Stochastic Methods in Variational Inequalities: Ergodicity, Bias and Refinements

Emmanouil V. Vlatakis Gkaragkounis
University of California, Berkeley

Angeliki Giannou
University of Wisconsin–Madison

Yudong Chen
University of Wisconsin–Madison

Qiaomin Xie
University of Wisconsin–Madison

Abstract

For min-max optimization and variational inequalities problems (VIP) encountered in diverse machine learning tasks, Stochastic Extragradient (SEG) and Stochastic Gradient Descent Ascent (SGDA) have emerged as preeminent algorithms. Constant step-size variants of SEG/SGDA have gained popularity, with appealing benefits such as easy tuning and rapid forgiveness of initial conditions, but their convergence behaviors are more complicated even in rudimentary bilinear models. Our work endeavors to elucidate and quantify the probabilistic structures intrinsic to these algorithms. By recasting the constant step-size SEG/SGDA as time-homogeneous Markov Chains, we establish a first-of-its-kind Law of Large Numbers and a Central Limit Theorem, demonstrating that the average iterate is asymptotically normal with a unique invariant distribution for an extensive range of monotone and non-monotone VIPs. Specializing to convex-concave min-max optimization, we characterize the relationship between the step-size and the induced bias with respect to the Von-Neumann's value. Finally, we establish that Richardson-Romberg extrapolation can improve proximity of the average iterate to the global solution for VIPs. Our probabilistic analysis, underpinned by experiments corroborating our theoretical discoveries, harnesses techniques from optimization, Markov chains, and operator theory.

1 Introduction

Variational inequalities problem (VIP) is a versatile framework that incorporates a broad range of problems including loss minimization, min-max optimization, bilinear games and various fixed point problems. Many problems in machine learning, such as training Generative Adversarial Networks (GANs) [16], Actor-Critic methods [40], multi-agent reinforcement learning [51] and robust learning [47], can be cast as VIPs.

In many applications of VIP, one is given only a stochastic oracle, typically constructed from finite data, that provides noisy access to the underlying operator. Various stochastic algorithms for VIP have been proposed and analyzed, with two prime examples being Stochastic Extragradient (SEG) [24] and Stochastic Gradient Descent Ascent (SGDA) methods [38]. It has been well recognized that convergence properties of stochastic VIP methods are more delicate than their deterministic and loss minimization counterparts. Nevertheless, much progress has been made in recent years, on both SEG [3, 18, 22, 25, 32, 35] and SGDA [4, 29, 31, 38, 49]. The closely related stochastic gradient descent (SGD) method [17], which can be viewed as a special case of SGDA, has an even larger and still growing literature. Classical results on these stochastic methods typically assume that a diminishing step-size is used, which allows for last-iterate convergence to the global solution [2, 11, 27, 42].

In this paper, we focus on the constant step-size variants of SEG and SGDA. Constant step-sizes are popular in practice, with several major benefits: the resulting algorithm is easy to tune with only a single

[*] Emails: emvlatakis@berkeley.edu, giannou@wisc.edu, yudong.chen@wisc.edu, qiaomin.xie@wisc.edu

parameter; it is insensitive to the initial condition, which is forgotten quickly; the algorithm makes substantial progress even in the first few iterations. Empirically, the use of constant step-size often leads to good performance in practical machine learning tasks and beyond.

The analysis of constant step-size SEG and SGDA, however, is more complicated, with various non-convergent behaviors even in rudimentary bilinear models [6, 7, 15, 22, 33]; see Figure 1 for an example. In particular, similar to SGD [10], these algorithms in general do not converge to the exact solution of the VIP. Rather, due to stochastic noise, the iterates fluctuate within a neighborhood of the solution. Existing theoretical results are typically in the form of an *upper bound* on the mean squared error or dual gap of the iterations. Such upper bounds often compound the deterministic and stochastic aspects of the convergence behavior.

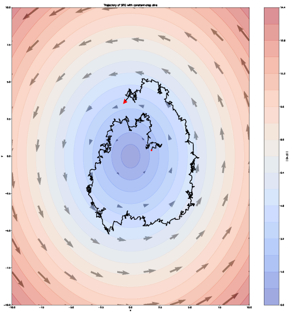


Figure 1: Example of divergent behavior in a constant step-size SEG over a quasi-bilinear Game: $\min_x \max_y \epsilon x^2 + xy - \epsilon y^2$, with $\epsilon \approx 10^{-4}$

In this work, we seek to elucidate and quantify the behaviors of SEG and SGDA with constant step-sizes. Rather than treating the stochastic fluctuation as a nuisance, we fully embrace the probabilistic nature of SEG and SGDA. By viewing them as time-homogeneous Markov chains, we study their fine-grained distributional properties, disentangling the deterministic and stochastic components. In particular, we show that while the iterate does not converge, its distribution does. Moreover, this perspective allows us to use the information provided by their fluctuation for uncertainty quantification.

Our contributions. We consider a class of VIPs with weak quasi strongly monotonicity, which encompasses a broad range of structured non-monotone and non-convex problems. Under appropriate regularity assumptions, we establish the following results.

- We prove that the iterates of SEG and SGDA form a Harris and positive recurrent Markov chain, which admits a unique stationary distribution. Our results quantify the relationship between the step-size and the regularity parameters of the VIP for ensuring recurrence.
- We show that the distribution of the iterates converges geometrically to the above stationary distribution. More generally, we establish geometric convergence of the expectation of any Lipschitz test function of the iterates.
- We derive an ergodic Law of Large Number and a Central Limit Theorem for the iterates, thereby establishing the asymptotic normality of the ergodic average of the iterates.
- We show that induced bias—the distance between the mean of the stationary distribution and the global solution of the VIP—is bounded by a linear function of the step-size and the weak monotonicity parameter. Specializing to convex-concave min-max optimization, we quantify the relationship between the step-size and the bias with respect to the Von-Neumann's value.
- For SGDA applied to quasi strongly monotone VIPs, we derive a first-order expansion of the induced bias in terms of the step-size. This characterization shows that an order-wise reduction the bias of SGDA can be achieved by the Richardson-Romberg refinement scheme.

Our challenges. Firstly, solving Variational Inequalities (VIs) consists in principal more daunting than standard minimization tasks, primarily due to the absence of a clear potential function to measure closeness to the optimal solution value. Furthermore, the stochastic Extragradient method, usually employed for smooth operators, intensifies the complexity of the analysis due to the double random steps inherent in each iteration. The interdependence between these steps, where the first random occurrence directly impacts the subsequent one, necessitates more intricate maneuvering within a high-dimensional probabilistic landscape. Amid this, our study into Richardson Extrapolation propels the discourse beyond the conventional confines of co-coercive noisy gradient oracles, revealing a more nuanced proof under milder assumptions exclusively for the expected gradient. This meticulous analysis expands the realm of what was previously comprehended also in minimization tasks. Lastly, we strive to unify the stochastic analysis across minimization, min-max scenarios, and generic VIs, paving the way in future work towards a more comprehensive understanding of constrained case and different algorithms.

Our techniques. This research provides a novel proof that the average behavior of Stochastic Extragradient (SEG) and Stochastic Gradient Descent Ascent (SGDA) methods, with a constant step-size, will converge towards a typical trajectory over time, regardless of the initial conditions. By considering these methods as continuous-state Markov Chains, the study exploits Markov Chain Central Limit Theorems, Richardson extrapolation, and Meyn & Tweedie's machinery to validate the existence of an invariant probability measure, thereby confirming the ergodic behavior. This validation is realized through the application of non-uniform versions of Doeblin's bound and the Foster-Lyapunov inequality within a well-defined "small set" around the solution set. Our study confirms that iterations will return to this small set infinitely many times, ensuring geometric convergence to a unique stationary distribution over time, regardless of the initial conditions.

1.1 Related work

Below we review prior work on VIP with a focus on stochastic methods with constant step-sizes.

Variational Inequalities. VIP and its various special cases has been studied extensively, especially in the deterministic setting where one has exact access to the operator. Many algorithms have been developed, with both asymptotic convergence and finite-time guarantees. It is beyond the scope of this paper to survey these results, but we mention that for VIPs with Lipschitz continuous and monotone operator, the works [37] study a variant of Extra Gradient algorithm [26] and establishes optimal convergence rates for ergodic average, and the work [15, 36] studies proximal point algorithm with geometric convergence results.

Most related to us are works for the stochastic setting, for which SEG [24] and SGDA [38] are two of the most prominent algorithms. Non-convergent phenomena are observed even in unconstrained bilinear games [6, 7, 15, 22, 33]. Complementarily, a growing line of work has been dedicated to better understanding of SEG and SGDA and bridging the gap between the deterministic and the stochastic cases. The work [24] provided the first analysis of SEG for monotone VIPs. Subsequent work has extended these results to other settings [3, 18, 22, 25, 32, 35]. A parallel line of work studies SGDA and its variants under different scenarios [29, 31, 38, 49]. Recently [4] proposed a unified convergence analysis that covers various SGDA methods for regularized VIPs, where the operator is either quasi-strongly monotone or ℓ-star-cocoercive. For a quantitative summary of existing results, we refer the readers to [18] for SEG and [4] for SGDA.

In this paper we consider *weakly quasi-strongly monotone* VIPs, which is a class of structured non-monotone operators under which one can bypass the the intractability issue that arises in general non-monotone regime [8, 9, 39]. Similar conditions have been considered in prior work to establish the convergence guarantee of various algorithms [18, 22, 31, 44, 49].

Constant step-size SGD and Stochastic Approximation. The literature on SGD and stochastic approximation (SA) is vast. Within this literature, our work is most related to, and in fact motivated by, a recent line of work that studies constant step-size SGD and SA through the lens of stochastic processes. The work [10] studies SGD for smooth and strongly convex functions. Extensions to non-convex functions are considered in [50], which establishes a central limit theorem that is similar in spirit to our results. More recently, [5] studies SGD for non-smooth non-convex functions. The work [12] considers constant step-size SA on Riemannian manifolds and studies the limiting behavior as the step-size approaches zero. The work [23] considers linear SA with Markovian noise; see the references therein for other recent results on SA. We mention that both [10] and [23] examine the Richardson-Romberg bias refinement scheme, which we also consider in this paper.

2 Problem setup

To provide a concrete foundation for our ensuing discussion, we first delineate the fundamental variational inequality framework that forms the backbone of our investigation in the subsequent sections.

2.1 Variational inequalities

Let $V : \mathbb{R}^d \to \mathbb{R}^d$ be a single-valued operator. The variational inequality problem related to the operator V, when no constraints are involved, is:

$$\text{Find } x^* \in \mathbb{R}^d \text{ such that } V(x^*) = 0. \tag{VI}$$

Below, we provide a series of examples which showcase potential interpretations of the operator V.

Example 2.1 (Non-linear Systems of Equations)**.** In this scenario, the operator V corresponds to the non-linear function $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^d$ that represents the system of equations. Formally, we write $V = \mathbf{F}$. The solution of (VI), denoted as x^*, is a root of $\mathbf{F}$, i.e., it satisfies $\mathbf{F}(x^*) = \mathbf{0}$.

Example 2.2 (Loss minimization)**.** In this case the operation V corresponds to the gradient of a function that we try to minimize. Formally, we have $V = \nabla f$ for some smooth loss function $f : \mathbb{R}^d \to \mathbb{R}$. Then, the solution of (VI), x^*, is a critical point of f, i.e., $\nabla f(x^*) = 0$.

Example 2.3 (Saddle-point problems)**.** Consider a smooth loss function $\mathcal{L} : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ which assigns a cost of $\mathcal{L}(x_1, x_2)$ to a player choosing $x_1 \in \mathbb{R}^{d_1}$ and a payoff $\mathcal{L}(x_1, x_2)$ to a player choosing $x_2 \in \mathbb{R}^{d_2}$. Then, the saddle-point problem associated with a $\mathcal{L}$ aims to find (x^*, y^*) such that

$$\mathcal{L}({x_1}^*, x_2) \leq \mathcal{L}({x_1}^*, x_2^*) \leq \mathcal{L}(x_1, x_2^*). \tag{1}$$

The pair $({x_1}^*, x_2^*)$ is a saddle point of $\mathcal{L}$. With $V = (\nabla_{x_1} \mathcal{L}, -\nabla_{x_2} \mathcal{L})$ the solutions of (VI) correspond to critical points of $\mathcal{L}$, while if $\mathcal{L}$ is also convex-concave it corresponds to a saddle point.

The above examples represent a broad spectrum of applications: Example 2.1 is related to Computational Fluid Dynamics and Physics, where Navier-Stokes or Maxwell equations encapsulate non-linear systems [19]; Example 2.2 is central to machine learning, reflecting model training via loss function minimization [28]; Example 2.3 garners more and more attention due to developments in GANs [7, 15, 16], Actor-Critic methods [40], and multi-agent Reinforcement Learning [51].

2.2 Assumptions

Our blanket assumptions concerning the operator V are the following:

Assumption 1. The set of solutions $\mathcal{X}^*$ of (VI) is non-empty and $\exists x^* \in \mathcal{X}^*, \mathrm{R} \in \mathbb{R}$ such that $\|x^*\| \leq \mathrm{R}$.

Assumption 2. The operator is λ-weak μ-quasi strongly monotone with $\lambda \geq 0$, $\mu > 0$, i.e.,

$$\langle V(x), x - x^* \rangle \geq \mu \|x - x^*\|^2 - \lambda \text{ for all } x \in \mathbb{R}^d \text{ and some } x^* \in \mathcal{X}^*. \tag{2}$$

Remark. Notice that Eq. (2) implies directly that $\|x_1^* - x_2^*\|^2 \leq \frac{\lambda}{\mu}$ for any $x_1^*, x_2^* \in \mathcal{X}^*$. Thus, Assumption 2 yields that $\mathcal{X}^*$ is actually contained in some ball of radius $\sqrt{\frac{\lambda}{\mu}}$.

Our next assumption pertains to the two algorithms Stochastic Gradient Descent Ascent (SGDA) and the Stochastic Extra Gradient (SEG), which are formally given in Section 3. Conforming to the customary convention in variational inequality literature, we make the presumption that when SEG is employed, we are dealing with a Lipschitz operator (so-called smooth case), while SGDA is used in scenarios that exhibit just linear growth (so-called non-smooth case).

Assumption 3. Unless we state it differently, we adopt the following convention for the Lipschitzness/bounded growth of the operator for different algorithms respectively:

- If (SEG) is run, we have that the operator V is ℓ-Lipschitz continuous, i.e.,

$$\|V(x') - V(x)\| \leq \ell \|x' - x\| \text{ for all } x, x' \in \mathbb{R}^d. \tag{3}$$

- If (SGDA) is run, we have that the operator V has at most L-linear growth, i.e.,

$$\|V(x)\| \leq L(1 + \|x\|) \text{ for all } x \in \mathbb{R}^d. \tag{4}$$

Assumption 4. In the ensuing discussion, we presuppose that our algorithms have access to V at each stage $t \geq 0$ through a stochastic oracle. Specifically, at each iteration t, the algorithm can pick a point x_t and call a black-box procedure that returns

$$V_t = V(x_t) + U_t(x_t). \tag{5}$$

Here, $(U_t(\cdot))_{t\geq 0}$ is a sequence of independent and identically distributed random fields that satisfy the following conditions: there exists a filtration (denoting the history of x_t) $(\mathcal{F}_t)_{t\geq 0}$ on a certain probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that $U_t(x_t)$ is $\mathcal{F}_{t+1}-$measurable, but not $\mathcal{F}_t-$measurable and corresponds to a noise with (i) *Zero mean*: $\mathbb{E}[U_t(x) \mid \mathcal{F}_t] = 0$ and (ii) *Bounded second moment*: $\mathbb{E}[\|U_t(x)\|^2 \mid \mathcal{F}_t] \leq \sigma^2$ for all $x \in \mathbb{R}^d$ and some constant $\sigma > 0$.

Additional remarks on the above assumptions: Assumption 1 is standard and widely adopted in the literature on VIP. Assumption 2 represents a further relaxation of μ-quasi strongly monotonicity, inspired by weakly dissipative dynamical systems and weakly convex optimization [13, 43]. This assumption is inclusive of special cases of non-monotone games. It is worth mentioning that for $\lambda > 0, \mu > 0$, it could encompass functions of the form $a_{\lambda,\mu}\|x\|^2 + b_{\lambda,\mu}\sin(\|x\|)$, as well as rescaled versions of the Rastrigin function or various non-monotone operators frequently encountered in statistical learning [46]. In the context of $\lambda = 0$, this assumption has been explored in the literature of VIPs under various names, e.g., quasi-strongly monotone problems [31], strong coherent VIPs [44], or VIPs satisfying the strong stability condition [32]. Assumption 3 corresponds to a well-established dichotomy on VIPs: we leverage (SEG) for its superior rates in smooth optimization scenarios, whereas (SGDA) is employed in cases of non-smooth optimization. Finally, Assumption 4 is standard for the analysis of stochastic algorithms in VIPs and optimization [21, 22, 32, 38, 49].

