METHODS FOR SOLVING VARIATIONAL INEQUALITIES WITH MARKOVIAN STOCHASTICITY

Vladimir Solodkin
Moscow Institute of Physics and Technology
Ivannikov Institute for System Programming of RAS

Michael Ermoshin
Moscow Institute of Physics and Technology

Roman Gavrilenko
Moscow Institute of Physics and Technology

Aleksandr Beznosikov
Moscow Institute of Physics and Technology
Ivannikov Institute for System Programming RAS
Innopolis University

ABSTRACT

In this paper, we present a novel stochastic method for solving variational inequalities (VI) in the context of Markovian noise. By leveraging Extragradient technique, we can productively solve VI optimization problems characterized by Markovian dynamics. We demonstrate the efficacy of proposed method through rigorous theoretical analysis, proving convergence under quite mild assumptions of L-Lipschitzness, strong monotonicity of the operator and boundness of the noise only at the optimum. In order to gain further insight into the nature of Markov processes, we conduct the experiments to investigate the impact of the mixing time parameter on the convergence of the algorithm.

1 INTRODUCTION

Stochastic gradient methods are crucial for solving a wide range of optimization problems, with various applications in machine learning Goodfellow et al. (2014; 2016), including areas ranging from traditional empirical risk minimization Vapnik (1998) to modern reinforcement learning Tomar et al. (2021); Schulman et al. (2015); Lan (2022). The majority of minimization problems in machine learning have a stochastic structure, and are typically addressed through SGD-like approaches Cotter et al. (2011); Vaswani et al. (2019); Taylor and Bach (2019); Aybat et al. (2019); Gorbunov et al. (2019); Woodworth and Srebro (2021). While such methods have been the focus of considerable research, most of the results surrounding the analysis of these methods rely heavily on the standard assumption of noise independence. However, this assumption become unrealistic in many practical scenarios. For instance, in distributed optimization, dependencies arise due to the communication delays and synchronization problems Lopes and Sayed (2007); Dimakis et al. (2010); Mao et al. (2020). Similarly, in reinforcement learning Bhandari et al. (2018); Srikant and Ying (2019); Durmus et al. (2021), data is generated by interacting with an environment, leading to highly-correlated dependencies. These observations underscore the necessity for the development of novel methodologies that can operate effectively in the presence of non-i.i.d. stochasticity.

At present, the research on stochastic methods with Markovian noise in minimization problems is less developed in comparison to research on methods with independent noise. This inconsistency can be attributed to several factors. Notably, methods incorporating Markovian noise often present more intricate mathematical challenges Duchi et al. (2012); Beznosikov et al. (2023b); Even (2023), as the dependencies between observations complicate the proof of fundamental algorithmic properties. In contrast, methods assuming independent noise generally have a more straightforward structure, facilitating their theoretical analysis (see Gorbunov et al. (2019) and references therein). Nevertheless, the investigation of Markovian noise in these methods is a rapidly evolving field, with a growing body of literature addressing the challenges associated with minimizing objective functions under such conditions Wang et al. (2022); Karimi et al. (2019); Doan (2023).

The majority of previous works utilizing Markovian stochastics have focused on the minimization problem. However, from a practical point of view, more generalized frameworks like variational inequalities (VI) are also of significant interest. VIs provide a unified formulation for a broad class of problems, encapsulating a wide range of applications from game theory to traffic flow analysis Scutari et al. (2010); Jofre et al. (2007). In particular, examples include economic market modeling, in which firms engage in competition and the objective is to identify a state of market balance; network routing, in which optimal paths are determined in the presence of competing flows; and auction theory, in which bidders strategies must converge to an equilibrium. Furthermore, solving saddle-point problems, which constitute a significant subset of variational inequality challenges, is a crucial aspect of training Generative Adversarial Networks (GANs) Gidel et al. (2018); Mertikopoulos et al. (2018); Chavdarova et al. (2019). These problems often involve optimizing a min-max objective function, wherein the generator and discriminator assume adversarial roles. The theoretical foundation provided by VI frameworks is vital for ensuring convergence and stability in such adversarial settings. This approach can be effectively used not only in GANs training Liang and Stokes (2019), but also in other applications where equilibrium conditions and competitive interactions prevail. As such, developing robust stochastic methods for VIs Juditsky et al. (2011); Mishchenko et al. (2020), especially in the presence of dependent noise, is an active area of research. These methods aim to extend the robustness and applicability of classical techniques to more complex and realistic settings, where noise dependencies are inherent and unavoidable.

In light of these insights, it becomes evident that there is a pressing need to develop a more practically applicable method. The objective of this paper is to incorporate the Markovian dynamics into the Extragradient algorithm Korpelevich (1977), thereby eliminating the limitations of traditional assumptions and expanding the scope of practical applications.

1.1 Related Work

Deterministic methods for solving VIs with a Lipschitz operator made significant progress with the advent of Extragradient method Korpelevich (1977). This classic method involves a two-step iterative process: first, an intermediate point is calculated using the current gradient operator, and then the actual update is performed by using operator at the intermediate point. This additional step is a meaningful improvement in the stability of convergence of the method, ultimately ensuring a more reliable optimization process. By introducing this, the method achieves greater robustness and improved performance in finding solutions. One of a more sophisticated modification of Extragradient method is Extrapolation From The Past Popov (1980). This single-call method calculates the operator only once per iteration, and the distinction lies in the definition of the intermediate point, which employs the operator in the previous intermediate point, not in the current, leading to halving the gradient calculations. Another significant approach is the Mirror-Prox algorithm Nemirovski (2004), which utilizes Bregman divergence to improve the optimization process. This idea adapts to the underlying geometry of the problem by employing Bregman divergence, permitting the use of non-Euclidean steps. By respecting the curvature of the solution space, it facilitates more efficient and stable convergence in complex, high-dimensional settings.

There were also significant developments in stochastic methods for solving variational inequalities. For instance, Juditsky et al. (2011) explored a stochastic version of the Mirror-Prox method, examining a scenario with constrained noise variance. This work served as the foundation for subsequent development in this field. To avoid the bounded variance assumption, Mishchenko et al. (2020) proposed Revisiting Stochastic Extragradient, another stochastic modification of the Extragradient algorithm with the same randomness. In case of less general settings, e.g. finite sum setup, the classical variance reduction technique is applicable. This is exemplified by the implementation of variance reduction in the context of variational inequalities, as demonstrated in the work Alacaoglu and Malitsky (2022), where variance reduction modifications of both classical Extragradient Korpelevich (1977) and Mirror-Prox Juditsky et al. (2011) were presented. Prior to mentioned work Alacaoglu and Malitsky (2022), efforts were also made to adapt the same technique, as evidenced by Chavdarova et al. (2019); Yang et al. (2020). Nevertheless, the majority of these results are predicated on the assumption of independent noise Palaniappan and Bach (2016); Chavdarova et al. (2019); Beznosikov et al. (2020); Yang et al. (2020); Alacaoglu and Malitsky (2022); Beznosikov et al. (2023a); Pichugin et al. (2023; 2024).

