox}_{\gamma r}(x) = \arg\min_{y\in\mathcal{X}}(\gamma r(y) + \frac{1}{2}\|x^+ - y\|^2)$ gives that $(x - x^+) \in \partial r(x^+)$. Therefore, using strong convexity (see A 6) for $r'(x^+) = (x - x^+) \in \partial r(x^+)$, we get

$$\gamma(r(u) - r(x^+)) \geq \langle x - x^+, u - x^+\rangle + \frac{\gamma\mu_r}{2}\|x^+ - u\|^2.$$

After small rearrangements we have what we need to prove. □

Theorem 10 (Theorem 4)**.** *Assume A 5, A 6 with $\mu_F + \mu_r > 0$, A 3, A 7. Let problem* (9) *be solved by Algorithm 3. Then for any $b \in \mathbb{N}^*$, and γ, M satisfying*

$$\gamma \leq \min\{(3\mu_F + 3\mu_r)^{-1}; (3L)^{-1}; (6\mu_F + \mu_r)\cdot[120(C_1\tau b^{-1} + (C_1+1)\tau^2 b^{-2})\Delta^2]^{-1}; \sqrt{(18C_1)^{-1}\Delta^{-2}\tau^{-1}b}\},$$
$$M = \max\{2; \sqrt{C_2\gamma^{-1}(\mu_F + \mu_r)^{-1}}\}, \quad B = \lceil b\log_2 M\rceil,$$

it holds that

$$\mathbb{E}\left[\|x^N - x^*\|^2\right] \lesssim \exp\left(-\frac{(\mu_F + \mu_r)\gamma N}{16}\right)\|x^0 - x^*\|^2 + \frac{\gamma}{\mu}(\sigma^2\tau b^{-1} + \sigma^2\tau^2 b^{-2}).$$

Proof. We start from Lemma 10 for $x^{k+1} = \text{prox}_{\gamma r}\left(x^k - \gamma g^k\right)$ with $x^+ = x^{k+1}$, $x = x^k - \gamma g^k$, $u = x^*$ and for $x^{k+1/2} = \text{prox}_{\gamma r}\left(x^k - \gamma B^{-1}\sum_{i=1}^B F(x^k, z_i^k)\right)$ with $x^+ = x^{k+1/2}$, $x = x^k - \gamma B^{-1}\sum_{i=1}^B F(x^k, z_i^k)$, $u = x^{k+1}$:

$$\langle x^{k+1} - x^k + \gamma g^k, x^* - x^{k+1}\rangle \geq \gamma\left(r(x^{k+1}) - r(x^*) + \frac{\mu_r}{2}\|x^{k+1} - x^*\|^2\right),$$

and

$$\begin{aligned}&\langle x^{k+1/2} - x^k + \gamma B^{-1}\sum_{i=1}^B F(x^k, z_i^k), x^{k+1} - x^{k+1/2}\rangle \\ &\qquad\qquad \geq \gamma\left(r(x^{k+1/2}) - r(x^{k+1}) + \frac{\mu_r}{2}\|x^{k+1} - x^{k+1/2}\|^2\right).\end{aligned}$$

Summing up these two inequalities, we get

$$\begin{aligned}&\langle x^{k+1} - x^k + \gamma g^k, x^* - x^{k+1}\rangle + \langle x^{k+1/2} - x^k + \gamma F(x^k, z^k), x^{k+1} - x^{k+1/2}\rangle \\ &\qquad \geq \gamma\left(r(x^{k+1/2}) - r(x^*) + \frac{\mu_r}{2}\|x^{k+1} - x^*\|^2 + \frac{\mu_r}{2}\|x^{k+1} - x^{k+1/2}\|^2\right).\end{aligned}$$

After some rearrangements, we have

$$\begin{aligned}&\langle x^{k+1} - x^k, x^* - x^{k+1}\rangle + \langle x^{k+1/2} - x^k, x^{k+1} - x^{k+1/2}\rangle \\ &+ \gamma\langle g^k - B^{-1}\sum_{i=1}^B F(x^k, z_i^k), x^{k+1/2} - x^{k+1}\rangle + \gamma\langle g^k, x^* - x^{k+1/2}\rangle \\ &\qquad \geq \gamma\left(r(x^{k+1/2}) - r(x^*) + \frac{\mu_r}{2}\|x^{k+1} - x^*\|^2 + \frac{\mu_r}{2}\|x^{k+1} - x^{k+1/2}\|^2\right).\end{aligned}$$