3 Algorithms

In this paper we focus on two of the most widely used algorithms for variational inequalities: Stochastic Gradient Descent Ascent (SGDA) and Stochastic Extra Gradient (SEG).

Stochastic Gradient Descent Ascent. At each time-step $t \in \mathbb{N}$, a vector $x_t \in \mathbb{R}^d$ is maintained and updated by accessing the stochastic oracle V_t, using a constant step-size $\gamma^{\text{SDGA}} \in (0, \infty)$. Formally,

$$x_{t+1} = x_t - \gamma^{\text{SGDA}} V_t = x_t - \gamma^{\text{SGDA}}(V(x_t) + U_t(x_t)), \tag{SGDA}$$

where V and $(U_t)_{t\geq 0}$ satisfy Assumptions 2–4.

Double Step-size Stochastic Extra Gradient. As previously delineated, the preferred approach for smooth variational inequality problems is the stochastic variants of the extragradient (EG) algorithm of Korpelevich [26], where at each step it uses an extra gradient "look-ahead" step $V_{t+1/2}$ to enhance convergence towards the solution. Formally, the incarnation of SEG with double constant step-size $(\alpha^{\text{SEG}}, \gamma^{\text{SEG}})$ can be defined as follows:

$$x_{t+1/2} = x_t - \gamma^{\text{SEG}} V_t, \qquad x_{t+1} = x_t - \alpha^{\text{SEG}} \gamma^{\text{SEG}} V_{t+1/2}, \tag{SEG}$$

where V and $(U_t, U_{t+1/2})_{t\geq 0}$ satisfy Assumptions 2–4 with intermediate step filtration satisfying $\mathcal{F}_{t+\frac{1}{2}} = \mathcal{F}_t$.

Inspired by seminal work on stochastic gradient descent [10], here we study the trajectories of both (SGDA) and (SEG) via the lens of Markov Chain theory. Indeed, their iterates $(x_t)_{t\geq 0}$ can be cast as time-homogeneous continuous Markov chains in $\mathbb{R}^d$.

Specifically, observe that:

(i) The iterates $(x_t)_{t\geq 0}$ of (SGDA) and (SEG) constitute respectively a Markov chain: the subsequent state x_{t+1} (post-update parameters) relies solely on the current state x_t.
(ii) The chain is time-homogeneous, meaning the transition kernel does not depend on time: this is attributed to the constant step-size in the update rule applied at each step with i.i.d. random fields $(U_t(x))_{t\geq 0}$.
(iii) The chains lie in the general continuous state space $\mathbb{R}^d$, in contrast to the typical discrete ones.

For a formal proof of the above claims, we direct interested readers to our appendix. In parallel to the study of Markov chains in a discrete finite state space, our analysis in the continuous state space primarily focuses on three fundamental properties: *irreducibility*, *aperiodicity*, and *recurrence* [34]. Building on these three properties, we establish limit theorems that shed light on the long-run behavior of the chains. The forthcoming sections aim to grapple with the amplified challenges that arise due to our chain trajectories navigating through multi-dimensional, uncountable domains.

3.1 Convergence up to constant factors

We begin by deriving a basic convergence result that resembles the classical descent inequalities. This result serves as a robust tool for understanding the recurrent behavior of our chains.

As established in prior work [4, 18] and highlighted in the introduction, when the operator V is Lipschitz and strongly monotone, the full-information/noiseless equivalent of SGDA/SEG attain exponential rate of convergence to some solution in the solution set $\mathcal{X}^*$. By relaxing the assumption of strong monotonicity to the assumption of weakly quasi strong monotonicity (Assumption 2), we show that this result can be achieved in the noisy setting as well up to an additive constant. The cornerstone of our proof hinges on the construction of a quasi-descent inequality [30] and the appropriate determination of a step-size in order to account for both the variance σ^2 and the shift λ of weakly quasi-monotonicity. The additive constant factor corresponds to the bias introduced by the stochasticity and non-monotonicity of V, and it depends on the constant step-sizes $\gamma^{\text{SGDA,SEG}}, \alpha^{\text{SEG}}$ used in the respective algorithms.

Formally, the following theorem holds:

Theorem 1. *Consider that either* (SGDA) *or* (SEG) *is run with a stochastic oracle satisfying Assumptions 1–4 respectively with step-sizes* $\gamma^{\text{SGDA}} < \frac{\mu}{L^2}$, $\gamma^{\text{SEG}} < \frac{1}{2\mu + \sqrt{3}\ell}$ *and* $\alpha^{\text{SEG}} \in (0,1)$ *and let* $(x_t)_{t\geq 0}$ *be the iterations generated. Then, there exists a pair of constants*[1] $(c_1, c_2)^{\{\text{SGDA,SEG}\}}$ *that depend on the choice of step-sizes, as well as the parameters of the model, with* $c_1^{\{\text{SGDA,SEG}\}} \in (0,1)$ *and* $c_2^{\{\text{SGDA,SEG}\}} \in (0,+\infty)$ *such that*

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq \left(1 - c_1^{\{\text{SGDA,SEG}\}}\right)^t \|x_0 - x^*\|^2 + c_2^{\{\text{SGDA,SEG}\}}, \tag{6}$$

for any initial point $x_0 \in \mathbb{R}^d$.

A byproduct of the above theorem's proof is the following one-step "quasi-descent" inequality:

Corollary 1. *Under the conditions of Theorem 1, for all* $x^* \in \mathcal{X}^*$ *there exists an extended real-valued function* $\mathcal{E} : \mathbb{R}^d \to [1, \infty]$ *and constants* $c_1^{\{\text{SGDA,SEG}\}} \in (0,1), c_2^{\{\text{SGDA,SEG}\}} \in (0,\infty)$ *such that*

$$\mathbb{E}[\mathcal{E}(x_{t+1}, x^*) \mid \mathcal{F}_t] \leq c_1^{\{\text{SGDA,SEG}\}} \mathcal{E}(x_t, x^*) + c_2^{\{\text{SGDA,SEG}\}}. \tag{7}$$

Specifically, $\mathcal{E}(x_t, x^*) = \|x_t - x^*\|^2 + 1$.

Remark. The function $\mathcal{E}$ is sometimes called an energy, potential or Lyapunov function. While the above corollary applies to any $x^* \in \mathcal{X}^*$, for the sake of conciseness, we will assume a fixed but arbitrary x^* and omit its reference. From now on, we will simply write the energy function as $\mathcal{E}(x_t)$.

[1] For the explicit formula of the constants, we refer the reader to the proof at the supplement.

Understanding Markov chains in continuous domains requires a grasp of different types of recurrences: *(null)-recurrence*, *Harris recurrence*, and *positive recurrence*, each progressively contributing to our insights on the chain behavior. Recurrence indicates a state will infinitely visit nearby regions on expectation, but without timing guarantees. Harris recurrence, specific to continuous state space Markov chains, ensures a state revisits the nearby areas infinitely often almost surely. Positive recurrence, an orthogonal refinement, promises a state's recurrent visits within a finite expected time. (For their formal definitions, we refer to our introductory appendix on Markov Chains.)

Harris and positive recurrence are the pivotal properties that underpin our key results on the existence of (a) an invariant measure, (b) a law of large numbers, and (c) an ergodic central limit theorem.

4 Main Results

The main result of this section can be summarized as follows:

Informal Theorem (Main Result)**.** *Under Assumptions 1–4, the Stochastic Extragradient Stochastic Extra Gradient and Stochastic Gradient Descent Ascent Stochastic Gradient Descent Ascent methods with constant step-size, behave as strong aperiodic, positive Harris recurrent continuous-state Markov Chains, converging to a unique stationary distribution over time regardless of the initial conditions. Moreover, their trajectory's ergodic averages adhere to the Law of Large Numbers and the Central Limit Theorem.*

Proof Sketch. Our main objective is to showcase that, under constant step-size, the average trajectory of SEG and SGDA methods converges to a typical path over time, validating their ergodic behavior. This endeavor necessitates the fusion of optimization and probabilistic techniques.

Our investigation commences by observing that both SEG and SGDA methods, when operating under a constant step-size, behave akin to continuous-state Markov Chains within the Euclidean space $\mathbb{R}^d$. To further exploit machinery such as Markov Chain Central Limit Theorems, Richardson extrapolation, etc., our primary objective is to ascertain the existence of an invariant probability measure. We achieve this by establishing properties like strong aperiodicity, positive Harris recurrence, and irreducibility—paralleling the standard approach for finite discrete-state Markov chains. Our proof for these properties leans heavily on a single-step probability minorization condition and arguments based on Lyapunov potential functions. In addition, the application of the SEG method to VIs brings added complexities due to its intricate update rule, contrasting the simpler case of Stochastic Gradient Descent (SGD) used for minimization task.

Focusing on our techniques, we extensively use a version of Doeblin's bound. In words this minorization condition posits that from any state, there's a positive probability that the chain will transition into a designated subset of states within one step. In mathematical terms, for all $x \in S$ and for all measurable subsets $A \subseteq S$ (where S is the state space), there's a positive probability that $P(x, A)$ is at least $\epsilon \cdot \mu(A)$ for some $\epsilon > 0$ and a probability distribution $\mu(\cdot)$. We then construct a coupling for two probability laws: Z_1 distributed according to $\nu(x) \cdot P^n(x, \cdot)$ and Z_2 according to $\pi(x) \cdot P^n(x, \cdot)$, for any arbitrary $x \in S$ and the stationary distribution $\pi(\cdot)$. This guarantees that the total variation distance between the laws of Z_1, Z_2 is bounded by $(1 - \epsilon)^n$ for any ν probability measure.

While in discrete settings we could consider the entire state space, it is not feasible to do so in continuous domains like $\mathbb{R}^d$. We navigate this challenge by applying the minorization condition within a bounded region around the solution set, referred to as $S^* := \mathrm{Ball}(X^*, r^*)$. Such regions are termed "small sets" in the literature of Markov Chains. In the context of Markov Chains literature, such regions are commonly referred to as "small sets". Given a state x that resides within S^*, Doeblin's condition ensures geometric convergence to the invariant probability. To extend this convergence rate to $\mathbb{R}^d$, we employ the Foster-Lyapunov (FL) inequality within a well-tailored small set. FL inequality —also known geometric drift property (See [45])—ensures that the distance from the solution set remains bounded in expectation and diminishes according to a quasi-descent inequality if the current state resides within a judiciously chosen attraction region. Using this inequality, we establish that iterations outside a small set S will converge on expectation to S exponentially fast, suggesting infinite visits to S and affirming geometric convergence to a unique stationary distribution, independent of the initial state.

In order to employ our stochastic analysis toolkit, we embrace the following standard regularity assumption regarding the nature of the noise [50].

Assumption 5. The random variable $U_t(x)$ can be decomposed as $U_t(x) = U_t^a(x) + U_t^b(x)$, such that the probability distribution of $U_t^a(x)$ has a density function, $\mathrm{pdf}_{U_t^a(x)}$, with respect to the Lebesgue measure satisfying $\inf_{x\in C} \mathrm{pdf}_{U_t^a(x)}(t) > 0$ for all bounded sets $C \subseteq \mathbb{R}^d$ and for all $t \in \mathbb{R}^d$.

Regarding the applicability of this assumption, observe that any Gaussian random field, among others, satisfies Assumption 5.

4.1 Minorization Condition, Geometric Drift Property & Recurrence Classification

Inspired by the Markov chain stability framework in [34], we prove two important properties: the *Minorization Condition* and the *Geometric Drift Property*. Both of them serve an important role in proving Harris and Positive Recurrence respectively.

Lemma 1. *Let the assumptions Assumptions 1–5 be satisfied for* (SGDA) *and* (SEG)*. Then given the step-sizes specified in Theorem 1, both algorithms satisfy the following minorization condition: there exist a constant $\delta > 0$, a probability measure ν and a set C dependent on the algorithm, such that $\nu(C) = 1$, $\nu(C^c) = 0$ and*

$$\Pr[x_{t+1} \in A | x_t = x] \geq \delta\, \mathbb{1}_C(x)\nu(A) \quad \textit{for all} \;\; A \in \mathcal{B}(\mathbb{R}^d),\; x \in \mathbb{R}^d. \tag{8}$$

If the set C encompassed the entire space, Eq. (8) would indicate that every subspace of $\mathbb{R}^d$ is reachable from any state. This would lead, through standard coupling arguments, to geometric convergence of the distribution of x_t towards a unique distribution. Although this scenario may not hold in our unbounded state space, a subset C that satisfies this condition, known as a "small/petite" set, can still ensure geometric convergence if a Foster-Lyapunov drift property is satisfied.

Corollary 2. *Under the setting of Lemma 1, the function $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}$ presented in Corollary 1 satisfies the following geometric drift property by* (SGDA) *or* (SEG)*: there exists a measurable set C, and constants $\beta > 0$, $b < \infty$ such that*

$$\Delta\mathcal{E}(x) \leq -\beta\mathcal{E}(x) + b\,\mathbb{1}_C(x), x \in \mathbb{R}^d, \tag{9}$$

where $\Delta\mathcal{E}(x) = \int_{y\in\mathbb{R}^d} P(z, dy)\mathcal{E}(y) - \mathcal{E}(x)$.

The above property is called the (V4) geometric drift property in [34]. In simple terms, the Foster-Lyapunov inequality (9) controls how quickly the energy function decreases as the Markov chain transitions between states. If r.h.s. of (9) is negative, it indicates an exponential rate of decrease, which in turn implies that the chain “forgets” its initial state and exhibiting predictable and stationary behavior around minimum of our energy function $\mathcal{E}(\cdot)$.

Equipped with the Minorization condition and the geometric drift property, we are ready to show all the necessary conditions for proving the ergodicity of (SGDA) and (SEG). Specifically,

Lemma 2. *The Markov chain sequences $(x_t)_{t\geq 0}$ corresponding to* (SGDA) *and* (SEG) *have the following properties:*

- *They are $\psi-$irreducible for some non-zero σ-finite measure ψ on $\mathbb{R}^d$ over Borel σ- algebra of $\mathbb{R}^d$.*
- *They are aperiodic.*
- *They are Harris and positive recurrent with an invariant measure.*

Thus using generalizations of aperiodic ergodic theorem for Markov chains satisfying the geometric drift property, we prove our first main result about the invariance measure. In the following, we let $\mathcal{P}_2(\mathbb{R}^d) := \{\nu : \int_{\mathbb{R}^d} \|x\|^2\nu(dx) < \infty\}$ denote the set of square-integrable probability measures.

4.2 Invariant Measure, Law of Large Numbers & Central Limit Theorem

Theorem 2. *Let Assumptions 1–5 be satisfied for (SGDA) and (SEG). Then given the step-sizes specified in Theorem 1, it holds that*

1. *(SGDA) and (SEG) iterates admit a unique stationary distribution* $\pi_\gamma^{\{\text{SGDA,SEG}\}} \in \mathcal{P}_2(\mathbb{R}^d)$.
2. *For each test function* $\phi : \mathbb{R}^d \to \mathbb{R}$ *satisfying that* $|\phi(x)| \le L_\phi(1 + \|x\|)$ *for all* $x \in \mathbb{R}^d$ *and some* $L_\phi > 0$ *and for any initialization* $x_0 \in \mathbb{R}^d$, *there exist* $\rho_{\phi,\gamma}^{\{\text{SGDA,SEG}\}} \in (0,1)$ *and* $\kappa_{\phi,x_0,\gamma}^{\{\text{SGDA,SEG}\}} \in (0,\infty)$ *such that:*
$$\left| \mathbb{E}_{x_t}[\phi(x_t)] - \mathbb{E}_{x \sim \pi_\gamma^{\{\text{SGDA,SEG}\}}}[\phi(x)] \right| \le \kappa_{\phi,x_0,\gamma}^{\{\text{SGDA,SEG}\}} (\rho_{\phi,\gamma}^{\{\text{SGDA,SEG}\}})^t. \tag{10}$$
Hence, (SGDA) and (SEG) converges geometrically under the total variation distance to $\pi_\gamma^{\{\text{SGDA,SEG}\}}$.
3. *For each test function* ϕ *that is* ℓ_ϕ*-Lipschitz, it holds that*
$$|\mathbb{E}_{x \sim \pi_\gamma^{\{\text{SGDA,SEG}\}}}[\phi(x)] - \phi(x^*)| \le \ell_\phi \sqrt{D^{\{\text{SGDA,SEG}\}}}, \tag{11}$$
for some constant $D^{\{\text{SGDA,SEG}\}} \propto \max(\lambda, \gamma^{\text{SGDA,SEG}})/\mu$.