Markovian noise, where the next iteration becomes dependent on the previous one, models real-world scenarios more accurately than i.i.d. noise scenario, capturing temporal dependencies and sequential decision processes. For instance, the work Beznosikov et al. (2023b) provided a comprehensive framework for analyzing first-order gradient methods in stochastic minimization and VIs involving Markovian noise. Their approach achieved optimal linear dependence on the mixing time of the noise sequence in the stochastic term through a randomized batching scheme based on the multilevel Monte Carlo method. This technique eliminated several limiting assumptions from previous research, such as the need for a bounded domain and uniformly bounded stochastic operator. Notably, their extension to VIs under Markovian noise represented a significant contribution, providing matching lower bounds for oracle complexity in VI problems. In another paper Wang et al. (2022), stochastic gradient-based Markov chain methods (MC-SGM) were analyzed for the min-max problem. The authors used algorithmic stability within the framework of statistical learning theory for both smooth and nonsmooth cases. However, the results in this paper are rather sparse and a continuation of this topic is needed.

1.2 Our Contributions

The main contributions of this paper are the following:

• Novel Stochastic Look at VI
We present a novel view on variational inequalities through the lens of Markovian stochasticity. This approach differs from traditional methods such as stochastic Extragradient and its variations, which typically impose restrictions on noise independence Chavdarova et al. (2019); Palaniappan and Bach (2016); Yang et al. (2020); Alacaoglu and Malitsky (2022); Mishchenko et al. (2020). Among the works dealing with the Markovian noise, our paper is distinguishable by a mild assumption on the noise variables bounding the operator only at the optimum. However, we necessitate the Lipschitzness and strong monotonicity of the operator for all realizations of a random variable. On the other hand, analogous assumptions are present in the work Mishchenko et al. (2020), wherein the noise is considered to be independent.

• Rates of Convergence
We provide sharp rates of convergence, being able to avoid the presence of mixing time in the deterministic term of the rate. The stochastic term is quadratic with respect to the mixing time of the underlying Markov chain, which is competitive with the exiting foundations in the literature.

• Experimental Analysis of Mixing Time Influence
In order to determine the actual convergence rate of the specified method and to verify the feasibility of the theoretical assessment in practice, we provide the numerical experiments. The aim of them is to demonstrate the effect of changes in the mixing time parameter on the convergence process.

1.3 Notation

We use $\langle x, y\rangle := \sum_{i=1}^d x_i y_i$ to denote standard inner product of vectors $x, y \in \mathbb{R}^d$. We introduce l_2-norm of vector $x \in \mathbb{R}^d$ as $\|x\| := \sqrt{\langle x, x\rangle}$. Let $(\mathsf{M}, \mathsf{d}_\mathsf{M})$ be a complete separable metric space endowed with its Borel σ-field $\mathcal{M}$. For $z^0, \dots, z^t$ being the iterates of any algorithm, we denote $\mathcal{F}_t = \sigma(z^j, j \leq t)$ and write $\mathbb{E}_t$ to denote conditional expectation $\mathbb{E}[\cdot|\mathcal{F}_t]$.

2 Technical preliminaries

In this paper, we are interested in solving the optimization problem of the form

$$\text{Find } z^* \in \mathbb{R}^d \text{ such that } \forall z \in \mathbb{R}^d \hookrightarrow \langle F(z^*), z - z^*\rangle \geq 0, \tag{1}$$

where $F(z) := \mathbb{E}_{\xi\sim\pi} F(z, \xi)$ is approximated by the stochastic oracle $F(z, \xi)$ and $\{\xi^t\}_{t=0}^\infty$ is a stationary Markov chain with a unique invariant distribution π, defined on $\mathcal{M}$.

The aforementioned formulation of the VI problem is a classical approach in optimization methods.

We begin by introducing two fundamental constraints on the operator $F(\cdot, \xi)$.

Assumption 1 ($L(\xi)$-Lipschitzness)**.** *The operator $F(\cdot, \xi)$ is $L(\xi)$-Lipschitz, i.e., there exists $L(\xi) > 0$ such that the following inequality holds for all $z', z'' \in \mathbb{R}^d$:*

$$\|F(z', \xi) - F(z'', \xi)\| \leq L(\xi)\|z' - z''\|.$$

We also define $L := \sup_{\xi \in \mathcal{M}} L(\xi) < +\infty$.

Assumption 2 ($\mu(\xi)$-strong monotonicity)**.** *The operator* $F(\cdot, \xi)$ *is* $\mu(\xi)$*-strongly monotone, i.e., there exists* $\mu(\xi) > 0$ *such that the following inequality holds for all* $z', z'' \in \mathbb{R}^d$*:*

$$\langle F(z', \xi) - F(z'', \xi), z' - z'' \rangle \geq \mu(\xi) \|z' - z''\|^2.$$

We also define $\mu := \inf_{\xi \in \mathcal{M}} \mu(\xi) > 0$.

The next assumption implies a uniform stochastic boundedness of gradient operator at the optimum point.

Assumption 3. *The oracle* $F(z, \cdot)$ *is bounded at the optimum* z^**, i.e., there exists* $\sigma_* > 0$ *such that the following inequality holds for all* $\xi \in \mathcal{M}$*:*

$$\|F(z^*, \xi)\| \leq \sigma_*.$$

To the best of our knowledge, this constraint seems to be unpopular in the majority of existing literature. Most of the existing work assumes either uniform boundness or boundness of the moments of distribution.

Now we introduce an important assumption related to the theory of Markov processes.

Assumption 4. $\{\xi_t\}_{t=0}^{\infty}$ *is a stationary Markov chain on* $(\mathsf{M}, \mathcal{M})$ *with unique invariant distribution* π*. Moreover,* $\{\xi_t\}_{t=0}^{\infty}$ *is uniformly geometrically ergodic with mixing time* $\tau_{mix}(\varepsilon)$*, i.e.*

$$\tau_{mix}(\varepsilon) := \inf\{t \geq 1 | \forall m_0, m \hookrightarrow |\mathbb{P}\{\xi_t = m | \xi_0 = m_0\} - \pi_m| \leq \varepsilon\}.$$

This kind of assumption is standard concerning the literature on Markovian noise Chavdarova et al. (2019); Palaniappan and Bach (2016); Yang et al. (2020); Alacaoglu and Malitsky (2022). It is quite obvious from definition, that the mixing time $\tau_{mix}(\varepsilon)$ is simply the number of steps of the Markov chain required for the distribution of the current state to be ε-close to the stationary probability π.

3 Main results

Now, we are ready to introduce our Algorithm 1. Following the idea of utilizing the intermediate step information, we incorporate the Markovian stochasticity into the classical `Extrapolated Gradient Method` Korpelevich (1977). An extrapolation step aims at stabilizing the convergence process, while the Markovian approach pursues to utilize the natural idea of using the previous time iterations. The aforestated algorithm outlines the steps involved.