With $2\langle a, b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$, we deduce

$$\begin{aligned}&\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2 - \|x^{k+1} - x^k\|^2 \\ &+ \|x^{k+1} - x^k\|^2 - \|x^{k+1/2} - x^k\|^2 - \|x^{k+1} - x^{k+1/2}\|^2 \\ &+ 2\gamma\langle g^k - B^{-1}\sum_{i=1}^B F(x^k, z_i^k), x^{k+1/2} - x^{k+1}\rangle + 2\gamma\langle g^k, x^* - x^{k+1/2}\rangle \\ &\qquad \geq 2\gamma\left(r(x^{k+1/2}) - r(x^*) + \frac{\mu_r}{2}\|x^{k+1} - x^*\|^2 + \frac{\mu_r}{2}\|x^{k+1} - x^{k+1/2}\|^2\right).\end{aligned}$$

After rewriting in a slightly different way,

$$\begin{aligned}\|x^{k+1} - x^*\|^2 + \|x^{k+1/2} - x^{k+1}\|^2 \leq &\|x^k - x^*\|^2 - 2\gamma\langle g^k, x^{k+1/2} - x^*\rangle \\ &- 2\gamma\langle B^{-1}\sum_{i=1}^B F(x^k, z_i^k) - g^k, x^{k+1/2} - x^{k+1}\rangle \\ &- \|x^{k+1/2} - x^k\|^2 - 2\gamma(r(x^{k+1/2}) - r(x^*)) \\ &- \mu_r\gamma\|x^{k+1} - x^*\|^2 - \mu_r\gamma\|x^{k+1} - x^{k+1/2}\|^2 \\ \leq &\|x^k - x^*\|^2 - 2\gamma\langle g^k, x^{k+1/2} - x^*\rangle \\ &+ \gamma^2\left\|B^{-1}\sum_{i=1}^B F(x^k, z_i^k) - g^k\right\|^2 + \|x^{k+1/2} - x^{k+1}\|^2 \\ &- \|x^{k+1/2} - x^k\|^2 - 2\gamma(r(x^{k+1/2}) - r(x^*)) \\ &- \mu_r\gamma\|x^{k+1} - x^*\|^2 - \mu_r\gamma\|x^{k+1} - x^{k+1/2}\|^2.\end{aligned}$$

In the last step, we used Cauchy-Schwartz inequality (43). Subtracting $\|x^{k+1} - x^{k+1/2}\|^2$ from both parts, we get

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - 2\gamma\langle g^k, x^{k+1/2} - x^*\rangle + \gamma^2\|B^{-1}\sum_{i=1}^B F(x^k, z_i^k) - g^k\|^2$$

$$
\begin{aligned}
&- \|x^{k+1/2} - x^k\|^2 - 2\gamma(r(x^{k+1/2}) - r(x^*)) \\
&- \mu_r\gamma\|x^{k+1} - x^*\|^2 - \mu_r\gamma\|x^{k+1} - x^{k+1/2}\|^2 \\
=&\|x^k - x^*\|^2 - 2\gamma\langle F(x^{k+1/2}), x^{k+1/2} - x^*\rangle \\
&- 2\gamma\langle \mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2}), x^{k+1/2} - x^*\rangle \\
&- 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^*\rangle \\
&+ \gamma^2\|F(x^k) - F(x^{k+1/2}) + F(x^k) - B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) + F(x^{k+1/2}) - g^k\|^2 \\
&- \|x^{k+1/2} - x^k\|^2 - 2\gamma(r(x^{k+1/2}) - r(x^*)) \\
&- \mu_r\gamma\|x^{k+1} - x^*\|^2 - \mu_r\gamma\|x^{k+1} - x^{k+1/2}\|^2.
\end{aligned}
$$

Again with Cauchy-Schwartz inequality (45), we conduct

$$
\begin{aligned}
\|x^{k+1} - x^*\|^2 \leq &\|x^k - x^*\|^2 - 2\gamma\langle F(x^{k+1/2}), x^{k+1/2} - x^*\rangle \\
&- 2\gamma\langle \mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2}), x^{k+1/2} - x^*\rangle \\
&- 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^*\rangle + 3\gamma^2\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 \\
&+ 3\gamma^2\|F(x^{k+1/2}) - g^k\|^2 + 3\gamma^2\|F(x^{k+1/2}) - F(x^k)\|^2 - \|x^{k+1/2} - x^k\|^2 \\
&- 2\gamma(r(x^{k+1/2}) - r(x^*)) - \mu_r\gamma\|x^{k+1} - x^*\|^2 - \mu_r\gamma\|x^{k+1} - x^{k+1/2}\|^2. \quad (41)
\end{aligned}
$$