The result outlined above provides critical insights into the behavior of constant step size Stochastic Extragradient Stochastic Extra Gradient and Stochastic Gradient Descent Ascent Stochastic Gradient Descent Ascent methods. Notably, it asserts the uniqueness of the stationary distribution of these methods, assuming it has a bounded second moment. It further offers an analysis of the fluctuation patterns of a test function ϕ across the Stochastic Extra Gradient/Stochastic Gradient Descent Ascent iterations, even in the face of non-smooth and non-convex objective functions. Elaborating on the convergence properties, the theorem elucidates that the Stochastic Extra Gradient/Stochastic Gradient Descent Ascent algorithm, irrespective of its initial point and provided the step size is suitably small, will gravitate towards its invariant distribution at an exponential rate (See Eq. (10)). This effectively confirms the robustness of these algorithms under various initialization scenarios and across a wide spectrum of step sizes. Lastly, for the class of smooth test functions, (See Eq. (11)) the above result constrains the deviation of the expected value of the test function's asymptotic behavior from its optimal value, offering an explicit bound. This bound delineates a 'ball of interest', providing a tangible limit to the bias, thus enhancing our understanding of the overall performance of these algorithms.

Following the influential work of Polyak and Juditcky [41], and having confirmed the uniqueness of the stationary distribution, we now focuses on the question of asymptotic normality of the two algorithms. To the best of our knowledge, such a result would be the first of its kind for stochastic approximation methods within the variational inequalities framework, especially for extrapolation techniques like (SEG). Establishing such results allows us to provide theoretical guarantees when constructing confidence intervals in game scenarios, surpassing the sole dependence on empirical evidence, i.e., [1, 22, Section 7]. To streamline our discussion, let us introduce a notation for any given function ϕ:

Definition 1. We denote the average iterate of our methods, also known as the Césaro mean [20], evaluated over a given function ϕ as $\overline{S_T(\phi)} := \frac{1}{T} S_T(\phi) := \frac{1}{T} \sum_{t=0}^{T} \phi(x_t)$.

Our inquiry begins with establishing a Law of Large Numbers (LLN) for (SGDA) and (SEG). By employing the analogue of the Birkhoff–Khinchin ergodic theorem for continuous state space ergodic Markov Chains, we can derive the ensuing LLN:

Theorem 3. *Let the Assumptions 1–5 hold. Then for the choice of step-sizes specified in Theorem 2 and any function* ϕ *satisfying* $\pi_\gamma(|\phi|) < \infty$, *where* $\pi_\gamma(|\phi|) = \mathbb{E}_{x \sim \pi_\gamma^{\{\text{SGDA,SEG}\}}}[|\phi(x)|]$, *it holds that*

$$\lim_{T\to\infty} \frac{1}{T} S_T(\phi) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \phi(x_t) = \pi_\gamma(\phi) \quad \textit{a.s.} \tag{Law of Large Numbers for (SGDA),(SEG)}$$

We next state a central limit theorem (CLT) for the sequences generated by (SGDA) and (SEG), establishing the asymptotic normality of their averaged iterates:

Theorem 4. *Let the Assumptions 1–5 hold. Then for the choice of step-sizes and a test function ϕ specified in Theorem 2, we have that*

$$T^{-1/2} S_T(\phi - \pi_\gamma(\phi)) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{\pi_\gamma}(\phi)), \qquad \text{(Central Limit Theorem for (SGDA),(SEG))}$$

where $\pi_\gamma(\phi) = \mathbb{E}_{x \sim \pi_\gamma^{\{SGDA,SEG\}}}[\phi(x)]$ *and* $\sigma^2_{\pi_\gamma}(\phi) := \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_{\pi_\gamma^{\{SGDA,SEG\}}}[S_T^2(\phi - \pi_\gamma(\phi))]$. *where* $\mathbb{E}_{\pi_\gamma^{\{SGDA,SEG\}}}$ *denotes that the initial distribution of the Markov chain is* $\pi_\gamma^{\{SGDA,SEG\}}$.

5 Applications and Experiments

In this section, we discuss the applications of our main theoretical results. We will focus our examination on two interesting subcategories of quasi-strongly monotone problems: (i) min-max convex-concave games, with locally quadratic region of attractions around the Nash Equilibria and (ii) the application of Richardson-Romberg (RR) bias refinement scheme for smooth quasi-strongly monotone operators. While the region of attraction in the first instance could potentially be an artifact of our analysis, it is noteworthy that the application of RR presupposes the existence of a unique solution to be viable. We conclude the section by presenting a series of experiments validating our theoretical establishments.

5.1 Min-Max Convex-Concave Games

We now explore a specific class of operators that lie in the merely monotone regime:

Assumption 6. we assume that the operator V is monotone in the sense that

$$\langle V(x) - V(x'), x - x' \rangle \geq 0 \text{ for all } x, x' \in \mathbb{R}^d. \tag{12}$$

Theorem 5. *Let Assumptions 1–6 hold. Then the iterates of* (SGDA), (SEG), *when run with the step-sizes given in Theorem 1, admit a stationary distribution* $\pi_\gamma^{\{SGDA,SEG\}}$ *such that*

$$\mathbb{E}_{x\sim\pi_\gamma^{\{SGDA,SEG\}}}[\mathrm{Gap}_V(x)] \leq c\gamma^{SGDA,SEG}, \tag{13}$$

where $\mathrm{Gap}_V(x)$ *is the restricted merit function* $\mathrm{Gap}_V(x) := \sup_{x^*\in\mathcal{X}^*}\langle V(x), x - x^*\rangle$ *and* $c \in \mathbb{R}$ *is a constant and depends on the parameters of the problem.*

For the particular case of convex-concave min-max games, the standard notion of *duality gap*, also known as *primal-dual optimality gap* or *Nash gap* defined as $\text{Duality-Gap}_f(\theta, \phi) = \max_{\phi' \in \mathbb{R}^{d_2}} f(\theta, \phi') - \min_{\theta' \in \mathbb{R}^{d_1}} f(\theta', \phi)$, is upper bounded by the aforementioned $\mathrm{Gap}_V(x)$. Here, $x = (\theta, \phi)$, $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is a convex function with respect to the first argument and concave with respect to the second one, and $V = (\nabla_\theta f, -\nabla_\phi f)$ as in Example 2.3.

Consequently, let $\mathrm{val}^* = \min_{\theta\in\mathbb{R}^{d_1}} \max_{\phi\in\mathbb{R}^{d_2}} f(\theta, \phi)$ denote the value of this convex-concave game. Then, for the unique stationary distribution $\pi_\gamma^{\{\mathrm{SGDA,SEG}\}}$ of the iterates of (SGDA) and (SEG), we have

$$|\mathbb{E}_{(\theta,\phi)\sim\pi_\gamma^{\{\mathrm{SGDA,SEG}\}}}[f(\theta, \phi)] - \mathrm{val}^*| \leq c\gamma^{\mathrm{SGDA,SEG}}. \tag{14}$$

From (13) and (14), we see that in this class of monotone games, (SGDA) and (SEG) converge to val^* –the unique value of the corresponding game at a Nash Equilibrium –within an expected error that is proportional to the stepsize $\gamma^{\mathrm{SGDA,SEG}}$, where the error is measured by the duality gap or the difference in the game value.

5.2 Bias Refinement in Quasi-Monotone Operators

Here we focus on the case of quasi-monotone operators (i.e., $\lambda = 0$ in Assumption 2), which encompasses a variety of non-monotone and non-convex optimization problems. In this regime, we provide a refined analysis of the stationary distribution induced by (SGDA) under some smoothness assumptions for the operator and the nature of noise. Specifically, we provide an explicit expansion of the steady-state expectation in terms of the stepsize, which allows us to employ the Richardson-Romberg (RR) bias refinement scheme [14] to construct a new estimate provably closer to the optimal solution. Our result is a strict generalization of [10], which requires co-coersive noisy first-order oracles.

Assumption 7. The operator V is ℓ-Lipschitz and $C^4(\mathbb{R}^d)$-smooth (i.e., $\sup_{x\in\mathbb{R}^d}\|\nabla^i V(x)\| < \infty$ for all $i = 1,\dots,4$). Furthermore, the noise has bounded kyrtosis, meaning that $\mathbb{E}[\|U_t(x)\|^4] < \delta^4_{\text{KYRT}}$ for all $x \in \mathbb{R}^d$ with the covariance tensor $x \mapsto \mathcal{C}(x) := \mathbb{E}[U_t(x)^{\otimes 2}]$ being 3 times smoothly differentiable, meaning $\|\mathcal{C}^{(i)}(x)\| < G, \forall x$, for $i \in \{1,2,3\}$.

Theorem 6. *Suppose Assumptions 1–5 and 7 hold. There exists a threshold θ such that if $\gamma \in (0,\theta)$, then (SGDA) admits a unique stationary distribution π_γ and*

$$\mathbb{E}_{x\sim\pi_\gamma}[x] - x^* = \gamma\Delta(x^*) + \mathcal{O}(\gamma^2), \tag{15}$$

where $\Delta(x^*)$ is a vector independent of the choice of step-size γ.

Note that Eq. (15) is an equality (up to a second order term). In the setting of Theorem 6, this equality gives a more precise characterization of the bias than the upper bound (11) applied to $\phi(x) = x$.

An immediate implication of Theorem 6 is that one can use the following RR refinement scheme to obtain a better estimate of x^*. Consider running two (SGDA) recursions with step-size γ and 2γ and denote the corresponding averaged iterates by $(\bar{x}_t^{\gamma})_{t\geq 0}$ and $(\bar{x}_t^{2\gamma})_{t\geq 0}$, respectively. Let us denote by π_γ and $\pi_{2\gamma}$ the resulting unique stationary distributions. By our result on LLN (cf. Theorem 3), the averaged iterates $(\bar{x}_t^{\gamma})_{t\geq 0}$ and $(\bar{x}_t^{2\gamma})_{t\geq 0}$ converges to $\mathbb{E}_{x\sim\pi_\gamma}[x]$ and $\mathbb{E}_{y\sim\pi_{2\gamma}}[y]$, respectively. Note that Eq. (15) implies that

$$\big(\mathbb{E}_{x\sim\pi_\gamma}[2x] - \mathbb{E}_{y\sim\pi_{2\gamma}}[y]\big) - x^* = \mathcal{O}(\gamma^2).$$

Therefore, the RR refinement of the averaged iterates, $(2\bar{x}_t^{\gamma} - \bar{x}_t^{2\gamma})_{t\geq 0}$, converge to a limit that is closer to the optimal solution x^* by a factor of γ.

5.3 Experiments

We conduct a series of experiments to empirically observe and validate our results. We focus on strongly convex-concave games with two players, for which we have adapted the code of the repository of [22]. In particular, for the first two sets of experiments (Figs. 2–4), we consider a strongly convex-concave min-max game, $\min_{x_1\in\mathbb{R}^d}\max_{x_2\in\mathbb{R}^d} f(x_1,x_2)$, with $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ given by

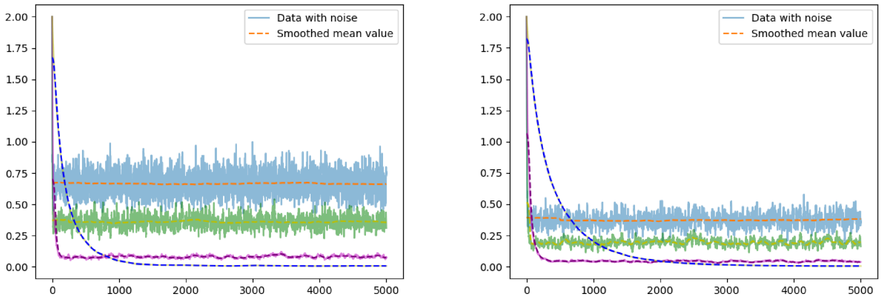
$$f(x_1,x_2) = x_1^\top A_1 x_1 - x_2^\top A_2 x_2 + (x_1^\top B_1 x_1)^2 - (x_2^\top B_2 x_2)^2 + x_1^\top C x_2,$$

where $d = 50$, each of $A_1, A_2, B_1, B_2 \in \mathbb{R}^{d\times d}$ is a random positive definite matrix, and C is a random matrix. Note that the global solution of the game is $x^* = (x_1^*, x_2^*) = (0,0)$ with value $f(x_1^*, x_2^*) = 0$. The operator associated with the above game is

$$V(x) = V((x_1,x_2)) = (\nabla_{x_1} f(x_1,x_2), -\nabla_{x_2} f(x_1,x_2)).$$

The stochastic oracle outputs $V(x) + Z$, where $Z \sim \mathcal{N}(0,\sigma^2 I)$ is Gaussian noise with $\sigma = 0.5$.

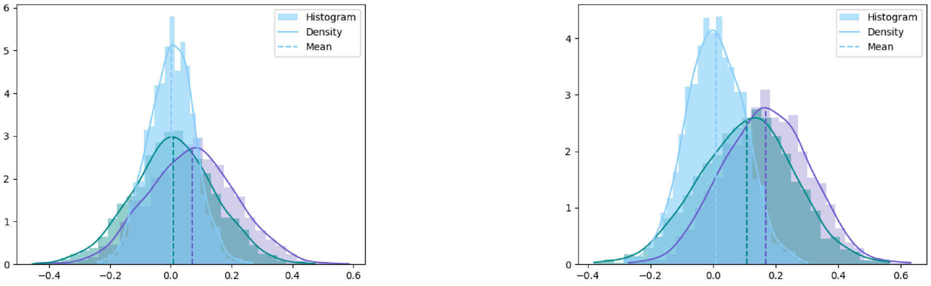
We started by plotting in Figs. 2a and 2b the squared error $\|x_t - x^*\|^2$ for (SGDA) and (SEG) for step-sizes $\gamma \in \{0.1, 0.05, 0.01, 0.001\}$, corresponding to the four curves from top to bottom; the parameter α^{SEG} for (SEG) is set to 0.5. We observe a decay of the steady-state error as a function of the step-size. In fact, the decay



(a) SGDA.

(b) SEG.

Figure 2: Convergence and squared error under different step-sizes for SGDA and SEG.



(a) SGDA

(b) SEG

Figure 3: Results for 100 (light purple), 200 (light green), 1000 (light blue) iterations (or from right to left).

is almost linear for both algorithms, which is consistent with our theoretical bound (11) applied to the test function $\phi(x) = \|x - x^*\|$.

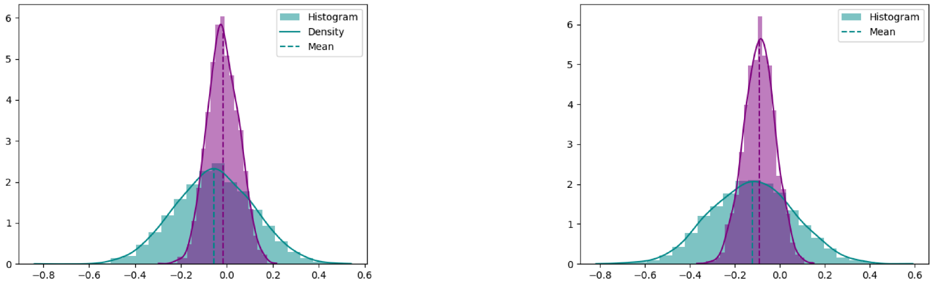
The second set of experiments examines the central limit theorem (CLT). We use as a test function the value of the game $f(x_t)$ evaluate at the iterate, and we observe the behavior of its averaged evaluations after $100, 200$ and 1000 iterations. To do so we run both algorithms with step-size $\gamma = 0.005$ for the aforementioned number of iterations and keep the sum of the evaluations, normalized with $\sqrt{\text{iterations}}$. We repeat this experiment 2000 times and report the histograms in Fig. 3. We observe how the distributions are concentrated closer to the actual value of the game (which is zero) as the number of iterations is increased. In Fig. 4 we run both algorithms in the previous setting for 1000 iterations and two different step-sizes 0.1 and 0.001. We observe how the histogram is concentrated closer to the actual value of the game for smaller step-size.