Algorithm 1 `Extrapolated Gradient Method with Markovian noise`

```
Parameters: step size $\gamma > 0$, number of iterations $T$
Initialization: choose $z^0 \in \mathcal{Z}$
for $t = 0$ to $T$ do
    $z^{t+\frac{1}{2}} = z^t - \gamma F(z^t, \xi^t)$
    $z^{t+1} = z^t - \gamma F(z^{t+\frac{1}{2}}, \xi^t)$
end for
```

In order to prove the main convergence theorem of Algorithm 1, we first establish three important lemmas. One of them is responsible for handling the deterministic step of the proof, while the other two aim to circumvent the difficulties introduced by the Markovian nature of stochasticity. Let us start with the classical Gidel et al. (2018); Mishchenko et al. (2020); Hsieh et al. (2019) descent lemma:

Lemma 1. *Let Assumptions 1, 2 be satisfied. Then for the iterates* $\{z_t\}_{t \geq 0}$ *of Algorithm 1 it holds that:*

$$\|z^{t+1} - z^*\|^2 \leq \|z^t - z^*\|^2 - 2\gamma\mu\|z^{t+\frac{1}{2}} - z^*\|^2 - 2\gamma\langle F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^* \rangle$$
$$+ \gamma^2 L^2 \|z^t - z^{t+\frac{1}{2}}\|^2 - \|z^t - z^{t+\frac{1}{2}}\|^2.$$

Proof. We start by using line 5 of Algorithm 1:

$$\begin{aligned}\|z^{t+1} - z^*\|^2 &= \|z^t - \gamma F(z^{t+\frac{1}{2}}, \xi^t) - z^*\|^2 \\ &= \|z^t - z^*\|^2 - 2\gamma\langle z^t - z^*, F(z^{t+\frac{1}{2}}, \xi^t)\rangle + \|\gamma F(z^{t+\frac{1}{2}}, \xi^t)\|^2 \\ &= \|z^t - z^*\|^2 - 2\gamma\langle z^{t+\frac{1}{2}} - z^*, F(z^{t+\frac{1}{2}}, \xi^t)\rangle \\ &\quad + 2\gamma\langle z^{t+\frac{1}{2}} - z^t, F(z^{t+\frac{1}{2}}, \xi^t)\rangle + \|\gamma F(z^{t+\frac{1}{2}}, \xi^t)\|^2.\end{aligned}$$

Note that

$$\begin{aligned}\|\gamma F(z^{t+\frac{1}{2}}, \xi^t)\|^2 = \|\gamma F(z^{t+\frac{1}{2}}, \xi^t) - \gamma F(z^t, \xi^t)\|^2 - \|\gamma F(z^t, \xi^t)\|^2 \\ + 2\langle \gamma F(z^{t+\frac{1}{2}}, \xi^t), \gamma F(z^t, \xi^t)\rangle.\end{aligned}$$

Using this and the fact $z^{t+\frac{1}{2}} - z^t = -\gamma F(z^t, \xi^t)$, we have

$$\begin{aligned}\|z^{t+1} - z^*\|^2 &= \|z^t - z^*\|^2 - 2\gamma\langle F(z^{t+\frac{1}{2}}, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle \\ &\quad - 2\langle \gamma F(z^t, \xi^t), \gamma F(z^{t+\frac{1}{2}}, \xi^t)\rangle \\ &\quad + \|\gamma F(z^{t+\frac{1}{2}}, \xi^t) - \gamma F(z^t, \xi^t)\|^2 - \|\gamma F(z^t, \xi^t)\|^2 \\ &\quad + 2\langle \gamma F(z^{t+\frac{1}{2}}, \xi^t), \gamma F(z^t, \xi^t)\rangle \\ &= \|z^t - z^*\|^2 - 2\gamma\langle F(z^{t+\frac{1}{2}}, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle - \|z^{t+\frac{1}{2}} - z^t\|^2 \\ &\quad + \gamma^2\|F(z^{t+\frac{1}{2}}, \xi^t) - F(z^t, \xi^t)\|^2.\end{aligned}$$

Next, we use Assumption 1:

$$\begin{aligned}\|z^{t+1} - z^*\|^2 \leq \|z^t - z^*\|^2 - 2\gamma\langle F(z^{t+\frac{1}{2}}, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle \\ + \gamma^2 L^2\|z^{t+\frac{1}{2}} - z^t\|^2 - \|z^{t+\frac{1}{2}} - z^t\|^2.\end{aligned} \tag{2}$$

Applying Assumption 2 to the second term, we get

$$\begin{aligned}\langle F(z^{t+\frac{1}{2}}, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle &= \langle F(z^{t+\frac{1}{2}}, \xi^t) - F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle \\ &\quad + \langle F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle \\ &\geq \mu\|z^{t+\frac{1}{2}} - z^*\|^2 + \langle F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle\end{aligned}$$

Substituting this into (2) completes the proof. □

One of the effective methods for addressing Markovian stochasticity is the technique of stepping back by $\mathcal{T} > \tau_{mix}(\varepsilon)$ steps. In contrast to the i.i.d. case, in which the term of the form $\mathbb{E}\langle F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle$ equals to zero, we here need to handle it in the following way:

$$\langle F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle = \langle F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^{t+\frac{1}{2}-\mathcal{T}}\rangle + \langle F(z^*, \xi^t), z^{t+\frac{1}{2}-\mathcal{T}} - z^*\rangle, \tag{3}$$

taking what is referred to as the step back. The first term in (3) is treated with the help of Cauchy-Schwarz inequality and the second lemma.

Lemma 2. *Let $\mathcal{T} \in \mathbb{N}$ be a fixed number and let $\{z_t\}_{t\geq 0}$ be the iterates of Algorithm 1. Then, for all $t \geq \mathcal{T}$ it holds that*

$$\|z^{t+\frac{1}{2}} - z^{t+\frac{1}{2}-\mathcal{T}}\| \leq \sum_{k=0}^{\mathcal{T}} [1 + \gamma L]\|z^{t+\frac{1}{2}-k} - z^{t-k}\|.$$

Proof. We start with the triangle inequality:

$$\begin{aligned}\|z^{t+\frac{1}{2}} - z^{t+\frac{1}{2}-\mathcal{T}}\| \leq \|z^{t+\frac{1}{2}} - z^t\| + \|[z^t - z^{t-1}] - [z^{t-\frac{1}{2}} - z^{t-1}]\| \\ + \|z^{t-\frac{1}{2}} - z^{t+\frac{1}{2}-\mathcal{T}}\|.\end{aligned}$$

Using lines 4 and 5 of Algorithm 1, we obtain:

$$\|z^{t+\frac{1}{2}} - z^{t+\frac{1}{2}-\mathcal{T}}\| \leq \|z^{t+\frac{1}{2}} - z^t\| + \gamma\|F(z^{t-\frac{1}{2}}, \xi^{t-1}) - F(z^{t-1}, \xi^{t-1})\|$$

$$
\begin{aligned}
&+ \|z^{t-\frac{1}{2}} - z^{t+\frac{1}{2}-\mathcal{T}}\| \\
&\leq \|z^{t+\frac{1}{2}} - z^{t}\| + \gamma L \|z^{t-\frac{1}{2}} - z^{t-1}\| + \|z^{t-\frac{1}{2}} - z^{t+\frac{1}{2}-\mathcal{T}}\|,
\end{aligned}
$$

where in the last inequality follows from Assumption 1. The same steps can now be applied to the $\|z^{t-\frac{1}{2}} - z^{t+\frac{1}{2}-\mathcal{T}}\|$ term. Finally, after performing the recursion, we obtain

$$
\begin{aligned}
\|z^{t+\frac{1}{2}} - z^{t+\frac{1}{2}-\mathcal{T}}\| \leq \|z^{t+\frac{1}{2}} - z^{t}\| &+ \sum_{k=1}^{\mathcal{T}-1} [1+\gamma L] \|z^{t+\frac{1}{2}-k} - z^{t-k}\| \\
&+ \gamma L \|z^{t+\frac{1}{2}-\mathcal{T}} - z^{t-\mathcal{T}}\|.
\end{aligned}
$$

This concludes the proof. □

The second term in (3) is resolved with the following lemma.