A 5 and the property of the solution (9): $-(r(x^{k+1/2}) - r(x^*)) \leq \langle F(x^*), x^{k+1/2} - x^*\rangle$, together give

$$
\begin{aligned}
\|x^{k+1} - x^*\|^2 \leq &\|x^k - x^*\|^2 - 2\gamma\langle F(x^{k+1/2}) - F(x^*), x^{k+1/2} - x^*\rangle \\
&- 2\gamma\langle \mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2}), x^{k+1/2} - x^*\rangle \\
&- 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^*\rangle \\
&+ 3\gamma^2\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 + 3\gamma^2\|F(x^{k+1/2}) - g^k\|^2 \\
&+ 3\gamma^2 L^2\|x^{k+1/2} - x^k\|^2 - \|x^{k+1/2} - x^k\|^2 \\
&- \mu_r\gamma\|x^{k+1} - x^*\|^2 - \mu_r\gamma\|x^{k+1} - x^{k+1/2}\|^2.
\end{aligned}
$$

Next, one can apply A 6 and have

$$
\begin{aligned}
\|x^{k+1} - x^*\|^2 \leq &\|x^k - x^*\|^2 - 2\mu_F\gamma\|x^{k+1/2} - x^*\|^2 \\
&- 2\gamma\langle \mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2}), x^{k+1/2} - x^*\rangle \\
&- 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^*\rangle + 3\gamma^2\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 \\
&+ 3\gamma^2\|F(x^{k+1/2}) - g^k\|^2 + 3\gamma^2 L^2\|x^{k+1/2} - x^k\|^2 - \|x^{k+1/2} - x^k\|^2 \\
&- \mu_r\gamma\|x^{k+1} - x^*\|^2 - \mu_r\gamma\|x^{k+1} - x^{k+1/2}\|^2.
\end{aligned}
$$

Using Cauchy-Schwartz inequality (43) one more time, we get

$$
\begin{aligned}
\|x^{k+1} - x^*\|^2 \leq &\|x^k - x^*\|^2 - 2\mu_F\gamma\|x^{k+1/2} - x^*\|^2 \\
&- \mu_r\gamma\|x^{k+1} - x^*\|^2 - \mu_r\gamma\|x^{k+1} - x^{k+1/2}\|^2 \\
&+ \frac{4\gamma}{\mu_F + \mu_r}\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2 + \frac{(\mu_F + \mu_r)\gamma}{4}\|x^{k+1/2} - x^*\|^2 \\
&- 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^*\rangle + 3\gamma^2\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+ 3\gamma^2\|F(x^{k+1/2}) - g^k\|^2 + 3\gamma^2 L^2\|x^{k+1/2} - x^k\|^2 - \|x^{k+1/2} - x^k\|^2 \\
\le &\|x^k - x^*\|^2 - \frac{7\mu_F\gamma}{4}\|x^{k+1/2} - x^*\|^2 \\
&- \mu_r\gamma\|x^{k+1} - x^*\|^2 - \mu_r\gamma\|x^{k+1} - x^{k+1/2}\|^2 \\
&+ \frac{\mu_r\gamma}{4}\|x^{k+1/2} - x^*\|^2 + \frac{4\gamma}{\mu_F + \mu_r}\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2 \\
&- 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^*\rangle + 3\gamma^2\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 \\
&+ 3\gamma^2\|F(x^{k+1/2}) - g^k\|^2 + 3\gamma^2 L^2\|x^{k+1/2} - x^k\|^2 - \|x^{k+1/2} - x^k\|^2.
\end{aligned}
$$

With Cauchy-Schwartz inequality in the form: $-\mu_r\gamma\|x^{k+1} - x^*\|^2 \le -\frac{\mu_r\gamma}{2}\|x^{k+1/2} - x^*\|^2 + \mu_r\gamma\|x^{k+1} - x^{k+1/2}\|^2$, one can deduce

$$
\begin{aligned}
\|x^{k+1} - x^*\|^2 \le &\|x^k - x^*\|^2 - \frac{(7\mu_F + \mu_r)\gamma}{4}\|x^{k+1/2} - x^*\|^2 \\
&+ \frac{4\gamma}{\mu_F + \mu_r}\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2 \\
&- 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^*\rangle + 3\gamma^2\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 \\
&+ 3\gamma^2\|F(x^{k+1/2}) - g^k\|^2 + 3\gamma^2 L^2\|x^{k+1/2} - x^k\|^2 - \|x^{k+1/2} - x^k\|^2.
\end{aligned}
$$