Lastly, to investigate the effect of the RR refinement scheme, we perform an experiment on a slightly more complicated game. Define the scalar function $h(z) := \log(1 + e^z)$, which is convex. Consider a strongly convex-concave min-max game with $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ given by

$$f(x_1, x_2) = h(x_1) + h(-2x_1) - h(x_2) - h(-2x_2) + 0.1x_1^2 - 0.1x_2^2 + 0.1x_1x_2.$$

The operator V and the stochastic oracle are defined in the same way as before. The global solution of this game is $x^* = (x_1^*, x_2^*) \approx (0.3268, 0.3801)$.

We run the (SGDA) algorithm with two different step-sizes γ and 2γ, where $\gamma = 0.1$. In Fig. 5, we plot the error $\|\bar{x}_t - x^*\|^2$ of the averaged iterate $\bar{x}_t := \frac{1}{t}\sum_{i=1}^t x_i$ with the two stepsizes, as well as that of the RR refinement scheme (cf. Section 5.2). The error achieved by the RR refinement is an order of magnitude better than vanilla (SGDA). This is consistent with the bias reduction effect predicted by our theoretical result in Section 5.2.



(a) SGDA

(b) SEG

Figure 4: Histograms for two different step-sizes. Green: $\gamma = 0.1$. Purple: $\gamma = 0.001$.

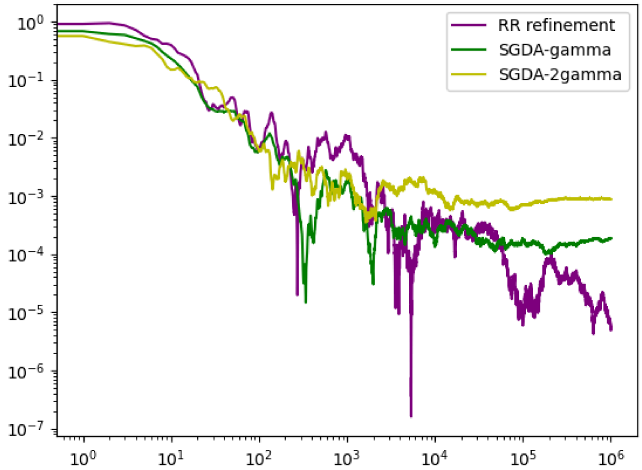


Figure 5: Errors of the average iterates of SGDA and RR refinement.

6 Concluding remarks

In this work, we delve into the probabilistic structures inherent in Stochastic Extragradient and Stochastic Gradient Descent Ascent algorithms, widely used in min-max optimization and variational inequalities problems. By treating constant step-size variants of SEG/SGDA as time-homogeneous Markov Chains, we establish a Law of Large Numbers and a Central Limit Theorem, revealing the existence of a unique invariant distribution and the asymptotic normality of the averaged iterate. For a wide class of convex-concave games, we characterize the intrinsic bias of these methods w.r.t. the game's value. Lastly, we demonstrate that the Richardson-Romberg refinement scheme enhances the proximity of the averaged iterate to the global solution for quasi-monotone variational inequalities.

As a result of this study, several intriguing open questions arise. The extension of Markovian analysis to broader operator families, and their potential applications in statistical inference, adversarial training, and robust machine learning present exciting research opportunities. Investigating how the methods used in this study can be applied to other established optimization algorithms, such as Optimistic Gradient Descent Ascent, which requires higher-order Markov process analysis, is another promising line of research. Exploring different geometries and studying robustness in reinforcement learning also offer interesting prospects.

Acknowledgments

Emmanouil V. Vlatakis-Gkaragkounis is grateful for financial support by the Post-Doctoral FODSI-Simons Fellowship. Yudong Chen is partially supported by NSF grants CCF-1704828 and CCF-2233152. Qiaomin Xie is supported in part by NSF grant CNS-1955997. This project started when EVVG, YC and QX were attending the Data-Driven Decision Processes program of the Simons Institute for the Theory of Computing.

References

[1] Antonakopoulos, K., Belmega, E. V., and Mertikopoulos, P. Adaptive extra-gradient methods for min-max optimization and games. In *ICLR '21: Proceedings of the 2021 International Conference on Learning Representations*, 2021.

[2] Benveniste, A., Metivier, M., and Priouret, P. *Adaptive Algorithms and Stochastic Approximations*. Springer Berlin Heidelberg, 1st edition, 2012. ISBN 9783642758942. doi: 10.1007/978-3-642-75894-2.

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

[4] Beznosikov, A., Gorbunov, E., Berard, H., and Loizou, N. Stochastic gradient descent-ascent: Unified theory and new efficient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 172–235. PMLR, 2023.

[5] Bianchi, P., Hachem, W., and Schechtman, S. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 30(3):1117–1147, 2022.

[6] Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. In *International Conference on Learning Representations (ICLR 2018)*, 2018.

[8] Daskalakis, C., Skoulakis, S., and Zampetakis, M. The complexity of constrained min-max optimization. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1466–1478, 2021.

[9] Diakonikolas, J., Daskalakis, C., and Jordan, M. I. Efficient methods for structured nonconvex-nonconcave min-max optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2746–2754. PMLR, 2021.

[10] Dieuleveut, A., Durmus, A., and Bach, F. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348 – 1382, 2020. doi: 10.1214/19-AOS1850. URL https://doi.org/10.1214/19-AOS1850.

[11] Douc, R., Moulines, E., Priouret, P., and Soulier, P. *Markov Chains*. Springer Cham, 1st edition, 2018. ISBN 9783319977041 (online). doi: https://doi.org/10.1007/978-3-319-97704-1.

[12] Durmus, A., Jiménez, P., Moulines, É., and Salem, S. On riemannian stochastic approximation schemes with fixed step-size. In *International Conference on Artificial Intelligence and Statistics*, pp. 1018–1026. PMLR, 2021.

[13] Erdogdu, M. A., Mackey, L., and Shamir, O. Global non-convex optimization with discretized diffusions. *Advances in Neural Information Processing Systems*, 31, 2018.

[14] Gautschi, W. *Numerical analysis*. Springer Science & Business Media, 2011.

[15] Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

[16] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[17] Gorbunov, E., Hanzely, F., and Richtárik, P. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pp. 680–690. PMLR, 2020.

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

[19] Hao, W. A gradient descent method for solving a system of nonlinear equations. *Appl. Math. Lett.*, 112:106739, 2021. doi: 10.1016/j.aml.2020.106739. URL https://doi.org/10.1016/j.aml.2020.106739.

[20] Hardy, G. and Series, D. Providence. *American Mathematical Society*, 1992.

[21] Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[22] Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *NeurIPS '20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[23] Huo, D. L., Chen, Y., and Xie, Q. Bias and extrapolation in markovian linear stochastic approximation with constant stepsizes, 2022. URL https://arxiv.org/abs/2210.00953.

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

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

[26] Korpelevich, G. M. The extragradient method for finding saddle points and other problems. *Èkonom. i Mat. Metody*, 12:747–756, 1976.

[27] Kushner, H. J. and Yin, G. G. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer, New York, NY, USA, 2nd edition, 2003. ISBN 9780387008943. doi: 10.1007/b97441. URL https://link.springer.com/book/10.1007/b97441.

[28] Lan, G. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Series in the Data Sciences. Springer International Publishing, 2020. ISBN 9783030395681. URL https://books.google.com/books?id=7dTkDwAAQBAJ.

[29] Lin, T., Zhou, Z., Mertikopoulos, P., and Jordan, M. Finite-time last-iterate convergence for multi-agent learning in games. In *International Conference on Machine Learning*, pp. 6161–6171. PMLR, 2020.

[30] Loizou, N., Berard, H., Jolicoeur-Martineau, A., Vincent, P., Lacoste-Julien, S., and Mitliagkas, I. Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pp. 6370–6381. PMLR, 2020.

[31] Loizou, N., Berard, H., Gidel, G., Mitliagkas, I., and Lacoste-Julien, S. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021.

[32] Mertikopoulos, P. and Zhou, Z. Learning in games with continuous action sets and unknown payoff functions. *Mathematical Programming*, 173:465–507, 2019.

[33] Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *ICLR 2019-7th International Conference on Learning Representations*, pp. 1–23, 2019.

[34] Meyn, S. P. and Tweedie, R. L. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2nd edition, 2009. ISBN 9780521731829. doi: 10.1017/CBO9780511626630. URL https://doi.org/10.1017/CBO9780511626630.

[35] Mishchenko, K., Kovalev, D., Shulgin, E., Richtárik, P., and Malitsky, Y. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 4573–4582. PMLR, 2020.

[36] Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507. PMLR, 2020.

[37] Nemirovski, A. 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.

[38] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[39] Papadimitriou, C. H., Vlatakis-Gkaragkounis, E., and Zampetakis, M. The computational complexity of multi-player concave games and kakutani fixed points. *CoRR*, abs/2207.07557, 2022. doi: 10.48550/arXiv.2207.07557. URL https://doi.org/10.48550/arXiv.2207.07557.

[40] Pfau, D. and Vinyals, O. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

[41] Polyak, B. T. New stochastic approximation type procedures. *Automation and Remote Control*, 51(7):98–107, Jul 1990.

[42] Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, Jul 1992. ISSN 0363-0129. doi: 10.1137/0330046. URL https://doi.org/10.1137/0330046.

[43] Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703. PMLR, 2017.

[44] Song, C., Zhou, Z., Zhou, Y., Jiang, Y., and Ma, Y. Optimistic dual extrapolation for coherent non-monotone variational inequalities. *Advances in Neural Information Processing Systems*, 33:14303–14314, 2020.

[45] Takimoto, E. and Warmuth, M. K. Path kernels and multiplicative updates. *Journal of Machine Learning Research*, 4(773-818), 2003.

[46] Tan, Y. S. and Vershynin, R. Online stochastic gradient descent with arbitrary initialization solves non-smooth, non-convex phase retrieval. *CoRR*, abs/1910.12837, 2019. URL http://arxiv.org/abs/1910.12837.

[47] Wen, J., Yu, C.-N., and Greiner, R. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, pp. 631–639. PMLR, 2014.

[48] Wikipedia. Wasserstein metric – wikipedia, the free encyclopedia, 2023. URL https://en.wikipedia.org/wiki/Wasserstein_metric. Accessed: 2023-05-22.

[49] Yang, J., Kiyavash, N., and He, N. Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33:1153–1165, 2020.

[50] Yu, L., Balasubramanian, K., Volgushev, S., and Erdogdu, M. A. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4234–4248. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/21ce689121e39821d07d04faab328370-Paper.pdf.

[51] Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

Organization of the appendix

A Background in Continuous-Space Markov Chains

In this preliminary segment, we furnish the basic concepts and tools for studying Markov chains defined on a continuous state space. These results subsequently form the foundational basis for the theorems we establish regarding our algorithms.

A.1 Basic Setup

To explain various concepts for a Markov chain, we first set up our space and identify the events of interest. This process is grounded in the conventional framework of a σ-algebra, which facilitates the comprehension of these events. Formally, we denote the (sub)-σ-algebra of $\mathcal{F}$ of events up to the t-th iteration with $\mathcal{F}_t$ (including the t-th iteration). We denote by $\mathcal{B}(C)$ the σ-algebra of Borel sets of C. We also denote the Markov kernel (Generalized Transition Matrix) on $\mathbb{R}^d$, $\mathcal{B}(\mathbb{R}^d)$ associated either with (SGDA) or (SEG) to be[2]

$$P(x, S) = \mathbb{P}(x_{t+1} \in S | x_t = x) \text{ almost surely } \forall S \in \mathcal{B}(\mathbb{R}^d), \forall x \in \mathbb{R}^d, \forall t \in \mathbb{N}. \tag{A.1}$$

We also define the m-th power of the kernel iteratively: $P^1(x, S) := P(x, S)$ and for $m > 1$, we define

$$P^{m+1}(x, S) = \int_{x' \in \mathbb{R}^d} P(x, dx') P^m(x', S) \text{ for all } x \in \mathbb{R}^d \text{ and } S \in \mathcal{B}(\mathbb{R}^d). \tag{A.2}$$

Additionally, for any function $\phi : \mathbb{R}^d \to \mathbb{R}$ and any $m \geq 1$, we define $P^m \phi : \mathbb{R}^d \to \mathbb{R}$ as

$$P^m \phi(x) = \int_{x' \in \mathbb{R}^d} \phi(x') P^m(x, dx') \text{ for all } x \in \mathbb{R}^d. \tag{A.3}$$

Definition A.1 (Time-homogeneous)**.** A stochastic process $\Phi = (\Phi_t)_{t=0}^{\infty}$ is called a time-homogeneous Markov chain with transition probability kernel $P(x, A)$ and initial distribution μ if the finite dimensional distributions of Φ satisfy

$$P_\mu(\Phi_0 \in A_0, \Phi_1 \in A_1, \ldots \Phi_n \in A_n) = \int_{y_0 \in A_0} \cdots \int_{y_{n-1} \in A_{n-1}} \mu(dy_0) P(y_0, dy_1) \cdots P(y_{n-1}, A_n) \tag{A.4}$$

for any n and all $A_i \in \mathcal{B}(\mathbb{R}^d)$.

A.2 Irreducibility, Recurrence, and Aperiodicity

Irreducibility.

Definition A.2 ($\psi-$irreducible)**.** A Markov chain is φ-irreducible if there exists a measure φ on $\mathcal{B}(\mathbb{R}^d)$ such that for all $x \in \mathbb{R}^d$ whenever $\varphi(A) > 0$, there exists $n > 0$, possible depending on x, A such that that $P^n(x, A) > 0$. Per convention, we always take φ to be a "maximal" irreducibility measure, denoted by ψ, and say that the chain is $\psi-$irreducible.

For this definition we combine Proposition 4.2.1 and Proposition 4.2.2 from [34]. Consider a $\psi-$irreducible Markov chain, we use $\mathcal{B}^+(\mathbb{R}^d)$ to denote the set of sets $A \in \mathcal{B}(\mathbb{R}^d)$ such that $\varphi(A) > 0$.

Recurrence.

Definition A.3 (Recurrent)**.** Consider a Markov chain $\Phi = (\Phi_t)_{t=0}^{\infty}$ with transition kernel P. Let $\eta_A := \sum_{t=0}^{\infty} \mathbb{1}\{\Phi_t \in A\}$ for some set A. Assume that Φ is ψ-irreducible, then we say that

- **(null)-Recurrent:** The set A is called recurrent if $\mathbb{E}[\eta_A \,|\, \Phi_0 = x] = \infty$ for all $x \in A$. If every set in $\mathcal{B}^+(\mathbb{R}^d)$ is recurrent then we call Φ recurrent.

[2] It would be clear from the context in which algorithm we refer to. If not we will specify it using subscripts.

- **Positive recurrent:** The set A is called positive if $\limsup_{n\to\infty} P^n(x, A) > 0$ for all $x \in A$. If every set $A \in \mathcal{B}^+(\mathbb{R}^d)$ is positive then Φ is called positive recurrent.
- **Harris recurrent:** The set A is called Harris recurrent if $\mathbb{P}(\eta_A = \infty \mid \Phi_0 = x) = 1$ for all $x \in A$. If every set $A \in \mathcal{B}^+(\mathbb{R}^d)$ is Harris recurrent, then Φ is called Harris recurrent.

Aperiodicity.

Definition A.4 (Strongly Aperiodic). An irreducible chain is called strongly aperiodic if there exists a set A, such that there exists a non-trivial measure ν_1 on $\mathcal{B}(\mathbb{R}^d)$ satisfying $\nu_1(A) > 0$, and for all $x \in A$ and $S \in \mathcal{B}(\mathbb{R}^d)$,

$$P(x, S) \geq \nu_1(S). \tag{A.5}$$

Looking at the bigger picture and drawing insight from traditional discrete space Markov chains, if we make a selection such that $S \leftarrow A$, then we achieve $P(x, A) \geq \nu_1(A) > 0$. This suggests that the set A is associated with a self-loop, as it has a positive probability of returning to itself.

A.3 Small Sets, Petite Sets, and Minorization Condition

We next introduce several concepts that pave the way for systematically and efficiently establishing the convergence rate of a Markov chain, other than in an ad-hoc manner.

We first introduce the Minorization Condition. Using this condition is similar in a way as thinking about coupling.