Lemma 3. *For any $\varepsilon > 0$, $\mathcal{T} > \tau_{mix}(\varepsilon)$, $t > \mathcal{T}$, for any $z^{t+\frac{1}{2}-\mathcal{T}} \in \mathbb{R}^d$, such that if we fix all randomness up to step* $t + \frac{1}{2} - \mathcal{T}$*, $z^{t+\frac{1}{2}-\mathcal{T}}$ becomes non-random, it holds that*

$$
\mathbb{E}\left[\langle F(z^*, \xi^t), z^* - z^{t+\frac{1}{2}-\mathcal{T}}\rangle\right] \leq \varepsilon \sigma_* \mathbb{E}\left[\|z^* - z^{t+\frac{1}{2}-\mathcal{T}}\|\right].
$$

Proof. We begin by using tower property:

$$
\begin{aligned}
\mathbb{E}\left[\langle F(z^*, \xi^t), z^* - z^{t+\frac{1}{2}-\mathcal{T}}\rangle\right] &= \mathbb{E}\left[\langle F(z^*, \xi^t) - F(x^*), z^* - z^{t+\frac{1}{2}-\mathcal{T}}\rangle\right] \\
&= \mathbb{E}\left[\langle \mathbb{E}_{t+\frac{1}{2}-\mathcal{T}}\left[F(z^*, \xi^t) - F(x^*)\right], z^* - z^{t+\frac{1}{2}-\mathcal{T}}\rangle\right],
\end{aligned}
$$

where $\mathbb{E}_{t+\frac{1}{2}-\mathcal{T}}\left[\cdot\right]$ is the conditional expectation with fixed randomness of all steps up to $t + \frac{1}{2} - \mathcal{T}$. Let us enroll the inner expectation:

$$
\begin{aligned}
&\mathbb{E}_{t+\frac{1}{2}-\mathcal{T}}\left[\langle F(z^*, \xi^t), z^* - z^{t+\frac{1}{2}-\mathcal{T}}\rangle\right] \\
&\qquad = \mathbb{E}\Big\langle \sum_{\xi \in \mathcal{M}} \Big(\mathbb{P}(\xi^t = \xi | z^{t+\frac{1}{2}-\mathcal{T}}) - \pi_\xi\Big) F(z^*, \xi), z^* - z^{t-\frac{1}{2}-\mathcal{T}}\Big\rangle \\
&\qquad \overset{①}{\leq} \mathbb{E}\| \sum_{\xi \in \mathcal{M}} \Big(\mathbb{P}(\xi^t = \xi | z^{t+\frac{1}{2}-\mathcal{T}}) - \pi_\xi\Big) F(z^*, \xi)\| \cdot \|z^* - z^{t-\frac{1}{2}-\mathcal{T}}\| \\
&\qquad \leq \mathbb{E}\left[\sum_{\xi \in \mathcal{M}} |\mathbb{P}(\xi^t = \xi | z^{t+\frac{1}{2}-\mathcal{T}}) - \pi_\xi| \|F(z^*, \xi)\| \|z^* - z^{t-\frac{1}{2}-\mathcal{T}}\|\right],
\end{aligned}
$$

where in ① we used Cauchy-Schwarz inequality (8). Now, using Assumption 4, which gives us $|\mathbb{P}(\xi^t = \xi | z^{t+\frac{1}{2}-\mathcal{T}}) - \pi_\xi| \leq \varepsilon \pi_\xi$, and Assumption 3, we obtain:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{\xi \in \mathcal{M}} \varepsilon \pi_\xi \|F(z^*, \xi)\| \cdot \|z^* - z^{t-\frac{1}{2}-\mathcal{T}}\|\right] &\leq \mathbb{E}\left[\sum_{\xi \in \mathcal{M}} \varepsilon \pi_\xi \sigma_* \|z^* - z^{t-\frac{1}{2}-\mathcal{T}}\|\right] \\
&= \mathbb{E}\left[\varepsilon \sigma_* \|z^* - z^{t-\frac{1}{2}-\mathcal{T}}\|\right].
\end{aligned}
$$

This finishes the proof. □

Now we are ready to prove the convergence of Algorithm 1. The proof scheme is to utilize Lemmas 1 and 2 and Cauchy-Schwarz inequalities, proceed by taking the full mathematical expectation on both sides of the resulting expressions and, with the aid of Lemma 3, sum these expectations from some index $\tau > \tau_{mix}(\varepsilon)$ up to iteration $T - 1$. Through these steps and by imposing specific conditions on the parameters γ and ε, we derive the final convergence result.

Theorem 1 (Convergence of Algorithm 1)**.** *Let Assumptions 1, 2, 3, 4 be satisfied. Let the problem* (1) *be solved by Algorithm 1. Then, for all* $\varepsilon > 0,\ \gamma > 0,\ T \geq \tau_{mix}(\varepsilon)$ *such that* $\gamma \leq \min\{\frac{1}{2L}, \frac{1}{4\mu}\},\ \varepsilon \leq \min\{\frac{\mu}{6}, \gamma\}$*, it holds that*

$$\mathbb{E}\|z^{T+1} - z^*\|^2 \leq \left(1 - \frac{\mu\gamma}{2}\right)^T \left[\left(1 - \frac{\mu\gamma}{2}\right)^{-\tau} \mathbb{E}\|z^\tau - z^*\|^2 + \Delta_\tau\right] + \frac{56\gamma\tau^2}{\mu}\sigma_*^2,$$

where $\Delta_\tau := 6\sum\limits_{t=0}^{\tau-1}\mathbb{E}\Big[\|z^{t+\frac{1}{2}} - z^t\|^2 + \|z^{t+\frac{1}{2}} - z^*\|^2\Big]$.

Proof. We start by using Lemma 1:

$$\|z^{t+1} - z^*\|^2 \leq \|z^t - z^*\|^2 - 2\gamma\mu\|z^{t+\frac{1}{2}} - z^*\|^2 - 2\gamma\langle F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle$$
$$+ \gamma^2 L^2\|z^t - z^{t+\frac{1}{2}}\|^2 - \|z^t - z^{t+\frac{1}{2}}\|^2.$$