Taking the expectation and using Lemma 3, Lemma 4 (with $\Delta^2\|x - x^*\|^2$ instead of $\delta^2\|\nabla f(x)\|^2$), we have

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \le &\mathbb{E}\left[\|x^k - x^*\|^2\right] - \frac{(7\mu_F + \mu_r)\gamma}{4}\mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right] \\
&+ \frac{4\gamma}{\mu_F + \mu_r}\mathbb{E}\left[\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2\right] \\
&+ 3\gamma^2\mathbb{E}\left[\mathbb{E}_k\left[\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2\right]\right] \\
&+ 3\gamma^2\mathbb{E}\left[\mathbb{E}_{k+1/2}\left[\|F(x^{k+1/2}) - g^k\|^2\right]\right] \\
&+ 3\gamma^2 L^2\mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] - \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] \\
\le &\mathbb{E}\left[\|x^k - x^*\|^2\right] - \frac{(7\mu_F + \mu_r)\gamma}{4}\mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right] \\
&+ \frac{4\gamma}{\mu_F + \mu_r}\cdot C_2\tau^2 M^{-2}B^{-2}\left(\sigma^2 + \Delta^2\mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right]\right) \\
&+ 3\gamma^2\cdot C_1\tau B^{-1}\left(\sigma^2 + \Delta^2\mathbb{E}\left[\|x^k - x^*\|^2\right]\right) \\
&+ 3\gamma^2\cdot\left(4C_1\tau B^{-1}\log_2 M + (4C_1 + 2)\tau^2 B^{-2}\right)\left(\sigma^2 + \Delta^2\|x^{k+1/2} - x^*\|^2\right) \\
&+ 3\gamma^2 L^2\mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] - \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right].
\end{aligned}
$$

With $M \ge \sqrt{C_2\gamma^{-1}(\mu_F + \mu_r)^{-1}}$, we have

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \le &\mathbb{E}\left[\|x^k - x^*\|^2\right] - \frac{(7\mu_F + \mu_r)\gamma}{4}\mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right] \\
&+ 4\gamma^2\cdot\tau^2 B^{-2}\left(\sigma^2 + \Delta^2\mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right]\right) \\
&+ 3\gamma^2\cdot C_1\tau B^{-1}\left(\sigma^2 + \Delta^2\mathbb{E}\left[\|x^k - x^*\|^2\right]\right) \\
&+ 3\gamma^2\cdot\left(4C_1\tau B^{-1}\log_2 M + (4C_1 + 2)\tau^2 B^{-2}\right)\left(\sigma^2 + \Delta^2\mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
&+ 3\gamma^2 L^2 \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] - \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] \\
\leq& \mathbb{E}\left[\|x^k - x^*\|^2\right] - \frac{(7\mu_F + \mu_r)\gamma}{4}\mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right] \\
&+ 3\gamma^2 \cdot C_1 \tau B^{-1}\Delta^2 \mathbb{E}\left[\|x^k - x^*\|^2\right] \\
&+ 12\gamma^2 \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\Delta^2 \mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right] \\
&+ 15\gamma^2 \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2 \\
&+ 3\gamma^2 L^2 \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] - \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right].
\end{aligned}
$$

Cauchy-Schwartz inequality (44) gives

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq& \mathbb{E}\left[\|x^k - x^*\|^2\right] - \frac{(7\mu_F + \mu_r)\gamma}{4}\mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right] \\
&+ 15\gamma^2 \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\Delta^2 \mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right] \\
&+ 15\gamma^2 \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2 \\
&+ 6\gamma^2 \cdot C_1\tau B^{-1}\Delta^2 \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] \\
&+ 3\gamma^2 L^2 \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] - \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right].
\end{aligned}
$$

Since $\gamma \leq (7\mu_F + \mu_r) \cdot \left[120\left(C_1\tau b^{-1} + (C_1+1)\tau^2 b^{-2}\right)\Delta^2\right]^{-1}$, $B = \lceil b\log_2 M\rceil$ and $M \geq 2$, one can obtain

$$
\begin{aligned}
\gamma &\leq (7\mu_F + \mu_r) \cdot \left[120\left(C_1\tau b^{-1} + (C_1+1)\tau^2 b^{-2}\right)\Delta^2\right]^{-1} \\
&\leq (7\mu_F + \mu_r) \cdot \left[120\left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\Delta^2\right]^{-1},
\end{aligned}
$$

and then,

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq& \mathbb{E}\left[\|x^k - x^*\|^2\right] - \frac{(7\mu_F + \mu_r)\gamma}{8}\mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right] \\
&+ 15\gamma^2 \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2 \\
&+ 6\gamma^2 \cdot C_1\tau B^{-1}\Delta^2 \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] \\
&+ 3\gamma^2 L^2 \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] - \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right].
\end{aligned}
$$