Definition A.5 (Minorization Condition). For some $\delta > 0$, some $C \in \mathcal{B}(X)$ and some probability measure ν with $\nu(C^c) = 0$ and $\nu(C) = 1$:

$$P(x, A) \geq \delta \mathbb{1}_C(x)\nu(A) \text{ for all } A \in \mathcal{B}(\mathbb{R}^d), x \in \mathbb{R}^d. \tag{A.6}$$

If C was the entire $\mathbb{R}^d$, the condition requires every state in the state space to be within reach of any other state. We could then minorize the transition probability with a density ν scaled by a parameter δ. This is equivalent to finding a sliver of a probability distribution where all the transition probabilities "overlap" with each other; see Figure 6 for an illustration. However, in continuous spaces having $C = \mathbb{R}^d$ is usually impossible. The set where such a condition holds is called "small".

Definition A.6 (Small Sets). A set $C \in \mathcal{B}(\mathbb{R}^d)$ is called a small set if there exists an $m \in \mathbb{N}_+$ and a non-trivial measure ν_m on $\mathcal{B}(\mathbb{R}^d)$ such that for all $x \in C$, $B \in \mathcal{B}(\mathbb{R}^d)$,

$$P^m(x, B) \geq \nu_m(B) \tag{A.7}$$

The set C is called ν_m-small.

Let $a = \{a(n)\}$ be a distribution or probability measure on $\mathbb{N}_+$ and consider the associated Markov chain Φ_a with probability transition kernel

$$K_a := \sum_{n=0}^{\infty} P^n(x, A)a(n) \; x \in \mathbb{R}^d, A \in \mathcal{B}(\mathbb{R}^d).$$

Φ_a is called the K_a-chain with sampling distribution a. We can interpret Φ_a as the chain Φ sampled in points according to the distribution a. When $a = \delta_m$ is the Dirac measure with $\delta_m(m) = 1$, then the K_{δ_m}-chain is called the m-skeleton with transitional kernel P^m. With this at hand we define below the petite sets.

Definition A.7 (Petite Sets). We will call a set $C \in \mathcal{B}(\mathbb{R}^d)$ ν_a-petite if the sampled chain satisfies the bound

$$K_a(x, B) \geq \nu_a(B) \tag{A.8}$$

for all $x \in C$, $B \in \mathcal{B}(\mathbb{R}^d)$, where ν_a is a non-trivial measure on $\mathcal{B}(\mathbb{R}^d)$.

Proposition A.1 (Proposition 5.5.3 in [34]). *If a set $C \in \mathcal{B}(\mathbb{R}^d)$ is ν_m-small then it is ν_{δ_m}-petite for some $\delta_m > 0$.*

A.4 Foster-Lyapunov Arguments

Given that only small sets can be found in our setting, in order to prove geometric convergence to a unique stationary distribution we will leverage the generalized version of Foster-Lyapunov condition, dubbed as (V4) in the cited book [34].

The following theorem gives a sufficient criterion for the positive recurrence and existence of an invariant distribution of a Markov chain in terms of a Lyapunov function V. Intuitively, the value $V(x)$ for any state x attained by Markov chain denotes "energy" or "potential" of that state. The idea is that if the mean energy decreases for all but some small set, the Markov chain keeps returning to level-sets close to minimum of the energy. That is, the Markov chain is positive recurrent.

Definition A.8 (Geometric Drift Property)**.** There exists an extended-real valued function $f : \mathbb{R}^d \to [1, \infty]$, a measurable set C, and constants $\beta > 0$, $b < \infty$ such that

$$\Delta f(x) \leq -\beta f(x) + b\,\mathbb{1}_C(x), x \in \mathbb{R}^d, \tag{A.9}$$

where $\Delta f(x) = \int_{y \in \mathbb{R}^d} P(z, dy) f(y) - f(x)$.

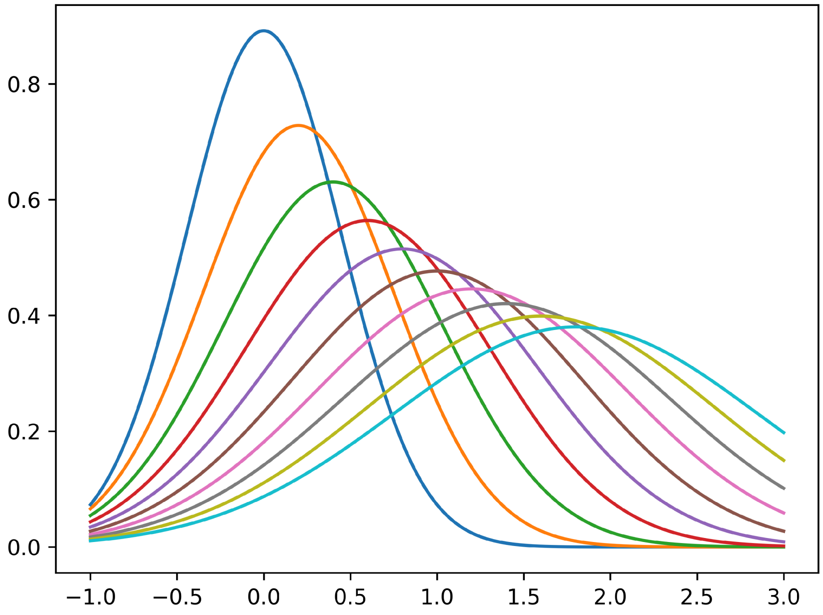


Figure 6: Example of transition kernel $P(x, C)$ for $x \in \mathbb{R}^d$

B Omitted Proofs of Section 3

B.1 (SGDA) and (SEG) are time-homogeneous Markov chains in $\mathbb{R}^d$

Lemma B.1. *Given a constant step-size, the stochastic gradient descent ascent and stochastic extra-gradient as described by Equation (SGDA) and (SEG) can be equivalently modeled as a time-homogeneous continuous Markov chain in $\mathbb{R}^d$.*

Proof. We start with (SGDA) simple case:

$$x_{t+1} = x_t - \gamma^{\text{SGDA}} V_t = x_t - \gamma^{\text{SGDA}}(V(x_t) + U_t(x_t)). \tag{SGDA}$$

By this definition we get that

$$\begin{aligned} P(x,B) &= \mathbb{P}(x_{t+1} \in B | x_t = x) \\ &= \mathbb{P}(x_t - \gamma(V(x_t) + U_t(x_t)) \in B | x_t = x) \\ &= \mathbb{P}(x - \gamma(V(x) + U_t(x)) \in B) \\ &= \mathbb{P}\Big(U(x) \in (\frac{x}{\gamma} - V(x)) + (-\frac{1}{\gamma}B)\Big), \end{aligned}$$

where $(U_t(x))_{t\geq 0} \sim U(x)$, since we assume i.i.d noise random fields. Hence, $P(x,B)$ is shown to be independent of both time t and preceding iterations, given the current state. This affirms that the stochastic gradient descent model described by Equation (SGDA) indeed exhibits the property of a time-homogeneity, substantiating its classification as a Markov chain.

For the case of (SEG), an equivalent form which will come at hand throughout our analysis is given below

$$\begin{aligned} x_{t+1} &= x_t - \alpha^{\text{SEG}}\gamma^{\text{SEG}} V_{t+1/2} \\ &= x_t - \alpha^{\text{SEG}}\gamma^{\text{SEG}} V_{t+1/2} \\ &= x_t - \alpha^{\text{SEG}}\gamma^{\text{SEG}} V_{t+1/2} \\ &= x_t - \alpha^{\text{SEG}}\gamma^{\text{SEG}}(V(x_{t+1/2}) + U_{t+1/2}(x_{t+1/2})) \\ &= x_t - \alpha^{\text{SEG}}\gamma^{\text{SEG}} V(x_t - \gamma^{\text{SEG}}(V(x_t) + U_t(x_t))) \\ &\quad - \alpha^{\text{SEG}}\gamma^{\text{SEG}} U_{t+1/2}(x_t - \gamma^{\text{SEG}}(V(x_t) + U_t(x_t))). \end{aligned} \tag{B.1}$$

Thus for the transition kernel we get that

$$\begin{aligned} P(x,B) =& \mathbb{P}(x_{t+1} \in B | x_t = x) \\ =& \mathbb{P}(x_t - \alpha\gamma V(x_t - \gamma V(x_t) - \gamma U_t(x_t)) \\ & - \alpha\gamma U_{t+1/2}(x_t - \gamma V(x_t) - \gamma U_t(x_t)) \in B | x_t = x) \\ =& \mathbb{P}(x - \alpha\gamma V(x - \gamma V(x) - \gamma U_t(x)) \\ & - \alpha\gamma U_{t+1/2}(x - \gamma V(x) - \gamma U_t(x)) \in B), \end{aligned}$$

where $U_t(x) \sim \text{law}(U^A(x))$, $U_{t+1/2}(x) \sim \text{law}(U^B(x))$ and $U^A(x) \perp U^B(x)$, identically distributed. Thus,

$$P(x,B) = \int_{\xi \in \mathbb{R}^d} \text{pdf}_{U^A(x)}(\xi)\, \mathbb{P}\Big(x - \alpha\gamma V(x - \gamma V(x) - \gamma\xi) - \alpha\gamma U^B(x - \gamma V(x) - \gamma\xi) \in B\Big)\, d\xi.$$

So again, $P(x,B)$ is shown to be independent of both time t and preceding iterations, given the current state. This affirms that the stochastic gradient descent model described by Equation (SGDA) indeed exhibits the property of a time-homogeneity, substantiating its classification as a Markov chain, completing the proof for the case of (SEG). ■

B.2 Geometric convergence up to constant factor

Fact 1. Let $a, b, c \in \mathbb{R}^d$, then the following holds

$$\|a+b+c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2). \tag{B.2}$$

We split Theorem 1 into two different lemmas for each of the algorithms. We start by presenting Eq. (SGDA).

Lemma B.2. *Suppose that Assumptions 1–4 hold then the iterations $(x_t)_{t\geq 0}$ of (SGDA), if the step-size is $\gamma < \frac{\mu}{L^2}$, satisfy:*

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 \mid \mathcal{F}_t] \leq (1-c)^t \|x_0 - x^*\|^2 + c'$$

for some constants $c \in (0,1)$ and $c' \in (0, +\infty)$ that depend on the choice of step-size, as well as the parameters of the problem.

Proof. For simplicity, we drop the exponent SGDA of the step-size and we write γ for the constant step-size used while the algorithm is run. We now start by writing

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &= \|x_t - \gamma V_t - x^*\|^2 \\ &= \|x_t - x^*\|^2 - 2\gamma\langle V_t, x_t - x^*\rangle + \gamma^2\|V_t\|^2.\end{aligned} \tag{B.3}$$

By taking the expectation condition on the filtration $\mathcal{F}_t$, we have that

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x^*\|^2 \mid \mathcal{F}_t] &= \|x_t - x^*\|^2 - 2\gamma\langle V(x_t), x_t - x^*\rangle + \gamma^2\,\mathbb{E}[\|V_t\|^2 \mid \mathcal{F}_t] \\ &= \|x_t - x^*\|^2 - 2\gamma\langle V(x_t), x_t - x^*\rangle + \gamma^2\,\mathbb{E}[\|V(x_t)\|^2] \\ &\quad + \gamma^2\,\mathbb{E}[\|U_t(x_t)\|^2 \mid \mathcal{F}_t],\end{aligned} \tag{B.4}$$

since x_t is $\mathcal{F}_t-$measurable and $\mathbb{E}[V_t \mid \mathcal{F}_t] = V(x_t)$. By Assumption 4 we have that

$$\mathbb{E}[\|U_t(x_t)\|^2 \mid \mathcal{F}_t] \leq \sigma^2, \tag{B.5}$$

while Assumption 2 implies that

$$-2\gamma\langle V(x_t), x_t - x^*\rangle \leq -2\mu\gamma\|x_t - x^*\|^2 + 2\lambda\gamma. \tag{B.6}$$

Finally, using the assumption that the operator has at most linear growth (Assumption 3) we have that for all $x \in \mathbb{R}^d$,

$$\begin{aligned}&\|V(x)\| \leq L(1 + \|x\|) \leq L(1 + \|x^*\| + \|x - x^*\|) \Rightarrow \\ &\|V(x)\|^2 \leq L^2(1 + \mathrm{R} + \|x - x^*\|)^2 \\ &\qquad\quad \leq 2L^2((1+\mathrm{R})^2 + \|x - x^*\|^2).\end{aligned} \tag{B.7}$$

By substituting Eqs. (B.5)–(B.7) to Eq. (B.4), we get that

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 \mid \mathcal{F}_t] \leq (1 - 2\mu\gamma + 2\gamma^2L^2)\|x_t - x^*\|^2 + (2\lambda\gamma + 2\gamma^2L^2(1+\mathrm{R}^2) + \gamma^2\sigma^2). \tag{B.8}$$

Now if

$$\begin{aligned}& 1 - 2\mu\gamma + 2\gamma^2L^2 < 1 \\ \Leftrightarrow\quad & 2\gamma^2L^2 < 2\mu\gamma \\ \Leftrightarrow\quad & 0 < \gamma < \frac{\mu}{L^2},\end{aligned}$$

and by letting $1 - c = 1 - 2\mu\gamma + 2\gamma^2 L^2 < 1$ and $c' = \frac{2\lambda\gamma + 2\gamma^2 L^2(1+\mathrm{R}^2) + \gamma^2\sigma^2}{c}$, we can rewrite Eq. (B.8) as

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 \mid \mathcal{F}_t] \leq (1-c)\|x_0 - x^*\|^2 + cc'.$$

Therefore, we have

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq (1-c)^t\|x_0 - x^*\|^2 + c' \text{ for all } t \geq 0.$$

■

We proceed on proving a similar lemma for the case of (SEG). To do so, we first introduce and analyze two intermediate steps.

Proposition B.1. *Consider that* (SEG) *is run and let* $x^* \in \mathcal{X}^*$, $g_t = V_{t+1/2} = V(x_{t+1/2}) + U_{t+1/2}(x_{t+1/2})$, *where* V, U *satisfy Assumptions 2–4,* $\gamma \in \mathbb{R}$ *is a constant step-size. If* $\gamma \leq \dfrac{1}{\sqrt{3}\ell}$ *then*

$$\gamma^2\,\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] \leq 2\gamma\,\mathbb{E}[\langle g_t, x_t - x^*\rangle \mid \mathcal{F}_t] + 2(\lambda\gamma + 3\sigma^2\gamma^2).$$

Proof. Consider the auxiliary variable $\hat{x}_{t+1} = x_t - \gamma g_t$, then we have

$$\|\hat{x}_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 - 2\gamma\langle g_t, x_t - x^*\rangle + \gamma^2\|g_t\|^2.$$

By taking the expectation given the filtration $\mathcal{F}_t$, we have

$$\mathbb{E}[\|\hat{x}_{t+1} - x^*\|^2 \mid \mathcal{F}_t] = \|x_t - x^*\|^2 - 2\gamma\,\mathbb{E}[\langle g_t, x_t - x^*\rangle \mid \mathcal{F}_t] + \gamma^2\,\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t]. \tag{B.9}$$

Notice that

$$\begin{aligned}
\mathbb{E}[\langle g_t, x_t - x^*\rangle \mid \mathcal{F}_t] &= \mathbb{E}[\langle V(x_{t+1/2}) + U_{t+1/2}(x_{t+1/2}), x_t - x^*\rangle \mid \mathcal{F}_t] \\
&= \mathbb{E}[\langle V(x_{t+1/2}), x_t - x^*\rangle \mid \mathcal{F}_t] \\
&= \mathbb{E}[\langle V(x_t - \gamma V_t), x_t - x^*\rangle \mid \mathcal{F}_t]
\end{aligned}$$

Thus, Eq. (B.9) becomes

$$\begin{aligned}
\mathbb{E}[\|\hat{x}_{t+1} - x^*\|^2 \mid \mathcal{F}_t] =& \|x_t - x^*\|^2 - 2\gamma\,\mathbb{E}[\langle V(x_t - \gamma V_t), x_t - \gamma V_t - x^*\rangle \mid \mathcal{F}_t] \\
&- 2\gamma^2\,\mathbb{E}[\langle V(x_t - \gamma V_t), V_t\rangle \mid \mathcal{F}_t] + \gamma^2\,\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t].
\end{aligned}$$

We can now use Assumption 2 and we get

$$\begin{aligned}
\mathbb{E}[\|\hat{x}_{t+1} - x^*\|^2 \mid \mathcal{F}_t] \leq& \|x_t - x^*\|^2 - 2\mu\gamma\,\mathbb{E}[\|x_t - \gamma V_t - x^*\|^2 \mid \mathcal{F}_t] + 2\lambda\gamma \\
&- 2\gamma^2\,\mathbb{E}[\langle V(x_{t+1/2}) + U_{t+1/2}(x_{t+1/2}), V_t\rangle \mid \mathcal{F}_t] \\
&+ \gamma^2\,\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] \\
\leq& \|x_t - x^*\|^2 + 2\lambda\gamma - 2\gamma^2\,\mathbb{E}[\langle g_t, V_t\rangle \mid \mathcal{F}_t] + \gamma^2\,\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t].
\end{aligned}$$