Take a look at the second term. We use the stepping back technique:

$$\langle F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^*\rangle = \langle F(z^*, \xi^t), z^{t+\frac{1}{2}-\tau} - z^*\rangle + \langle F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^{t+\frac{1}{2}-\tau}\rangle,$$

where τ is an arbitrary number such that $\tau \geq \tau_{mix}(\varepsilon)$ with ε to be specified later. Using the last bound, one can obtain:

$$\begin{aligned}\|z^{t+1} - z^*\|^2 &\leq \|z^t - z^*\|^2 - 2\gamma\mu\|z^{t+\frac{1}{2}} - z^*\|^2 - 2\gamma\langle F(z^*, \xi^t), z^{t+\frac{1}{2}-\tau} - z^*\rangle\\ &\quad - 2\gamma\langle F(z^*, \xi^t), z^{t+\frac{1}{2}} - z^{t+\frac{1}{2}-\tau}\rangle - \|z^{t+\frac{1}{2}} - z^t\|^2\\ &\quad + \gamma^2L^2\|z^{t+\frac{1}{2}} - z^t\|^2\\ &= \|z^t - z^*\|^2 - 2\gamma\mu\|z^{t+\frac{1}{2}} - z^*\|^2 + 2\gamma\langle F(z^*, \xi^t), -z^{t+\frac{1}{2}-\tau} + z^*\rangle\\ &\quad + 2\gamma\langle F(z^*, \xi^t), -z^{t+\frac{1}{2}} + z^{t+\frac{1}{2}-\tau}\rangle - \|z^{t+\frac{1}{2}} - z^t\|^2\\ &\quad + \gamma^2L^2\|z^{t+\frac{1}{2}} - z^t\|^2.\end{aligned}$$

Now we use Cauchy-Schwarz inequality (8):

$$\|z^{t+1} - z^*\|^2 \leq \|z^t - z^*\|^2 - 2\gamma\mu\|z^{t+\frac{1}{2}} - z^*\|^2 - 2\gamma\langle F(z^*, \xi^t), z^{t+\frac{1}{2}-\tau} - z^*\rangle$$
$$+ 2\gamma\|F(z^*, \xi^t)\|\,\|z^{t+\frac{1}{2}} - z^{t+\frac{1}{2}-\tau}\| + \gamma^2L^2\|z^{t+\frac{1}{2}} - z^t\|^2 - \|z^{t+\frac{1}{2}} - z^t\|^2.$$

Applying Lemma 2 and Assumption 3, we get:

$$\|z^{t+1} - z^*\|^2 \leq \|z^t - z^*\|^2 - 2\gamma\mu\|z^{t+\frac{1}{2}} - z^*\|^2 - 2\gamma\langle F(z^*, \xi^t), z^{t+\frac{1}{2}-\tau} - z^*\rangle$$
$$+ 2\gamma\sigma_*\sum_{k=0}^{\tau}[1 + \gamma L]\|z^{t+\frac{1}{2}-k} - z^{t-k}\| + \gamma^2L^2\|z^{t+\frac{1}{2}} - z^t\|^2 - \|z^{t+\frac{1}{2}} - z^t\|^2.$$

Using Cauchy-Schwarz inequality (7) with $\beta = \beta_1$ to be specified later, one can obtain:

$$\|z^{t+1} - z^*\|^2 \leq \|z^t - z^*\|^2 - 2\gamma\mu\|z^{t+\frac{1}{2}} - z^*\|^2 - 2\gamma\langle F(z^*, \xi^t), z^{t+\frac{1}{2}-\tau} - z^*\rangle$$
$$+ \frac{\gamma}{\beta_1}{\sigma_*}^2 + \gamma(1 + \gamma L)^2\beta_1\left[\sum_{k=0}^{\tau}\|z^{t+\frac{1}{2}-k} - z^{t-k}\|\right]^2$$
$$+ \gamma^2L^2\|z^{t+\frac{1}{2}} - z^t\|^2 - \|z^{t+\frac{1}{2}} - z^t\|^2.$$

We now take the full mathematical expectation from both sides of the last inequality:

$$\mathbb{E}\|z^{t+1} - z^*\|^2 \leq \mathbb{E}\|z^t - z^*\|^2 - 2\gamma\mu\mathbb{E}\|z^{t+\frac{1}{2}} - z^*\|^2 - \mathbb{E}\|z^{t+\frac{1}{2}} - z^t\|^2$$
$$+ \frac{\gamma}{\beta_1}{\sigma_*}^2 + \gamma(1 + \gamma L)^2\beta_1\mathbb{E}\left[\sum_{k=0}^{\tau}\|z^{t+\frac{1}{2}-k} - z^{t-k}\|\right]^2$$
$$+ \gamma^2L^2\mathbb{E}\|z^{t+\frac{1}{2}} - z^t\|^2 - 2\gamma\mathbb{E}\langle F(z^*, \xi^t), z^{t+\frac{1}{2}-\tau} - z^*\rangle.$$

Using the convexity of the squared norm, we get:

$$\begin{aligned}\mathbb{E}\|z^{t+1}-z^*\|^2 \le{}& \mathbb{E}\|z^t-z^*\|^2 - 2\gamma\mu\mathbb{E}\|z^{t+\frac{1}{2}}-z^*\|^2 - \mathbb{E}\|z^{t+\frac{1}{2}}-z^t\|^2 \\ &+ \frac{\gamma}{\beta_1}{\sigma_*}^2 + \gamma(1+\gamma L)^2\beta_1\tau\sum_{k=0}^{\tau}\mathbb{E}\|z^{t+\frac{1}{2}-k}-z^{t-k}\|^2 \\ &+ \gamma^2L^2\mathbb{E}\|z^{t+\frac{1}{2}}-z^t\|^2 - 2\gamma\mathbb{E}\langle F(z^*,\xi^t), z^{t+\frac{1}{2}-\tau}-z^*\rangle.\end{aligned}$$

Applying Lemma 3 to the last inequality leads to:

$$\begin{aligned}\mathbb{E}\|z^{t+1}-z^*\|^2 \le{}& \mathbb{E}\|z^t-z^*\|^2 - 2\gamma\mu\mathbb{E}\|z^{t+\frac{1}{2}}-z^*\|^2 + 2\gamma\varepsilon\sigma_*\mathbb{E}\|z^{t+\frac{1}{2}-\tau}-z^*\| \\ &+ \frac{\gamma}{\beta_1}{\sigma_*}^2 + \gamma(1+\gamma L)^2\beta_1\tau\sum_{k=0}^{\tau}\mathbb{E}\|z^{t+\frac{1}{2}-k}-z^{t-k}\|^2 \\ &+ \gamma^2L^2\mathbb{E}\|z^{t+\frac{1}{2}}-z^t\|^2 - \mathbb{E}\|z^{t+\frac{1}{2}}-z^t\|^2.\end{aligned}$$

Using Cauchy-Schwarz inequality (7) with $\beta = 1$, we obtain:

$$\begin{aligned}\mathbb{E}\|z^{t+1}-z^*\|^2 \le{}& \mathbb{E}\|z^t-z^*\|^2 - 2\gamma\mu\mathbb{E}\|z^{t+\frac{1}{2}}-z^*\|^2 + \gamma\varepsilon\mathbb{E}\|z^{t+\frac{1}{2}-\tau}-z^*\|^2 \\ &+ (\frac{\gamma}{\beta_1}+\gamma\varepsilon){\sigma_*}^2 + \gamma(1+\gamma L)^2\beta_1\tau\sum_{k=0}^{\tau}\mathbb{E}\|z^{t+\frac{1}{2}-k}-z^{t-k}\|^2 \\ &+ (\gamma^2L^2-1)\mathbb{E}\|z^{t+\frac{1}{2}}-z^t\|^2.\end{aligned}$$