With Cauchy-Schwartz inequality in the form: $-\|x^{k+1/2} - x^*\|^2 \leq -\frac{1}{2}\|x^k - x^*\|^2 + \|x^k - x^{k+1/2}\|^2$, we have

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq& \left(1 - \frac{(7\mu_F + \mu_r)\gamma}{16}\right)\mathbb{E}\left[\|x^k - x^*\|^2\right] \\
&- \left(1 - (\mu_F + \mu_r)\gamma - 3\gamma^2 L^2 - 6\gamma^2 \cdot C_1\tau B^{-1}\Delta^2\right)\mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right] \\
&+ 15\gamma^2 \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2.
\end{aligned}
$$

Since $\gamma \leq \min\left\{(3\mu_F + 3\mu_r)^{-1}; (3L)^{-1}; \sqrt{(18C_1)^{-1}\tau^{-1}b\Delta^{-2}}\right\}$, we get

$$
\begin{aligned}
\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq& \left(1 - \frac{(7\mu_F + \mu_r)\gamma}{16}\right)\mathbb{E}\left[\|x^k - x^*\|^2\right] \\
&+ 15\gamma^2 \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2.
\end{aligned}
$$

Next, we perform the recursion

$$
\begin{aligned}
\mathbb{E}\left[\|x^N - x^*\|^2\right] \leq& \left(1 - \frac{(7\mu_F + \mu_r)\gamma}{16}\right)^N \|x^0 - x^*\|^2 \\
&+ \frac{240\gamma}{(\mu_F + \mu_r)} \cdot \left(C_1\tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2
\end{aligned}
$$

$$
\begin{aligned}
&\leq \exp\left(-\frac{(\mu_F+\mu_r)\gamma N}{16}\right)\|x^0 - x^*\|^2 \\
&\quad + \frac{240\gamma}{(\mu_F+\mu_r)} \cdot \left(C_1 \tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}\right)\sigma^2.
\end{aligned}
$$

Substituting $B = \lceil b \log_2 M \rceil$ and using $M \geq 2$ finish the proof. $\square$

B.9 Proof of Theorem 5

Theorem 11 (Theorem 5)**.** *Assume A 5, A 6 with $\mu_F + \mu_r = 0$, A 8, A 3, A 7. Let problem* (9) *be solved by Algorithm 3. Then for any $B \in \mathbb{N}^*$, and γ, M satisfying $\gamma \lesssim L^{-1}$, $M = \sqrt{N}$, it holds that*

$$
\mathbb{E}\left[\mathit{Gap}(\bar{x}^N)\right] \lesssim \frac{D^2}{\gamma N} + \gamma(\tau B^{-1}\log_2 N + \tau^2 B^{-2})(\sigma^2 + \Delta^2 D^2)\,,
$$

where $\bar{x}^N = \frac{1}{N}\sum\limits_{k=0}^{N-1} x^{k+1/2}$.

Proof. We start from (41) with arbitrary $x \in \mathcal{X}$ instead of x^* and $\mu_r = 0$:

$$
\begin{aligned}
\|x^{k+1} - x\|^2 \leq &\|x^k - x\|^2 - 2\gamma\langle F(x^{k+1/2}), x^{k+1/2} - x\rangle \\
&- 2\gamma\langle \mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2}), x^{k+1/2} - x\rangle \\
&- 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x\rangle + 3\gamma^2\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 \\
&+ 3\gamma^2\|F(x^{k+1/2}) - g^k\|^2 + 3\gamma^2\|F(x^{k+1/2}) - F(x^k)\|^2 - \|x^{k+1/2} - x^k\|^2 \\
&- 2\gamma(r(x^{k+1/2}) - r(x)).
\end{aligned}
$$

After small rearrangements, we get

$$
\begin{aligned}
&2\gamma\big(\langle F(x^{k+1/2}), x^{k+1/2} - x\rangle + r(x^{k+1/2}) - r(x)\big) \\
&\qquad\leq \|x^k - x\|^2 - \|x^{k+1} - x\|^2 - 2\gamma\langle \mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2}), x^{k+1/2} - x\rangle \\
&\qquad\quad - 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x\rangle + 3\gamma^2\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 \\
&\qquad\quad + 3\gamma^2\|F(x^{k+1/2}) - g^k\|^2 + 3\gamma^2\|F(x^{k+1/2}) - F(x^k)\|^2 - \|x^{k+1/2} - x^k\|^2.
\end{aligned}
$$