By using the identity $\|a - b\|^2 = \|a\|^2 + \|b\|^2 - 2\langle a, b\rangle$ we get

$$\mathbb{E}[\|\hat{x}_{t+1} - x^*\|^2 \mid \mathcal{F}_t] \leq \|x_t - x^*\|^2 + 2\lambda\gamma + \gamma^2\,\mathbb{E}[\|g_t - V_t\|^2 \mid \mathcal{F}_t] - \gamma^2\,\mathbb{E}[\|V_t\|^2 \mid \mathcal{F}_t]. \tag{B.10}$$

Furthermore, by using Fact 1 and Assumption 3 we have that

$$\begin{aligned}
\|V_t - g_t\|^2 &= \|V(x_t) - V(x_t - \gamma V_t) + U_t(x_t) - U_{t+1/2}(x_{t+1/2})\| \\
&\leq 3\Big(\|V(x_t) - V(x_t - \gamma V_t)\|^2 + \|U_t(x_t)\|^2 + \|U_{t+1/2}(x_{t+1/2})\|^2\Big) \\
&\leq 3\Big(\ell^2\gamma^2\|V_t\|^2 + \|U_t(x_t)\|^2 + \|U_{t+1/2}(x_{t+1/2})\|^2\Big).
\end{aligned}$$

Thus Eq. (B.10) becomes

$$\mathbb{E}[\|\hat{x}_{t+1} - x^*\|^2 \mid \mathcal{F}_t] \leq \|x_t - x^*\|^2 + \gamma^2(3\ell^2\gamma^2 - 1)\,\mathbb{E}[\|V_t\|^2 \mid \mathcal{F}_t] + 2(\lambda\gamma + 3\sigma^2\gamma^2),$$

where we also used Assumption 4 to bound the variance of the noises $U_t, U_{t+1/2}$. Now if $\gamma \leq \dfrac{1}{\sqrt{3}\ell}$ we have that

$$\mathbb{E}[\|\hat{x}_{t+1} - x^*\|^2 \mid \mathcal{F}_t] \leq \|x_t - x^*\|^2 + 2(\lambda\gamma + 3\sigma^2\gamma^2).$$

Finally, notice that

$$\begin{aligned}\mathbb{E}[\|\hat{x}_{t+1} - x^*\|^2 \mid \mathcal{F}_t] &= \|x_t - x^*\|^2 - 2\gamma\,\mathbb{E}[\langle g_t, x_t - x^*\rangle \mid \mathcal{F}_t] + \gamma^2\,\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t]\\ &\leq \|x_t - x^*\|^2 + 2(\lambda\gamma + 3\sigma^2\gamma^2).\end{aligned}$$

Thus,

$$\gamma^2\,\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] \leq 2\gamma\,\mathbb{E}[\langle g_t, x_t - x^*\rangle \mid \mathcal{F}_t] + 2(\lambda\gamma + 3\sigma^2\gamma^2).$$

■

The above proposition shows how the energy descent inequality is weaken due to noise introduced by the noisy oracle. The next proposition aims to analyze the drift, i.e., $\text{drift}_t = \gamma\,\mathbb{E}[\langle g_t, x_t - x^*\rangle \mid \mathcal{F}_t]$.

Proposition B.2. *Consider that (SEG) is run and let $drift_t = \gamma\,\mathbb{E}[\langle g_t, x_t - x^*\rangle \mid \mathcal{F}_t]$, where g_t is defined as in Proposition B.1. If $\gamma < \dfrac{1}{2\mu + \sqrt{3}\ell}$ and Assumptions 2–4 holds then*

$$-drift_t \leq -\frac{\mu\gamma}{2}\|x_t - x^*\|^2 + (\gamma\lambda + 3\gamma^2\sigma^2). \tag{B.11}$$

Proof. Recall that $g_t = V_{t+1/2} = V(x_{t+1/2}) + U_{t+1/2}(x_{t+1/2})$. We have

$$\begin{aligned}-\text{drift}_t &= -\gamma\,\mathbb{E}[\langle g_t, x_t - x^*\rangle \mid \mathcal{F}_t]\\ &= -\gamma\,\mathbb{E}[\langle V(x_{t+1/2}) + U_{t+1/2}(x_{t+1/2}), x_t - x^*\rangle \mid \mathcal{F}_t]\\ &= -\gamma\,\mathbb{E}[\langle V(x_t - \gamma V_t), x_t - x^*\rangle \mid \mathcal{F}_t]\\ &= -\gamma\,\mathbb{E}[\langle V(x_t - \gamma V_t), x_t - \gamma V_t - x^*\rangle \mid \mathcal{F}_t] - \gamma^2\,\mathbb{E}[\langle V(x_t - \gamma V_t), V_t\rangle \mid \mathcal{F}_t]\\ &\leq -\mu\gamma\,\mathbb{E}[\|x_t - \gamma V_t - x^*\|^2 \mid \mathcal{F}_t] + \gamma\lambda - \gamma^2\,\mathbb{E}[\langle V(x_t - \gamma V_t), V_t\rangle \mid \mathcal{F}_t],\end{aligned}$$

where we used the fact that $x_{t+1/2} = x_t - \gamma V_t$ and the property of weakly quasi strongly monotone (Assumption 2). We now use again the identity $\|a - b\|^2 = \|a\|^2 + \|b\|^2 - 2\langle a, b\rangle$, for all $a, b \in \mathbb{R}^d$, Fact 1 and we get

$$\begin{aligned}-\text{drift}_t \leq& -\mu\gamma\,\mathbb{E}[\|x_t - \gamma V_t - x^*\|^2 \mid \mathcal{F}_t] + \gamma\lambda\\ &- \frac{\gamma^2}{2}\Big(\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] + \mathbb{E}[\|V_t\|^2 \mid \mathcal{F}_t] - \mathbb{E}[\|g_t - V_t\|^2 \mid \mathcal{F}_t]\Big)\\ \leq& -\mu\gamma\,\mathbb{E}[\|x_t - \gamma V_t - x^*\|^2 \mid \mathcal{F}_t] + \gamma\lambda\\ &- \frac{\gamma^2}{2}\Big(\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] + \mathbb{E}[\|V_t\|^2 \mid \mathcal{F}_t]\Big)\\ &+ \frac{\gamma^2}{2}\,\mathbb{E}[3\Big(\|V(x_{t+1/2}) - V(x_t)\|^2 + \|U_t(x_t)\|^2 + \|U_{t+1/2}(x_{t+1/2}\|^2\Big) \mid \mathcal{F}_t]\\ \leq& -\mu\gamma\,\mathbb{E}[\|x_t - \gamma V_t - x^*\|^2 \mid \mathcal{F}_t] + \gamma\lambda\\ &- \frac{\gamma^2}{2}\Big(\mathbb{E}[\|g_t\|^2 \mid \mathcal{F}_t] + \mathbb{E}[\|V_t\|^2 \mid \mathcal{F}_t]\Big)\end{aligned}$$

$$+\frac{3\gamma^2}{2}\Big(\ell^2\gamma^2\,\mathbb{E}[\|V_t\|^2\,|\,\mathcal{F}_t]+2\sigma^2\Big).$$

Furthermore, it holds that $\|a-b\|^2 \geq \dfrac{\|a\|^2}{2} - \|b\|^2$ for all $a, b \in \mathbb{R}^2$; thus by using this inequality and rearranging we have

$$\begin{aligned}
-\text{drift}_t \leq &-\frac{\mu\gamma}{2}\|x_t-x^*\|^2+\gamma\lambda+3\gamma^2\sigma^2\\
&-\frac{\gamma^2}{2}\Big(1-2\mu\gamma-3\gamma^2\ell^2\Big)\,\mathbb{E}[\|V_t\|^2\,|\,\mathcal{F}_t]\\
&-\frac{\gamma^2}{2}\,\mathbb{E}[\|g_t\|^2\,|\,\mathcal{F}_t]\\
\leq &-\frac{\mu\gamma}{2}\|x_t-x^*\|^2+\gamma\lambda+3\gamma^2\sigma^2\\
&-\frac{\gamma^2}{2}\Big(1-2\mu\gamma-3\gamma^2\ell^2\Big)\,\mathbb{E}[\|V_t\|^2\,|\,\mathcal{F}_t].
\end{aligned}$$

In order to cancel out the last term of the above inequality, we require that $1-2\mu\gamma-3\gamma^2\ell^2 \geq 0$ or equivalently $\gamma \in (-\frac{\mu+\sqrt{\mu^2+3\ell^2}}{3\ell^2}, \frac{-\mu+\sqrt{\mu^2+3\ell^2}}{3\ell^2})$. Since $\gamma > 0$, we need that

$$\begin{aligned}
0<\gamma &\leq \frac{-\mu+\sqrt{\mu^2+3\ell^2}}{3\ell^2}\\
&= \frac{3\ell^2}{3\ell^2(\mu+\sqrt{\mu^2+3\ell^2})}\\
&= \frac{1}{\mu+\sqrt{\mu^2+3\ell^2}}.
\end{aligned}$$

Thus, if $\gamma \leq \dfrac{1}{2\mu+\sqrt{3}\ell}$ we get

$$-\text{drift}_t \leq -\frac{\mu\gamma}{2}\|x_t-x^*\|^2+(\gamma\lambda+3\gamma^2\sigma^2).$$

■

With this machinery at hand we proceed to prove the following lemma.

Lemma B.3. *Suppose that Assumptions 1–4 hold then the iterations* $(x_t)_{t\geq 0}$ *of* (SEG)*, if the step-size* $\gamma \leq \dfrac{1}{2\mu+\sqrt{3}\ell}$*, satisfy:*

$$\mathbb{E}[\|x_{t+1}-x^*\|^2\,|\,\mathcal{F}_t] \leq (1-c)^t\|x_0-x^*\|^2+c' \tag{B.12}$$

for some constants $c \in (0,1)$ *and* $c' \in (0,+\infty)$ *that depend on the choice of step-size, as well as the parameters of the problem.*

Proof. We start by analyzing the norm of the difference between the iteration x_{t+1} and the solution x^*. For the updates of (SEG) we use γ to denote the step-size and α to denote the scaling parameter and drop the exponent (SEG) for simplicity.

$$\begin{aligned}
\|x_{t+1}-x^*\|^2 &= \|x_t-\alpha\gamma V_{t+1/2}-x^*\|^2\\
&= \|x_t-x^*\|^2-2\alpha\gamma\langle V_{t+1/2}, x_t-x^*\rangle+\alpha^2\gamma^2\|V_{t+1/2}\|^2.
\end{aligned}$$

Now by taking the expectation on both sides given the filtration $\mathcal{F}_t$ we get

$$\mathbb{E}[\|x_{t+1}-x^*\|^2\,|\,\mathcal{F}_t] = \|x_t-x^*\|^2-2\alpha\text{drift}_t+\alpha^2\gamma^2\,\mathbb{E}[\|g_t\|^2\,|\,\mathcal{F}_t].$$

where g, drift were defined in Propositions B.1 and B.2. Now from the same propositions we get that

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x^*\|^2 \mid \mathcal{F}_t] &\leq \|x_t - x^*\|^2 - 2\alpha \mathrm{drift}_t + 2\alpha^2 \mathrm{drift}_t + 2\alpha^2(\lambda\gamma + 3\sigma^2\gamma^2)\\ &\leq \|x_t - x^*\|^2 - 2\alpha(1-\alpha)\mathrm{drift}_t + 2\alpha^2(\lambda\gamma + 3\sigma^2\gamma^2)\\ &\leq \|x_t - x^*\|^2(1 - \alpha(1-\alpha)\gamma\mu) + 2\alpha(3\gamma^2\sigma^2 + \gamma\lambda). \end{aligned} \tag{B.13}$$

Now let $c = \alpha(1-\alpha)\gamma\mu$ and $c' = \frac{2\alpha(3\gamma^2\sigma^2+\gamma\lambda)}{c}$. Since $\gamma \leq \dfrac{1}{2\mu + \sqrt{3}\ell} < \dfrac{1}{2\mu}$, it holds that $c < 1$.Thus, we have

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq (1-c)^t \|x_0 - x^*\|^2 + c' \tag{B.14}$$

and the proof is completed. ■

Theorem B.1 (Restated Theorem 1)**.** *Consider that either (SGDA) or (SEG) is run with a stochastic oracle satisfying Assumptions 1–4 respectively with step-sizes* $\gamma^{\text{SGDA}} < \dfrac{\mu}{L^2}$, $\gamma^{\text{SEG}} < \dfrac{1}{2\mu + \sqrt{3}\ell}$ *and* $\alpha^{\text{SEG}} \in (0,1)$ *and let* $(x_t)_{t\geq 0}$ *be the iterations generated. Then, there exists a pair of constants* $(c_1, c_2)^{\{\text{SGDA,SEG}\}}$ *that depend on the choice of step-sizes, as well as the parameters of the model, with* $c_1^{\{\text{SGDA,SEG}\}} \in (0,1)$ *and* $c_2^{\{\text{SGDA,SEG}\}} \in (0, +\infty)$ *such that*

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq \left(1 - c_1^{\{\text{SGDA,SEG}\}}\right)^t \|x_0 - x^*\|^2 + c_2^{\{\text{SGDA,SEG}\}}, \tag{B.15}$$

for any initial point $x_0 \in \mathbb{R}^d$.

Proof. Proof follows by combining Lemma B.2 and B.3. ■

B.3 One-step quasi-descent inequality

In this subsection, we provide the proof for one-step "quasi-descent" inequality stated in Corollary 1.

Corollary B.1 (Restated Corollary 1)**.** *Under the conditions of Theorem 1, for all* $x^* \in \mathcal{X}^*$ *there exists an extended real-valued function* $\mathcal{E} : \mathbb{R}^d \to [1, \infty]$ *and constants* $c_1^{\{\text{SGDA,SEG}\}} \in (0,1), c_2^{\{\text{SGDA,SEG}\}} \in (0,\infty)$ *such that*

$$\mathbb{E}[\mathcal{E}(x_{t+1}, x^*) \mid \mathcal{F}_t] \leq c_1^{\{\text{SGDA,SEG}\}} \mathcal{E}(x_t, x^*) + c_2^{\{\text{SGDA,SEG}\}}.$$

Specifically, $\mathcal{E}(x_t, x^*) = \|x_t - x^*\|^2 + 1$.

Proof. For (SGDA), by Eq. (B.8) in the proof of Lemma B.2, we have

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 + 1 \mid \mathcal{F}_t] \leq (1 - 2\mu\gamma + 2\gamma^2 L^2)\left(\|x_t - x^*\|^2 + 1\right) + (2\lambda\gamma + 2\mu\gamma + 2\gamma^2 L^2 \mathrm{R}^2 + \gamma^2\sigma^2).$$

Let $c_1 = 1 - 2\mu\gamma + 2\gamma^2 L^2$ and $c_2 = 2\lambda\gamma + 2\mu\gamma + 2\gamma^2 L^2 \mathrm{R}^2 + \gamma^2\sigma^2$. By the step-size condition, we have $c_1 \in (0,1)$ and $c_2 \in (0,\infty)$ and thus complete the proof for (SGDA).

For (SEG), by Eq. (B.13) in the proof of Lemma B.3, we have

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 + 1 \mid \mathcal{F}_t] \leq (1 - \alpha(1-\alpha)\gamma\mu)\left(\|x_t - x^*\|^2 + 1\right) + \alpha(6\gamma^2\sigma^2 + 2\gamma\lambda + (1-\alpha)\gamma\mu).$$

Now let $c_1 = 1 - \alpha(1-\alpha)\gamma\mu$ and $c_2 = \alpha(6\gamma^2\sigma^2 + 2\gamma\lambda + (1-\alpha)\gamma\mu)$. Similarly, by the step-size condition, we have $c_1 \in (0,1)$ and $c_2 \in (0,\infty)$ and thus complete the proof for (SEG). ■

C Omitted Proofs of Section 4

C.1 Minorization Condition and Geometric Drift Property

Lemma C.1 (Restated Lemma 1)**.** *Let Assumptions 1–5 be satisfied for* (SGDA) *and* (SEG)*. Then given the step-sizes specified in Theorem 1 it holds that for both algorithms the minorization condition is satisfied. Namely there exist constant $\delta > 0$, probability measure ν and set C, dependent on the algorithm such that $\nu(C) = 1$ and $\nu(C^c) = 0$ such that*

$$\Pr[x_{t+1} \in A | x_t = x] \geq \delta \mathbb{1}_C(x)\nu(A) \quad \textit{for all} \quad A \in \mathcal{B}(\mathbb{R}^d), x \in \mathbb{R}^d. \tag{C.1}$$

Proof. We again split the proof in two different parts for each one of the two algorithms. For the sequence we fix a point $x^* \in \mathcal{X}^*$ and we consider the energy function defined as $\mathcal{E}(x) = \|x - x^*\|^2 + 1$.