For $t \ge 0$, let $p_t = p^t$ and $p = (1-\mu\gamma/2)^{-1}$. Here we multiply the above expression by p_t and sum for $\tau \le t < T$, hoping for cancellations:

$$\begin{aligned}\sum_{t=\tau}^{T-1}p_t\mathbb{E}\|z^{t+1}-z^*\|^2 \le{}& \sum_{t=\tau}^{T-1}p_t\mathbb{E}\|z^t-z^*\|^2 - 2\gamma\mu\sum_{t=\tau}^{T-1}p_t\mathbb{E}\|z^{t+\frac{1}{2}}-z^*\|^2 \\ &+ (\frac{\gamma}{\beta_1}+\gamma\varepsilon){\sigma_*}^2\sum_{t=\tau}^{T-1}p_t + \gamma\varepsilon\sum_{t=\tau}^{T-1}p_t\mathbb{E}\|z^{t+\frac{1}{2}-\tau}-z^*\|^2 \\ &+ (\gamma^2L^2-1)\sum_{t=\tau}^{T-1}p_t\mathbb{E}\|z^{t+\frac{1}{2}}-z^t\|^2 \\ &+ \gamma(1+\gamma L)^2\beta_1\tau\sum_{t=\tau}^{T-1}p_t\sum_{k=0}^{\tau}\mathbb{E}\|z^{t+\frac{1}{2}-k}-z^{t-k}\|^2.\end{aligned}$$

Using the fact that $(1-a/x)^{-x} \le 2e^a \le 2e$ for any $x \ge 2$ and $0 \le a \le 1$, we can estimate $p_\tau = (1-\mu\gamma_1/(2\tau))^{-\tau} \le 2e \le 6$, where $\gamma_1 = \gamma\tau, \gamma \le (\mu\tau)^{-1}$. Then, we can rearrange the sums:

$$\begin{aligned}\sum_{t=\tau}^{T-1}p_t\sum_{k=0}^{\tau}\mathbb{E}\|z^{t+\frac{1}{2}-k}-z^{t-k}\|^2 \le{}& p^\tau\sum_{t=\tau}^{T-1}\sum_{k=0}^{\tau}p_k\mathbb{E}\|z^zt+\frac{1}{2}-k-z^{t-k}\|^2 \\ \le{}& 6\tau\sum_{t=0}^{T-1}p_t\mathbb{E}\|z^{t+\frac{1}{2}}-z^t\|^2.\end{aligned}$$

Now we can estimate:

$$\begin{aligned}\sum_{t=\tau}^{T-1}p_t\mathbb{E}\|z^{t+1}-z^*\|^2 \le{}& \sum_{t=\tau}^{T-1}p_t\mathbb{E}\|z^t-z^*\|^2 - 2\gamma\mu\sum_{t=\tau}^{T-1}p_t\mathbb{E}\|z^{t+\frac{1}{2}}-z^*\|^2 \\ &+ (\frac{\gamma}{\beta_1}+\gamma\varepsilon){\sigma_*}^2\sum_{t=\tau}^{T-1}p_t + \gamma\varepsilon\sum_{t=\tau}^{T-1}p_t\mathbb{E}\|z^{t+\frac{1}{2}-\tau}-z^*\|^2 \qquad (4) \\ &+ (\gamma^2L^2-1)\sum_{t=\tau}^{T-1}p_t\mathbb{E}\|z^{t+\frac{1}{2}}-z^t\|^2\end{aligned}$$

$$+ 6\gamma(1+\gamma L)^2\beta_1\tau^2 \sum_{t=0}^{T-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}} - z^t\|^2.$$

Assume the following notation:

$$\Delta_\tau := \gamma\varepsilon \sum_{t=\tau}^{2\tau-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}-\tau} - z^*\|^2 + 6\gamma(1+\gamma L)^2\beta\tau^2 \sum_{t=0}^{\tau-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}} - z^t\|^2.$$

Then, rearranging (4), one can obtain

$$\sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^{t+1} - z^*\|^2 \leq \sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^t - z^*\|^2 - 2\gamma\mu \sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}} - z^*\|^2$$
$$+ (\frac{\gamma}{\beta_1} + \gamma\varepsilon){\sigma_*}^2 \sum_{t=\tau}^{T-1} p_t + \gamma\varepsilon \sum_{t=2\tau}^{T-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}-\tau} - z^*\|^2$$
$$+ (\gamma^2 L^2 - 1) \sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}} - z^t\|^2$$
$$+ 6\gamma(1+\gamma L)^2\beta_1\tau^2 \sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}} - z^t\|^2 + \Delta_\tau.$$

It follows from Cauchy-Schwarz inequality (7) with $\beta = 1$, that:

$$-\gamma\mu\|z^{t+\frac{1}{2}} - z^*\|^2 \leq -\frac{\gamma\mu}{2}\|z^t - z^*\|^2 + \gamma\mu\|z^{t+\frac{1}{2}} - z^t\|^2.$$

Combining this with the fact that

$$\gamma\varepsilon \sum_{t=2\tau}^{T-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}-\tau} - z^*\|^2 \leq 6\gamma\varepsilon \sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}} - z^*\|^2,$$

we get:

$$\sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^{t+1} - z^*\|^2 \leq \Big(1 - \frac{\mu\gamma}{2}\Big) \sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^t - z^*\|^2$$
$$+ \gamma(6\varepsilon - \mu) \sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}} - z^*\|^2 + (\frac{\gamma}{\beta_1} + \gamma\varepsilon) \sum_{t=\tau}^{T-1} p_t {\sigma_*}^2 + \Delta_\tau$$
$$+ (6\gamma(1+\gamma L)^2\beta_1\tau^2 + \gamma^2 L^2 + \mu\gamma - 1) \sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^{t+\frac{1}{2}} - z^t\|^2.$$

Taking

$$\gamma \leq \min\{\frac{1}{2L}, \frac{1}{4\mu\tau}\}, \qquad \varepsilon \leq \min\{\frac{\mu}{6}, \gamma\}, \qquad \beta_1 = \frac{1}{27\gamma\tau^2},$$

we obtain:

$$\gamma(6\varepsilon - \mu) \leq 0,$$

$$6\gamma(1+\gamma L)^2\beta_1\tau^2 + \gamma^2 L^2 + \mu\gamma - 1 \leq 0.$$

Therefore, we can claim that:

$$\sum_{t=\tau}^{T-1} p_t \mathbb{E}\|z^{t+1} - z^*\|^2 \leq \sum_{t=\tau}^{T-1} \left[1 - \frac{\gamma\mu}{2}\right] p_t \mathbb{E}\|z^t - z^*\|^2 + 28\gamma^2\tau^2 \sum_{t=\tau}^{T-1} p_t {\sigma_*}^2 + \Delta_\tau.$$

Finally, we substitute $p_t := \big(1 - \frac{\gamma\mu}{2}\big)^{-t}$:

$$\sum_{t=\tau}^{T-1} \left[1 - \frac{\gamma\mu}{2}\right]^{-t} \mathbb{E}\|z^t - z^*\| \leq \sum_{t=\tau}^{T-1} \left[1 - \frac{\gamma\mu}{2}\right]^{-t+1} \mathbb{E}\|z^{t+1} - z^*\|^2$$

$$+ 28\gamma^2\tau^2\sigma_*^2 \sum_{t=\tau}^{T-1} \left[1 - \frac{\gamma\mu}{2}\right]^{-t} + \Delta_\tau;$$