Applying Cauchy-Schwartz inequality and making more rearrangements, we get

$$
\begin{aligned}
&2\gamma\big(\langle F(x^{k+1/2}), x^{k+1/2} - x\rangle + r(x^{k+1/2}) - r(x)\big) \\
&\qquad\leq \|x^k - x\|^2 - \|x^{k+1} - x\|^2 + \gamma^2 N\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2 + \frac{1}{N}\|x^{k+1/2} - x\|^2 \\
&\qquad\quad - 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^0\rangle - 2\gamma\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^0 - x\rangle \\
&\qquad\quad + 3\gamma^2\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 + 3\gamma^2\|F(x^{k+1/2}) - g^k\|^2 \\
&\qquad\quad + 3\gamma^2\|F(x^{k+1/2}) - F(x^k)\|^2 - \|x^{k+1/2} - x^k\|^2.
\end{aligned}
$$

Summing over all k from 0 to $N-1$ and dividing by N, we have

$$
\begin{aligned}
&2\gamma \cdot \frac{1}{N}\sum_{k=0}^{N-1}\big(\langle F(x^{k+1/2}), x^{k+1/2} - x\rangle + r(x^{k+1/2}) - r(x)\big) \\
&\leq \frac{\|x^0 - x\|^2 - \|x^K - x\|^2}{N} + \gamma^2\sum_{k=0}^{N-1}\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2 + \frac{1}{N^2}\sum_{k=0}^{N-1}\|x^{k+1/2} - x\|^2
\end{aligned}
$$

$$
\begin{aligned}
&- 2\gamma \cdot \frac{1}{N}\sum_{k=0}^{N-1} \langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^0\rangle - 2\gamma \langle N^{-1}\sum_{k=0}^{N-1}\left[g^k - \mathbb{E}_{k+1/2}[g^k]\right], x^0 - x\rangle \\
&+ 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1} \|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 + 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\|F(x^{k+1/2}) - g^k\|^2 \\
&+ 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\|F(x^{k+1/2}) - F(x^k)\|^2 - \frac{1}{N}\sum_{k=0}^{N-1}\|x^{k+1/2} - x^k\|^2.
\end{aligned}
$$

Using monotonicity and Jensen's inequality (46) for convex function r, we get (with notation $\bar{x}^N = \frac{1}{N}\sum\limits_{k=0}^{N-1} x^{k+1/2}$)

$$
\begin{aligned}
&2\gamma\big(\langle F(x), \bar{x}^N - x\rangle + r(\bar{x}^N) - r(x)\big) \\
&\quad \le \frac{\|x^0 - x\|^2 - \|x^N - x\|^2}{N} + \gamma^2\sum_{k=0}^{N-1}\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2 + \frac{1}{N^2}\sum_{k=0}^{N-1}\|x^{k+1/2} - x\|^2 \\
&\quad - 2\gamma \cdot \frac{1}{N}\sum_{k=0}^{N-1} \langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^0\rangle - 2\gamma \langle N^{-1}\sum_{k=0}^{N-1}\left[g^k - \mathbb{E}_{k+1/2}[g^k]\right], x^0 - x\rangle \\
&\quad + 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1} \|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 + 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\|F(x^{k+1/2}) - g^k\|^2 \\
&\quad + 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\|F(x^{k+1/2}) - F(x^k)\|^2 - \frac{1}{N}\sum_{k=0}^{N-1}\|x^{k+1/2} - x^k\|^2.
\end{aligned}
$$

Applying Cauchy-Schwartz inequality (43) one more time,

$$
\begin{aligned}
&2\gamma\big(\langle F(x), \bar{x}^N - x\rangle + r(\bar{x}^N) - r(x)\big) \\
&\quad \le \frac{2\|x^0 - x\|^2}{N} + \gamma^2\sum_{k=0}^{N-1}\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2 + \frac{1}{N^2}\sum_{k=0}^{N-1}\|x^{k+1/2} - x\|^2 \\
&\quad - 2\gamma \cdot \frac{1}{N}\sum_{k=0}^{N-1} \langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^0\rangle + \gamma^2 \|N^{-1}\sum_{k=0}^{N-1}\left[g^k - \mathbb{E}_{k+1/2}[g^k]\right]\|^2 \\
&\quad + 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1} \|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2 + 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\|F(x^{k+1/2}) - g^k\|^2 \\
&\quad + 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\|F(x^{k+1/2}) - F(x^k)\|^2 - \frac{1}{N}\sum_{k=0}^{N-1}\|x^{k+1/2} - x^k\|^2.
\end{aligned}
$$