SGDA: We start by observing that the Energy/Lyapunov function $\mathcal{E}(x) := \|x - x^*\|^2 + 1$ is a function unbounded off small sets, i.e., the sublevel sets $C(r) := \{x \in \mathbb{R}^d | \mathcal{E}(x) \leq r\}$ are either empty or small for all $r > 0$. Indeed assume that $C(r) = \{x \in \mathbb{R}^d | \mathcal{E}(x) \leq r\}$ is non-empty ($r > 1$), then the sublevel sets correspond to some ball $\mathbb{B}(x^*, \sqrt{r-1})$ for $r > 1$. We will prove that the ball $\mathbb{B}(x^*, \sqrt{r-1})$ for $r > 1$ is actually ν_1-small for $m = 1$ (see Definition A.6).

$$\begin{aligned}
P(x, B) &= \mathbb{P}(x_{t+1} \in B | x_t = x) \\
&= \mathbb{P}(x_t - \gamma(V(x_t) + U_t(x_t)) \in B | x_t = x) \\
&= \mathbb{P}(x - \gamma(V(x) + U_t(x)) \in B) \\
&= \mathbb{P}\left(U_t(x) \in (\frac{x}{\gamma} - V(x)) + (-\frac{1}{\gamma}B)\right),
\end{aligned}$$

where $U_t(x) \sim \operatorname{law}(U(x))$ for all $t \geq 0$. With this notation we want to emphasize that once x_t is fixed the distribution of the noise is independent of the time-step, since we have assumed that at each time-step the noises are independent and identically distributed random fields. Thus, we have

$$P(x, B) = \int_{\beta \in B} \operatorname{pdf}_{U(x)}(\frac{x - \beta}{\gamma} - V(x))\, d\beta \tag{C.2}$$

$$\geq \int_{\beta \in B} \inf_{x \in C(r)} \operatorname{pdf}_{U(x)}(\frac{x - \beta}{\gamma} - V(x))\, d\beta \tag{C.3}$$

$$:= \nu_r^{\text{SGDA}}(B). \tag{C.4}$$

Notice that ν_r^{SGDA} is a non-trivial measure since if we set $B = C(r)$, which is a non-empty and bounded set, we have

$$\nu_r^{\text{SGDA}}(C(r)) = \int_{x' \in C(r)} \inf_{x \in C(r)} \operatorname{pdf}_{U(x)}\left(\frac{x - x'}{\gamma} - V(x)\right) dx' > 0,$$

which follows from Assumption 5.

We now fix $r = r_0 > 1$ and proceed in proving the minorization property. Consider the measure $\tilde{\nu}_{r_0}^{\text{SGDA}}(X) = \mathbb{1}(X \subseteq C(r_0))\dfrac{\nu_{r_0}^{\text{SGDA}}(X)}{\nu_{r_0}^{\text{SGDA}}(C(r_0))}$ for all $X \in \mathcal{B}(\mathbb{R}^d)$. It is easy to verify that $\tilde{\nu}_{r_0}^{\text{SGDA}}(C(r_0)) = 1$ and $\tilde{\nu}_{r_0}^{\text{SGDA}}(C(r_0)^c) = 0$. Additionally, if $\{x \notin C(r_0)$ or $A \nsubseteq C(r_0)\}$ we have that $P(x, A) \geq \delta \mathbb{1}_{C(r_0)}(x)\tilde{\nu}_{r_0}^{SGDA}(A) = 0$. Also, if $\{x \in C(r_0)$ and $A \subseteq C(r_0)\}$ we have $P(x, A) \geq \nu_{r_0}^{\text{SGDA}}(A) = \delta \mathbb{1}_{C(r_0)}(x)\tilde{\nu}_{r_0}^{SGDA}(A)$, where $\delta = \nu_{r_0}^{\text{SGDA}}(C(r_0)) > 0$ and thus the proof is completed.

SEG: We continue with the proof when (SEG) is run. Similarly as before we have that

$$\begin{aligned}
P(x, B) &= \mathbb{P}(x_{t+1} \in B | x_t = x) \\
&= \mathbb{P}(x_t - \alpha\gamma V(x_t - \gamma V(x_t) - \gamma U_t(x_t))
\end{aligned}$$

$$
\begin{aligned}
&- \alpha\gamma U_{t+1/2}(x_t - \gamma V(x_t) - \gamma U_t(x_t)) \in B | x_t = x) \\
=& \mathbb{P}(x - \alpha\gamma V(x - \gamma V(x) - \gamma U_t(x)) \\
&- \alpha\gamma U_{t+1/2}(x - \gamma V(x) - \gamma U_t(x)) \in B)
\end{aligned}
$$

where $U_t(x) \sim \operatorname{law}(U^A(x))$, $U_{t+1/2}(x) \sim \operatorname{law}(U^B(x))$ and $U^A(x) \perp U^B(x)$, identically distributed. Thus,

$$
\begin{aligned}
P(x, B) &= \int_{\xi \in \mathbb{R}^d} \mathrm{pdf}_{U^A(x)}(\xi)\, \mathbb{P}\Big(x - \alpha\gamma V(x - \gamma V(x) - \gamma\xi) - \alpha\gamma U^B(x - \gamma V(x) - \gamma\xi) \in B\Big)\, d\xi \\
&= \int_{\beta \in B} \int_{\xi \in \mathbb{R}^d} \mathrm{pdf}_{U^A(x)}(\xi) \mathrm{pdf}_{U^B(x - \gamma V(x) - \gamma\xi)}\left(\frac{x - \beta}{\alpha\gamma} - V(x - \gamma V(x) - \gamma\xi)\right) d\xi\, d\beta \\
&\geq \int_{\beta \in B} \int_{\xi \in \mathbb{B}(0,1)} \mathrm{pdf}_{U^A(x)}(\xi) \mathrm{pdf}_{U^B(x - \gamma V(x) - \gamma\xi)}\left(\frac{x - \beta}{\alpha\gamma} - V(x - \gamma V(x) - \gamma\xi)\right) d\xi\, d\beta.
\end{aligned}
$$

Notice that since $x \in C(r)$, we have that $x - \gamma V(x) - \gamma\xi \in C(r) - \gamma V(C(r)) - \gamma\mathbb{B}(0,1)$. Thus $\mathrm{pdf}_{U^A(x)}(t) \geq \inf_{x \in C(r)} \mathrm{pdf}_{U^A(x)}(t) > 0$ for all $t \in \mathbb{R}^d$ and $\mathrm{pdf}_{U^B(x - \gamma V(x) - \gamma\xi)}(t) \geq \inf_{\rho \in C(r) - \gamma V(C(r)) - \gamma\mathbb{B}(0,1)} \mathrm{pdf}_{U^B(\rho)}(t)$. Hence, we can define the following measure for any set B:

$$
\nu_{r_0}^{\text{SEG}}(B) := \int_{\beta \in B} \int_{\xi \in \mathbb{B}(0,1)} \inf_{x \in C(r)} \mathrm{pdf}_{U^A(x)}(\xi) \inf_{\rho \in C'} \mathrm{pdf}_{U^B(\rho)}\left(\frac{x - \beta}{\alpha\gamma} - V(\rho)\right) d\xi\, d\beta,
$$

where $C' = C(r) - \gamma V(C(r)) - \gamma\mathbb{B}(0,1)$. Notice that the measure is non-trivial since for some fixed $r = r_0 > 1$ we have that $\nu_{r_0}^{\text{SEG}}(C(r_0)) > 0$ since $C(r_0)$ is non-empty. As in the case of SGDA we define

$$
\tilde{\nu}_{r_0}^{\text{SEG}}(X) = \mathbb{1}(X \subseteq C(r_0)) \frac{\nu_{r_0}^{\text{SEG}}(X)}{\nu_{r_0}^{\text{SEG}}(C(r_0))}.
$$

Thus, we have that

$$
P(x, B) \geq \tilde{\nu}_{r_0}^{\text{SEG}}(B).
$$

By repeating the exact same methodology as before the result follows. ■

Corollary C.1 (Improved version of Corollary 2). *Under the setting of Lemma 1 the functions $f_1 := \mathcal{E}, f_2 := \sqrt{\mathcal{E}}$, $f_1, f_2 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ presented in Corollary 1 satisfies the (V4) Geometric Drift Property of [34] for the Markov Chain generated either by* (SGDA) *or* (SEG)*. Namely it holds that there exist a measurable set C, and constants $\beta > 0$, $b < \infty$ such that*

$$
\Delta f_i(x) \leq -\beta f_i(x) + b\, \mathbb{1}_C(x), x \in \mathbb{R}^d, \tag{C.5}
$$

where $\Delta f_i(x) = \int_{y \in \mathbb{R}^d} P(z, dy) f_i(y) - f_i(x)$ for $i \in \{1, 2\}$.

Proof. Based on Definition A.8 we need to show that there exists a function $f : \mathbb{R}^d \to [1, \infty)$, a measurable set C and constants $\beta > 0, b < \infty$ such that $\Delta f(x) \leq -\beta f(x) + b\, \mathbb{1}_C(x)$ for all $x \in \mathbb{R}^d$. We start with the observation that

$$
\Delta f(x) = \int_{y \in \mathbb{R}^d} P(x,\, dy) f(y) - f(x) = \mathbb{E}[f(x_{t+1}) - f(x_t) \,|\, \mathcal{F}_t : \{x_t = x\}]
$$

where x_t that is generated either through (SGDA) or (SEG). Furthermore, notice that the function defined in Corollary 1, $\mathcal{E} : \mathbb{R}^d \to [1, \infty)$ is extended-real valued and also it holds that

$$
\mathbb{E}[\mathcal{E}(x_{t+1}) \,|\, \mathcal{F}_t : \{x_t = x\}] \leq c_1^{\{\text{SGDA,SEG}\}} \mathcal{E}(x) + c_2^{\{\text{SGDA,SEG}\}}
$$

with $c_1^{\{\text{SGDA,SEG}\}} \in (0, 1)$ and $c_2^{\{\text{SGDA,SEG}\}} \in (0, +\infty)$.

Similarly, for the function $\sqrt{\mathcal{E}}$ we have that

$$\begin{aligned}\mathbb{E}[\sqrt{\mathcal{E}(x_{t+1})}\,|\,\mathcal{F}_t : \{x_t = x\}] &\leq \sqrt{\mathbb{E}[\mathcal{E}(x_{t+1})\,|\,\mathcal{F}_t : \{x_t = x\}]}\\ &\leq \sqrt{c_1^{\{\text{SGDA,SEG}\}}\mathcal{E}(x) + c_2^{\{\text{SGDA,SEG}\}}}\\ &\leq \sqrt{c_1^{\{\text{SGDA,SEG}\}}}\sqrt{\mathcal{E}(x)} + \sqrt{c_2^{\{\text{SGDA,SEG}\}}}.\end{aligned}$$

Now notice that for any function $\mathcal{E}$ which is unbounded off small sets and for all $x \in \mathbb{R}^d$ satisfies

$$\mathbb{E}[\mathcal{E}(x_{t+1})\,|\,\mathcal{F}_t : \{x_t = x\}] \leq c\mathcal{E}(x) + c',$$

or equivalently

$$\mathbb{E}[\mathcal{E}(x_{t+1})\,|\,\mathcal{F}_t : \{x_t = x\}] - \mathcal{E}(x) \leq -(1-c)\mathcal{E}(x) + c',$$

we have that it satisfies the geometric drift property for any set $C = \{x \in \mathbb{R}^d : \mathcal{E}(x) \leq \frac{2c'}{(1-c)}\}$ and constants $\beta = \frac{1-c}{2}$ and $b = c'$. Indeed,

$$c' \leq \mathbb{1}_C(x)c' + \mathbb{1}_{C^c}(x)\frac{1-c}{2}\mathcal{E}(x) \text{ for all } x \in \mathbb{R}^d.$$

Thus,

$$\begin{aligned}\mathbb{E}[\mathcal{E}(x_{t+1})\,|\,\mathcal{F}_t : \{x_t = x\}] - \mathcal{E}(x) &\leq -(1-c)\mathcal{E}(x) + \mathbb{1}_C(x)c' + \mathbb{1}_{C^c}(x)\frac{1-c}{2}\mathcal{E}(x)\\ &\leq -\frac{1-c}{2}\mathcal{E}(x) + \mathbb{1}_C(x)c'.\end{aligned}$$

The last inequality follows from the fact that $\mathbb{1}_{C^c}(x) \leq 1$ and $c \in (0,1)$. ■

C.2 Invariant Measure, Total Variation Convergence and Limit Theorems

Lemma C.2 (Restated Lemma 2)**.** *The corresponding Markov chain sequences $(x_t)_{t\geq 0}$ for (SGDA) and (SEG) have the following properties:*

- *They are ψ−irreducible for some non-zero σ-finite measure ψ on $\mathbb{R}^d$ over Borel σ-algebra of $\mathbb{R}^d$.*
- *They are strongly aperiodic.*
- *They are Harris and positive recurrent with an invariant measure.*

Proof. We prove each one of the properties above separately.

- **(Irreducible):** Consider any non-zero σ-finite measure φ in Borel σ-algebra of $\mathbb{R}^d$. From the proof of Lemma C.1 for (SGDA) we have

$$\mathbb{P}(x_{t+1} \in A | x_t = x) = \int_{a\in A} \text{pdf}_{U(x)}(\frac{x-a}{\gamma} - V(x))\, da.$$

  By Assumption 5 and for any $A \subseteq \mathcal{B}(\mathbb{R}^d)$ with $\psi(A) > 0$ we have that $\{x\} \subseteq \mathbb{B}(x,1)$ and there exists $\varepsilon > 0$ such that $\mathbb{B}(a_0,\varepsilon) \subseteq A$, for some $a_0 \in A$. Thus,

$$\begin{aligned}P(x,A) &\geq \int_{\tilde{a}\in\mathbb{B}(a_0,\varepsilon)} \text{pdf}_{U(x)}(\frac{x-\tilde{a}}{\gamma} - V(x))\, d\tilde{a}\\ &\geq \int_{\tilde{a}\in\mathbb{B}(a_0,\varepsilon)} \inf_{\tilde{x}\in\mathbb{B}(x,1)} \text{pdf}_{U(\tilde{x})}(\frac{\tilde{x}-\tilde{a}}{\gamma} - V(x))\, d\tilde{a}\\ &> 0.\end{aligned}$$

Similarly, for the case of (SEG) and by repeating the same argument for some non-zero σ-finite measure φ in $\mathcal{B}\mathbb{R}^d$ algebra, we have that

$$
\begin{aligned}
P(x, A) &= \int_{a\in A} \int_{\xi\in\mathbb{B}(0,1)} \mathrm{pdf}_{U^A(x)}(\xi)\mathrm{pdf}_{U^B(x-\gamma V(x)-\gamma\xi)}\left(\frac{x-a}{\alpha\gamma} - V(x-\gamma V(x)-\gamma\xi)\right) d\xi\, da \\
&\geq \int_{\tilde{a}\in\mathbb{B}(a_0,\varepsilon)} \int_{\xi\in\mathbb{B}(0,1)} \inf_{\tilde{x}\in\mathbb{B}(x,1)} \mathrm{pdf}_{U^A(\tilde{x})}(\xi) \inf_{\rho\in C} \mathrm{pdf}_{U^B(\rho)}\left(\frac{\tilde{x}-\tilde{a}}{\alpha\gamma} - V(\rho)\right) d\xi\, d\tilde{a} \\
&> 0,
\end{aligned}
$$

where $C = \mathbb{B}(x,1) - \gamma V(\mathbb{B}(x,1)) - \gamma\mathbb{B}(0,1)$. The strict positivity for both cases follows from Assumption 5. Thus, by Definition A.2 the sequences are ψ-irreducible.