Upon removing the contracting terms, the following remains:

$$\begin{aligned}\mathbb{E}\|z^T - z^*\| \leq \left[1 - \frac{\gamma\mu}{2}\right]^{T-\tau} \mathbb{E}\|z^\tau - z^*\|^2 + \left[1 - \frac{\gamma\mu}{2}\right]^{T} \Delta_\tau \\ + 28\gamma^2\tau^2\sigma_*^2 \sum_{t=\tau}^{T} \left[1 - \frac{\gamma\mu}{2}\right]^{T-t}.\end{aligned}$$

Next, we use that:

$$\sum_{t=\tau}^{T} \left[1 - \frac{\gamma\mu}{2}\right]^{T-t} = \sum_{t=0}^{T-\tau} \left[1 - \frac{\gamma\mu}{2}\right]^{t} \leq \sum_{t=0}^{\infty} \left[1 - \frac{\gamma\mu}{2}\right]^{t} = \frac{2}{\gamma\mu},$$

and bound Δ_τ:

$$\Delta_\tau \leq 6 \sum_{t=0}^{\tau-1} \mathbb{E}\Big[\|z^{t+\frac{1}{2}} - z^t\|^2 + \|z^{t+\frac{1}{2}} - z^*\|^2\Big],$$

leading us to:

$$\mathbb{E}\|z^{T+1} - z^*\|^2 \leq \left(1 - \frac{\mu\gamma}{2}\right)^{T} \left[\left(1 - \frac{\mu\gamma}{2}\right)^{-\tau} \mathbb{E}\|z^\tau - z^*\|^2 + \Delta_\tau\right] + \frac{56\gamma\tau^2}{\mu}\sigma_*^2.$$

This finishes the proof. □

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

$$\gamma \leq \min\left\{\frac{1}{2L}, \frac{1}{4\mu}, \frac{2\log(\max\{2, \frac{\mu^2 r_\tau T^2}{112\tau^2\sigma_*^2}\})}{\mu T}\right\}$$

in order to achieve the ϵ-approximate solution in terms of $\mathbb{E}\|z^T - z^*\|^2 \leq \epsilon^2$ it takes

$$T = \tilde{\mathcal{O}}\Bigg(\left(\tau + \frac{L}{\mu}\right)\log\frac{1}{\epsilon} + \frac{\tau^2\sigma_*^2}{\mu^2\epsilon}\Bigg) \quad \textit{oracle calls.} \tag{5}$$

4 Discussion.

In spite of the fact that the Markov noise setup in the context of a classical optimization problem has a quite wide representation in the literature, there is considerably less work for the VI case. To the best of our knowledge, there are only three existing works on the topic of VI with Markovian stochasticity. Two of them, Wang et al. (2022) and Solodkin et al. (2024), consider only monotone setting, obtaining results of the form $T = \widetilde{\mathcal{O}}\left(\frac{L\sigma^2}{\epsilon^2} + \frac{\tau^2\sigma^4}{\epsilon^2}\right)$ and $T = \widetilde{\mathcal{O}}\left(\frac{LD^2}{\epsilon} + \frac{\tau D^2\sigma^2}{\epsilon^2}\right)$ respectively. The third work Beznosikov et al. (2023b) provides the following convergence guarantees: $T = \tilde{\mathcal{O}}\left(\frac{\tau L}{\mu}\log\frac{1}{\epsilon} + \frac{\tau\sigma^2}{\mu^2\epsilon}\right)$. This result is nearly identical to that of Corollary 1, with the exception of the mixing time entry. However, the authors of Beznosikov et al. (2023b) utilize the batching technique with batches of size $\tilde{\mathcal{O}}(\tau)$, resulting in a significant increase in gradient evaluations. Moreover, this bound is contingent upon the assumption of uniformly bounded gradient differences, while our analysis is considerably less demanding with the necessity in only bound at the optimal point.

5 Numerical Experiments

In this section, we present numerical experiments that are designed to investigate the effect of mixing time on the convergence rate.

5.1 Problem formulation

Let λ, ν be positive real numbers. Let $b, c \in \mathbb{R}^d$ be vectors with randomly generated from $[-1, 1]$ real components. Finally, define $P \in \mathbb{R}^{d \times d}$ as a matrix of randomly generated real values such that $\mathbf{spec}(P) \subset [0.1, 10]$.

Introducing this notation, we consider the following formulation of the problem (1):

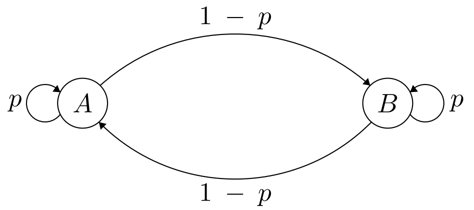
$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^d} \left[f(x, y) := x^T P y + b^T x + c^T y + \frac{\lambda}{2}\|x\|_2^2 - \frac{\nu}{2}\|y\|_2^2 \right]. \tag{6}$$

For this problem, the operator has the following form:

$$F(z) := \begin{pmatrix} \nabla_x f(x, y) \\ -\nabla_y f(x, y) \end{pmatrix} = \begin{pmatrix} Py + b + \lambda x \\ -P^T x - c - \nu y \end{pmatrix}$$

5.2 Setup

In the numerical experiments, we consider the problem described above on the different ergodic Markov chains. In order to compare the outcomes properly, we let all the Markov chains have the similar structure:



Here p is a unique parameter of the Markov chain, arrows denote the transitions, and A and B represent the states. Each state implies a unique current distribution that generates the noise. In our experiments, we assume that both of the states have a normal distribution, in particular, state A generates a value from the distribution $\mathcal{N}(0.1, \sigma^2)$, state B generates a value from $\mathcal{N}(-0.1, \sigma^2)$, where σ is a varying parameter for deeper study. The generated noise ξ is considered to be additive:

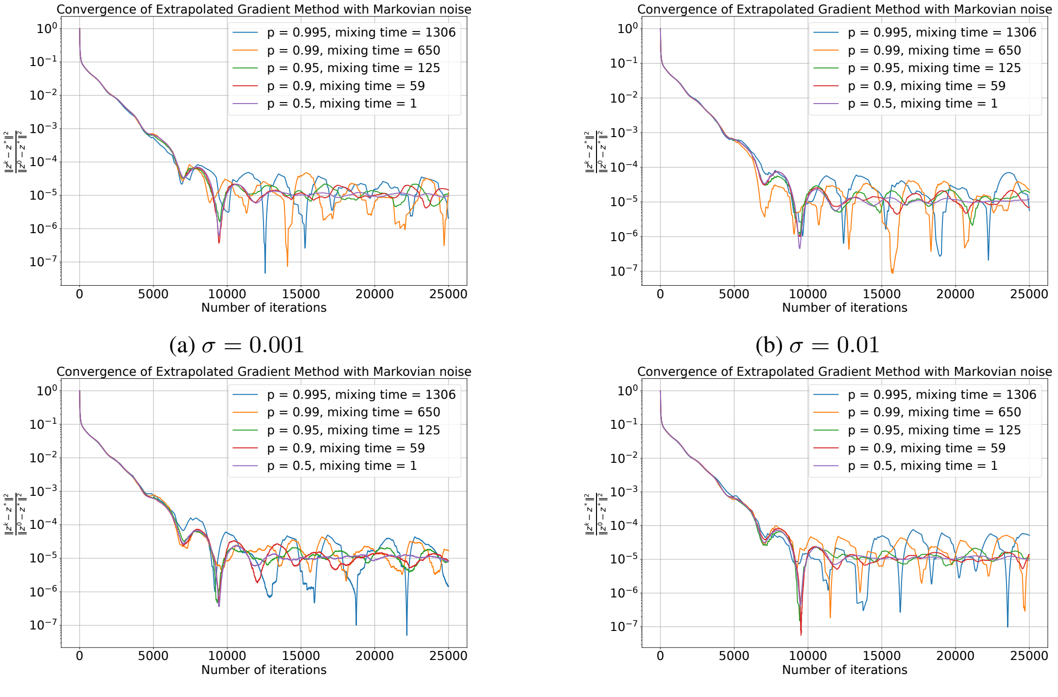
$$[F(z, \xi)]_i := [F(z)]_i + \xi_i.$$

5.3 Results

We first performed the experiments with $\sigma = 0.001$ (Figure 1a), $\sigma = 0.01$ (Figure 1b), $\sigma = 0.1$ (Figure 1c) and $\sigma = 0$ (Figure 1d) as the standard deviations of normal distributions described in the setup.

As illustrated in Figure 1, the oscillation amplitude of the method is dependent on the mixing time. Furthermore, stochasticity does not impact this dependence, as evidenced by the consistent character of convergence observed in experiments with varying standard deviations. At this juncture, it seems appropriate to examine the nature of this dependence in greater detail.

Now, we fix the values of standard deviation $\sigma = 0.1$ and expectation $\mu = 0.1$ for the noise variables. In order to most accurately assess the impact of the mixing time parameter on the convergence region, we ran Algorithm 1 $K = 14$ times for each value of p with identical hyperparameters $\gamma = \frac{1}{2L}$. Subsequently, the sample variance was calculated for each solution, i.e. denoting the results of j-th run by the $\{z_t^j\}_{t=0}^T$, we calculate $\frac{1}{K}\sum\limits_{j=1}^{K}\left[\frac{1}{T}\sum\limits_{t=1}^{T}\left(z_t^j - \overline{z}^j\right)^2\right]$, where $\overline{z}^j = \frac{1}{T}\sum\limits_{t=1}^{T} z_t^j$. As a consequence, we were able to establish a reliable correlation between the convergence region and the mixing time, which serves to substantiate the conclusions reached in the theoretical analysis. However, the results of Figure 2 indicate that the nature of this dependence is tend to be linear, whereas the theoretical estimation (5) implies the quadratic correlation. This raises a slight research gap, namely whether it is feasible to design a more intricate proof that would yield a more precise assessment of the



(a) $\sigma = 0.001$

(b) $\sigma = 0.01$

(c) $\sigma = 0.1$

(d) $\sigma = 0$

Figure 1: Comparison of Algorithm 1 with varying mixing time parameter for the different values of the variance σ.

dependence on mixing time. Alternatively, it may be possible to modify the experimental procedure in order to more clearly observe the effect of mixing time parameter. In any case, this seems to us as the perfect topic for future research.

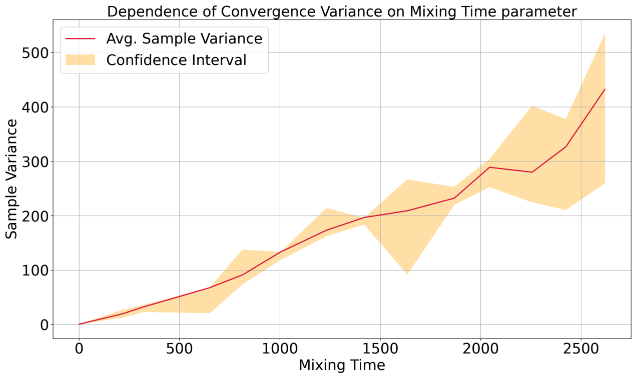


Figure 2: Comparison of convergence regions of Algoritm 1 for different values of mixing time. For the sake of clarity, the values are normalized on the variance for $\tau = 1$.

Acknowledgements

The work of V. Solodkin was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Moscow Institute of Physics and Technology dated November 1, 2021 No. 70-2021-00138.

REFERENCES

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

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.

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

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, 2023a.

Aleksandr Beznosikov, Sergey Samsonov, Marina Sheshukova, Alexander Gasnikov, Alexey Naumov, and Eric Moulines. First order methods with markovian noise: from acceleration to variational inequalities. *Advances in Neural Information Processing Systems*, 36, 2023b. doi: 10.48550/arXiv.2305.15938. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/8c3e38ce55a0fa44bc325bc6fdb7f4e5-Abstract-Conference.html. NeurIPS 2023.

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.

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

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

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.

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

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

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.

Mathieu Even. Stochastic gradient descent under markovian sampling schemes. *arXiv preprint arXiv:2302.14428*, 2023.

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

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.

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

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. *arXiv preprint arXiv:1905.11261*, 2019.

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.

Alejandro Jofre, R Rockafellar, and Roger Wets. Variational inequalities and economic equilibrium. *Mathematics of Operations Research*, 32, 2007. doi: 10.1287/moor.1060.0233.

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

Belhal Karimi, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. Non-asymptotic analysis of biased stochastic approximation scheme, 2019. URL `https://arxiv.org/abs/1902.00629`.

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

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes, 2022. URL `https://arxiv.org/abs/2102.00135`.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 907–915. PMLR, 2019.

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

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.

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.

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

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.

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

Alexander Pichugin, Maksim Pechin, Aleksandr Beznosikov, Alexander Gasnikov, and Andrey Savchenko. Optimal analysis of method with batching for monotone stochastic finite-sum variational inequalities. In *Doklady Mathematics*, volume 108, pages S348–S359. Springer, 2023.

Alexander Pichugin, Maksim Pechin, Aleksandr Beznosikov, Vasilii Novitskii, and Alexander Gasnikov. Method with batching for stochastic finite-sum variational inequalities in non-euclidean setting. *Chaos, Solitons & Fractals*, 187:115396, 2024.

L. D. Popov. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR*, 28(5):845–848, 1980.

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.

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, 2010.

Vladimir Solodkin, Andrew Veprikov, and Aleksandr Beznosikov. Methods for optimization problems with markovian stochasticity and non-euclidean geometry, 2024. URL `https://arxiv.org/abs/2408.01848`.

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

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.

Manan Tomar, Lior Shani, Yonathan Efroni, and Mohammad Ghavamzadeh. Mirror descent policy optimization, 2021. URL `https://arxiv.org/abs/2005.09814`.

V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

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.

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.

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.

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.

Supplementary Material

A Auxiliary Facts

Lemma 4 (Cauchy-Schwarz inequality)**.** *For any $a, b \in \mathbb{R}^d$ and $\beta > 0$ the following inequalities hold*

$$2\langle a, b\rangle \leq \frac{\|a\|^2}{\beta} + \beta\|b\|^2, \tag{7}$$

$$|\langle a, b\rangle| \leq \|a\|\|b\|. \tag{8}$$