Taking supermom on x from $\mathcal{X}$ and then the full expectation, we get

$$
\begin{aligned}
2\gamma\mathbb{E}\left[\text{Gap}(\bar{x}^N)\right] \le & \frac{2\max_{x\in\mathcal{X}}\|x^0 - x\|^2}{N} + \frac{1}{N^2}\sum_{k=0}^{N-1}\mathbb{E}\left[\max_{x\in\mathcal{X}}\|x^{k+1/2} - x\|^2\right] \\
& - 2\gamma \cdot \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^0\rangle\right] \\
& + \gamma^2\sum_{k=0}^{N-1}\mathbb{E}\left[\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2\right] \\
& + \gamma^2\mathbb{E}\left[\|N^{-1}\sum_{k=0}^{N-1}\left[g^k - \mathbb{E}_{k+1/2}[g^k]\right]\|^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\|B^{-1}\sum_{i=1}^{B}F(x^k,z_i^k) - F(x^k)\|^2\right] \\
&+ 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\|F(x^{k+1/2}) - g^k\|^2\right] \\
&+ 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\|F(x^{k+1/2}) - F(x^k)\|^2\right] - \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right].
\end{aligned}
$$

One can note that

$$
\begin{aligned}
\mathbb{E}\left[\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^0\rangle\right] &= \mathbb{E}\left[\mathbb{E}_{k+1/2}[\langle g^k - \mathbb{E}_{k+1/2}[g^k], x^{k+1/2} - x^0\rangle]\right] \\
&= \mathbb{E}\left[\langle \mathbb{E}_{k+1/2}[g^k - \mathbb{E}_{k+1/2}[g^k]], x^{k+1/2} - x^0\rangle\right] \\
&= 0,
\end{aligned}
$$

and (here we also need Cauchy-Schwartz inequality (44))

$$
\begin{aligned}
\mathbb{E}\left[\|N^{-1}\sum_{k=0}^{N-1}\left[g^k - \mathbb{E}_{k+1/2}[g^k]\right]\|^2\right] =& \frac{1}{N^2}\sum_{k=0}^{N-1}\mathbb{E}\left[\|g^k - \mathbb{E}_{k+1/2}[g^k]\|^2\right] \\
&+ \frac{1}{N^2}\sum_{k\neq j}\mathbb{E}\left[\langle g^k - \mathbb{E}_{k+1/2}[g^k], g^j - \mathbb{E}_{j+1/2}[g^j]\rangle\right] \\
=& \frac{1}{N^2}\sum_{k=0}^{N-1}\mathbb{E}\left[\|g^k - \mathbb{E}_{k+1/2}[g^k]\|^2\right] \\
&+ \frac{2}{N^2}\sum_{k>j}\mathbb{E}\left[\langle \mathbb{E}_{k+1/2}[g^k - \mathbb{E}_{k+1/2}[g^k]], g^j - \mathbb{E}_{j+1/2}[g^j]\rangle\right] \\
=& \frac{1}{N^2}\sum_{k=0}^{N-1}\mathbb{E}\left[\|g^k - \mathbb{E}_{k+1/2}[g^k]\|^2\right] \\
\leq& \frac{2}{N^2}\sum_{k=0}^{N-1}\mathbb{E}\left[\|g^k - F(x^{k+1/2})\|^2\right] \\
&+ \frac{2}{N^2}\sum_{k=0}^{N-1}\mathbb{E}\left[\|F(x^{k+1/2}) - \mathbb{E}_{k+1/2}[g^k]\|^2\right].
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
2\gamma\mathbb{E}\left[\mathrm{Gap}(\bar{x}^N)\right] \leq& \frac{2\max_{x\in\mathcal{X}}\|x^0 - x\|^2}{N} + \frac{1}{N^2}\sum_{k=0}^{N-1}\mathbb{E}\left[\max_{x\in\mathcal{X}}\|x^{k+1/2} - x\|^2\right] \\
&+ 2\gamma^2\sum_{k=0}^{N-1}\mathbb{E}\left[\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2\right] \\
&+ 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\|B^{-1}\sum_{i=1}^{B}F(x^k,z_i^k) - F(x^k)\|^2\right] \\
&+ 5\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\|F(x^{k+1/2}) - g^k\|^2\right] \\
&+ 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\|F(x^{k+1/2}) - F(x^k)\|^2\right] - \frac{1}{N}\sum_{k=0}^{N-1}\mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right].
\end{aligned}
$$

With A 6 and A 8, we obtain

$$
2\gamma\mathbb{E}\left[\mathrm{Gap}(\bar{x}^N)\right] \leq \frac{3D^2}{N} + 2\gamma^2\sum_{k=0}^{N-1}\mathbb{E}\left[\|\mathbb{E}_{k+1/2}[g^k] - F(x^{k+1/2})\|^2\right]
$$

$$+ 3\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}\left[\|B^{-1}\sum_{i=1}^{B} F(x^k, z_i^k) - F(x^k)\|^2\right]$$
$$+ 5\gamma^2 \cdot \frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}\left[\|F(x^{k+1/2}) - g^k\|^2\right]$$
$$- (1 - 3\gamma^2 L^2)\frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right].$$

Using Lemma 3 and Lemma 4, we have

$$\begin{aligned}
2\gamma\mathbb{E}\left[\text{Gap}(\bar{x}^N)\right] \leq& \frac{3D^2}{N} + 2\gamma^2 C_2 \tau^2 M^{-2} B^{-2} \sum_{k=0}^{N-1}\left(\sigma^2 + \Delta^2 \mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right]\right) \\
&+ 3\gamma^2 C_1 \tau B^{-1} \cdot \frac{1}{N}\sum_{k=0}^{N-1}\left(\sigma^2 + \Delta^2 \mathbb{E}\left[\|x^k - x^*\|^2\right]\right) \\
&+ 20\gamma^2 (C_1 \tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}) \cdot \frac{1}{N}\sum_{k=0}^{N-1}\left(\sigma^2 + \Delta^2 \mathbb{E}\left[\|x^{k+1/2} - x^*\|^2\right]\right) \\
&- (1 - 3\gamma^2 L^2)\frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right].
\end{aligned}$$

Again with A 8, we get

$$\begin{aligned}
2\gamma\mathbb{E}\left[\text{Gap}(\bar{x}^N)\right] \leq& \frac{3D^2}{N} + 2\gamma^2 C_2 \tau^2 M^{-2} B^{-2} N\left(\sigma^2 + \Delta^2 D^2\right) \\
&+ 3\gamma^2 C_1 \tau B^{-1}\left(\sigma^2 + \Delta^2 D^2\right) \\
&+ 20\gamma^2 (C_1 \tau B^{-1}\log_2 M + (C_1+1)\tau^2 B^{-2}) \cdot \left(\sigma^2 + \Delta^2 D^2\right) \\
&- (1 - 3\gamma^2 L^2)\frac{1}{N}\sum_{k=0}^{N-1} \mathbb{E}\left[\|x^{k+1/2} - x^k\|^2\right].
\end{aligned}$$

With $M = \sqrt{N}$ and $\gamma \leq (3L)^{-1}$, one can deduce

$$2\gamma\mathbb{E}\left[\text{Gap}(\bar{x}^N)\right] \leq \frac{3D^2}{N} + 25\gamma^2 (C_1 \tau B^{-1}\log_2 M + (C_1 + C_2 + 1)\tau^2 B^{-2}) \cdot \left(\sigma^2 + \Delta^2 D^2\right).$$

Substituting $M = \sqrt{N}$ finishes the proof. □

C Basic Facts

Lemma 11 (see Lemma 1.2.3 and Theorem 2.1.5 from [76])**.** *If f is L-smooth in $\mathbb{R}^d$, then for any $x, y \in \mathbb{R}^d$*

$$f(x) - f(y) - \langle \nabla f(y), x - y\rangle \leq \frac{L}{2}\|x - y\|^2. \tag{42}$$

Lemma 12 (Cauchy Schwartz inequality)**.** *For any $a, b, x_1, \ldots, x_n \in \mathbb{R}^d$ and $c > 0$ the following inequalities hold:*

$$2\langle a, b\rangle \leq \frac{\|a\|^2}{c} + c\|b\|^2, \tag{43}$$

$$\|a + b\|^2 \leq \left(1 + \frac{1}{c}\right)\|a\|^2 + (1 + c)\|b\|^2, \tag{44}$$

$$\left\|\sum_{i=1}^{n} x_i\right\|^2 \leq n \cdot \sum_{i=1}^{n}\|x_i\|^2. \tag{45}$$

Lemma 13 (Jensen's inequality)**.** *If f is a convex function, then for any $n \in \mathbb{N}^*$ and $x_1, \ldots, x_n \in \mathbb{R}^d$ the following inequality holds:*

$$f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \leq \frac{1}{n}\sum_{i=1}^{n} f(x_i). \tag{46}$$