- **(Strongly Aperiodic):** This is an immediate consequence of the proof of Lemma C.1, since the sets $C(r)$ are small and have positive measure for the measure we constructed.
- **(Recurrent with invariant measure):** Given that the Markov chain is ψ-irreducible and aperiodic, from Theorem 15.0.1 (Geometric Ergodic Theorem) in [34] we have that the chain is positive recurrent and has an invariant measure. This is true since we have proven the geometric drift property (cf. Corollary C.1) for a small set, which is also a petite set by Proposition A.1.

  The fact that the Markov chain is also Harris is a consequence of Theorem 9.1.8 of [34]. For completeness, we mention here that if a chain is ψ-irreducible and there exists a function f that is unbounded off petite sets such that $\Delta f \leq 0$ then the chain is Harris recurrent. All these requirements are direct implications of the results presented so far, particularly the proof of Corollary C.1 and the current lemma. As such, the Markov chains induced by the stochastic gradient descent models in Equations (SGDA) and (SEG) are demonstrably Harris recurrent.

■

Theorem C.1 (Restated Theorem 2)**.** *Let Assumptions 1–5 be satisfied for* (SGDA) *and* (SEG)*. Then given the step-sizes specified in Theorem 1 it holds that*

1. (SGDA) *and* (SEG) *iterates admit a unique stationary distribution* $\pi_\gamma^{\{\text{SGDA,SEG}\}} \in \mathcal{P}_2(\mathbb{R}^d)$*.*
2. *For a test function* $\phi : \mathbb{R}^d \to \mathbb{R}$ *satisfying that* $|\phi(x)| \leq L_\phi(1 + \|x\|)$ *for all* $x \in \mathbb{R}^d_{\geq 0}$*, for some* $L_\phi > 0$ *and for any initialization* $x_0 \in \mathbb{R}^d$ *there exist* $\rho_{\phi,\gamma}^{\{\text{SGDA,SEG}\}} \in (0,1)$ *and* $\kappa_{\phi,x_0,\gamma}^{\{\text{SGDA,SEG}\}} \in (0,\infty)$ *such that:*
$$\left| \mathbb{E}_{x_t}[\phi(x_t)] - \mathbb{E}_{x\sim\pi_\gamma^{\{\text{SGDA,SEG}\}}}[\phi(x)] \right| \leq \kappa_{\phi,x_0,\gamma}^{\{\text{SGDA,SEG}\}} (\rho_{\phi,\gamma}^{\{\text{SGDA,SEG}\}})^t. \tag{C.6}$$
Hence, (SGDA) *and* (SEG) *converge geometrically under the total variation distance to* $\pi_\gamma^{\{\text{SGDA,SEG}\}}$*.*
3. *Finally, for any test function* ϕ *that is* ℓ_ϕ*-Lipschitz we have that*
$$|\mathbb{E}_{x\sim\pi_\gamma^{\{\text{SGDA,SEG}\}}}[\phi(x)] - \phi(x^*)| \leq \ell_\phi\sqrt{D^{\{\text{SGDA,SEG}\}}}, \tag{C.7}$$
for some constant $D^{\{\text{SGDA,SEG}\}} \propto \max(\lambda,\gamma)/\mu$*.*

Proof. The first part of the theorem follows from the fact that the induced Markov chains are Harris recurrent and aperiodic with invariant measure and have the geometric drift property; thus from Strong Aperiodic Ergodic Theorem (See Theorem 13.0.1 in [34]) the measure is unique and finite. Additionally assume that $x_0 \sim \pi_\gamma^{\{\text{SGDA,SEG}\}}$. Then by the invariance property $(x_t)_{t\geq 0} \sim \pi_\gamma^{\{\text{SGDA,SEG}\}}$. Using Corollary 1 for some arbitrary fixed $x^* \in \mathcal{X}^*$, there exist two corresponding constants $(c_1^{\{\text{SGDA,SEG}\}}, c_2^{\{\text{SGDA,SEG}\}})$ such that $c_1^{\{\text{SGDA,SEG}\}} \in (0,1)$ and $c_2^{\{\text{SGDA,SEG}\}} \in (0,\infty)$ that satisfy

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 + 1 \,|\, \mathcal{F}_t] \leq c_1^{\{\text{SGDA,SEG}\}}(\|x_t - x^*\|^2 + 1) + c_2^{\{\text{SGDA,SEG}\}} \Rightarrow \tag{C.8}$$

$$\mathbb{E}_{x\sim\pi_\gamma^{\{\text{SGDA,SEG}\}}}[\|x-x^*\|^2] \le \frac{c_1^{\{\text{SGDA,SEG}\}}+c_2^{\{\text{SGDA,SEG}\}}-1}{1-c_1^{\{\text{SGDA,SEG}\}}} = \mathcal{O}(\max(\lambda,\gamma)/\mu) < \infty. \tag{C.9}$$

Since $\|x^*\| \le$ R, the above inequality implies that $\pi_\gamma^{\{\text{SGDA,SEG}\}} \in \mathcal{P}_2(\mathbb{R}^d)$.

For the second part, we will use the geometric convergence theorem for Harris positive strongly aperiodic Markov Chains endowed with geometric drift property (See 16.0.1 in [34])

$$\begin{aligned}|\phi(x)| &\le L_\phi(1+\|x\|) \le L_\phi((R+1)+\|x-x^*\|) \le L_\phi(R+1)(1+\|x-x^*\|)\\ &\le \sqrt{2}L_\phi(R+1)\sqrt{\mathcal{E}(x)} \le \max(1,\sqrt{2}L_\phi(R+1))\cdot\mathcal{E}'(x) = c'\mathcal{E}'(x)\end{aligned}$$

where $c' := \max(1,\sqrt{2}L_\phi(R+1))$ and $\mathcal{E}'(x) := \sqrt{\mathcal{E}(x)}$. Notice that Corollary C.1 certifies that $\mathcal{E}'$ also satisfies geometric drift property. Additionally, since $c' \ge 1$, $\mathcal{E}''(x) := c'\mathcal{E}'(x)$ also satisfies the geometric drift property. Hence we can prove that (SEG),(SGDA) are $\mathcal{E}''$-uniformly ergodic (Theorem 16.0.1 Condition (iv) in [34]). Therefore, from the equivalent condition (ii) of the aforementioned theorem, there exist $r_{\ell_\phi,\gamma} \in (0,1)$, $R_{\ell_\phi,\gamma} \in (0,\infty)$ such that

$$|P^k\phi(x_0) - \mathbb{E}_{x\sim\pi_\gamma^{\{\text{SGDA,SEG}\}}}[\phi(x)]| \le R_{\ell_\phi,\gamma} r^k_{\ell_\phi,\gamma}|\mathcal{E}''(x_0)|,$$

thus by setting $\kappa_{\phi,x_0,\gamma}^{\{\text{SGDA,SEG}\}} := R_{\ell_\phi,\gamma}|\mathcal{E}''(x_0)|$ and $\rho_{\phi,\gamma}^{\{\text{SGDA,SEG}\}} := r_{\ell_\phi,\gamma}$ we get the requirement. Finally for the total variation distance it suffices to address only test functions that are bounded by 1. Thus there exist constants $r_\gamma \in (0,1)$, $R_\gamma \in (0,\infty)$ independent of the function such that

$$\sup_{|\phi|\le 1}|P^k\phi(x_0) - \mathbb{E}_{x\sim\pi_\gamma^{\{\text{SGDA,SEG}\}}}[\phi(x)]| \le R_\gamma r_\gamma^k|\mathcal{E}''(x_0)|,$$

which implies the geometric convergence under total variation distance via the dual representation of Radon metric for bounded initial conditions [48].

For the last part, we start by linearity of expectation and Lipschitzness of ϕ:

$$\begin{aligned}|\mathbb{E}_{x\sim\pi_\gamma^{\{\text{SGDA,SEG}\}}}[\phi(x)] - \phi(x^*)| &= |\mathbb{E}_{x\sim\pi_\gamma^{\{\text{SGDA,SEG}\}}}[\phi(x)-\phi(x^*)]|\\ &\le \mathbb{E}_{x\sim\pi_\gamma^{\{\text{SGDA,SEG}\}}}[|\phi(x)-\phi(x^*)|]\\ &\le \mathbb{E}_{x\sim\pi_\gamma^{\{\text{SGDA,SEG}\}}}[\ell_\phi\|x-x^*\|]\\ &\le \ell_\phi\sqrt{\mathbb{E}_{x\sim\pi_\gamma^{\{\text{SGDA,SEG}\}}}[\|x-x^*\|^2]}\\ &\le \ell_\phi\sqrt{D^{\{\text{SGDA,SEG}\}}}\end{aligned}$$

where $D^{\{\text{SGDA,SEG}\}} \propto \max(\lambda,\gamma)/\mu$ by Eq. (C.8).

■

Below we use the following notations. The distribution π refers to $\pi_\gamma^{\{\text{SGDA,SEG}\}}$ for respective algorithms. For any function $\phi' : \mathbb{R}^d \to \mathbb{R}$, we introduce the shorthand

$$S_T(\phi') := \sum_{t=1}^{T}\phi'(x_t);$$

in addition, we use $\pi(\phi')$ to denote the expected value of ϕ' over π, i.e., $\pi(\phi') = \mathbb{E}_{x\sim\pi}[\phi'(x)]$.

Theorem C.2 (Restated Theorems 3 and 4)**.** *Let Assumptions 1–5 hold. Then for choice of step-sizes specified in Theorem 2 and any function $\phi : \mathbb{R}^d \to \mathbb{R}$ satisfying $\pi(|\phi|) < \infty$, we have that*

$$\lim_{T\to\infty} \frac{1}{T} S_T(\phi) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \phi(x_t) = \pi(\phi) \quad a.s., \qquad \text{(Law of Large Numbers for (SGDA),(SEG))}$$

and that

$$T^{-1/2} S_T(\phi - \pi(\phi)) \xrightarrow{d} \mathcal{N}(0, \sigma_\pi^2(\phi)), \qquad \text{(Central Limit Theorem for (SGDA),(SEG))}$$

where $\sigma_\pi^2(\phi) := \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_\pi[S_T^2(\phi - \pi(\phi))]$.

Proof. According to Theorem 17.0.1 in [34], the Law of Large Numbers and the Central Limit Theorem, as described in Theorem C.2, hold for positive Harris chains with invariant measures, given that they exhibit $\mathcal{E}^*$-uniform ergodicity. To complete the proof, it is necessary to demonstrate that a function ϕ with linear growth fulfills the conditions of Theorem 17.0.1. This can be achieved by proving the existence of an energy function $\mathcal{E}^*(\cdot)$ satisfying (i) the (V4) geometric drift property in [34] and (ii) $|\phi(x)|^2 \le \mathcal{E}^*(x)$.

$$\begin{aligned}
|\phi(x)|^2 &\le L_\phi^2(1 + \|x\|)^2 \le L_\phi^2(1 + R + \|x - x^*\|)^2 \le L_\phi^2(1+R)^2(1 + \|x - x^*\|)^2 \\
&\le \sqrt{2} L_\phi^2 (1+R)^2 \sqrt{(1 + \|x - x^*\|^2)} \\
&\le \max(1, \sqrt{2} L_\phi^2 (1+R)^2) \sqrt{(1 + \|x - x^*\|^2)} := \mathcal{E}^*(x)
\end{aligned}$$

By Corollary C.1, we get that $\mathcal{E}^*$ satisfies geometric drift property, thus proving that (SEG) and (SGDA) are $\mathcal{E}^*$-uniformly ergodic. We complete the proof of Theorem C.2. ■

D Omitted Proofs of Section 5

D.1 Min-Max Convex-Concave Games

Theorem D.1 (Restated Theorem 5)**.** *Let Assumptions 1–5 hold then the iterates of* (SGDA), (SEG) *when run with the step-sizes given in Theorem 1 admit a stationary distribution* $\pi_\gamma^{\{SGDA,SEG\}}$ *such that*

$$\mathbb{E}_{x\sim\pi_\gamma^{\{SGDA,SEG\}}}[\mathrm{Gap}_V(x)] \leq c\gamma^{SGDA,SEG}, \tag{D.1}$$

where $\mathrm{Gap}_V(x)$ *is the restricted merit function* $\mathrm{Gap}_V(x) := \sup_{x^*\in\mathcal{X}^*}\langle V(x), x-x^*\rangle$ *and* $c\in\mathbb{R}$ *is a constant and depends on the parameters of the problem.*

Proof. From the analysis of (SGDA) in Lemma B.2 (cf. Eqs. (B.4) and (B.7)) we have that

$$\|x_{t+1}-x^*\|^2 \leq \|x_t-x^*\|^2 - 2\gamma\langle V(x_t), x_t-x^*\rangle - 2\gamma\langle U_t(x_t), x_t-x^*\rangle + \gamma^2\|V(x_t)+U_t(x_t)\|^2,$$
$$\|V(x)\|^2 \leq 2L^2((1+\mathrm{R})^2+\|x-x^*\|^2).$$

Since $\mathbb{E}_{x_{t+1}\sim\pi_\gamma}[\|x_{t+1}-x^*\|^2] = \mathbb{E}_{x_t\sim\pi_\gamma}[\|x_t-x^*\|^2]$ we have that

$$\begin{aligned}\frac{1}{\gamma}\mathbb{E}_{x_t\sim\pi_\gamma}[\langle V(x_t), x_t-x^*\rangle] &\leq 2\,\mathbb{E}_{x_t\sim\pi_\gamma}[L^2((1+\mathrm{R})^2+\|x_t-x^*\|^2)] + 2\,\mathbb{E}_{x_t\sim\pi_\gamma}[\|U_t(x_t)\|^2]) \\ &\leq 2L^2((1+\mathrm{R})^2 + 2\,\mathbb{E}_{x_t\sim\pi_\gamma}[\|x_t-x^*\|^2]) + 2\sigma^2 \\ &\leq 2L^2((1+\mathrm{R})^2+2c_2^{\text{SGDA}}) + 2\sigma^2 \\ &\leq \max_{\gamma\in(0,\frac{\mu}{\ell^2})} 2L^2((1+\mathrm{R})^2+2c_2^{\text{SGDA}}) + 2\sigma^2 \\ &\leq C\end{aligned}$$

where $C = \max_{\gamma\in(0,\frac{\mu}{\ell^2})}[2L^2((1+\mathrm{R})^2+2c_2^{\text{SGDA}})+2\sigma^2]$ (Recall that c_2^{SGDA} depends on the step-size).

For the case of (SEG), it easy to see that $\mathrm{Gap}_V(x) \leq \ell\|x_t-x^*\|^2$. So the rest of the proof is derived by Theorem 1, using dominant convergence theorem for $\mathbb{E}_{x_{t+1}\sim\pi_\gamma}[\|x_{t+1}-x^*\|^2]$, as well as the invariance property that $x_\infty\sim\pi_\gamma$ if we initialize $x_0\sim\pi_\gamma$. ■

We next show the connection of Duality-Gap_f and Gap_V for a convex-concave function f and $V = (\nabla_\theta f, -\nabla_\phi f)$:

$$\begin{aligned}\text{Duality-Gap}_f(\theta,\phi) &= \max_{\phi'\in\mathbb{R}^{d_2}} f(\theta,\phi') - \min_{\theta'\in\mathbb{R}^{d_1}} f(\theta',\phi) \\ &= (f(\theta,\phi) - \min_{\theta'\in\mathbb{R}^{d_1}} f(\theta',\phi)) - (f(\theta,\phi) - \max_{\phi'\in\mathbb{R}^{d_2}} V(\theta,\phi')) \\ &\leq \langle V(\theta,\phi), (\theta,\phi)-(\theta^*,\phi^*)\rangle,\end{aligned}$$

where the last step holds since f is convex (resp. concave) in its first (resp. second) argument. Thus if we call $x = (\theta,\phi)$, $x^* = (\theta^*,\phi^*)$, we have

$$\text{Duality-Gap}_f(\theta,\phi) \leq \mathrm{Gap}_V(x).$$

Additionally, it is easy to see that

$$V(\theta,\phi) \leq \max_{\phi'\in\mathbb{R}^{d_2}} V(\theta,\phi') = \text{Duality-Gap}(\theta,\phi) + \min_{\theta'\in\mathbb{R}^{d_1}} V(\theta',\phi) \leq \text{Duality-Gap}(\theta,\phi) + \max_{\phi'\in\mathbb{R}^{d_2}}\min_{\theta'\in\mathbb{R}^{d_1}} V(\theta',\phi')$$

and

$$V(\theta, \phi) \geq \min_{\theta' \in \mathbb{R}^{d_1}} V(\theta', \phi) = \max_{\phi' \in \mathbb{R}^{d_2}} V(\theta, \phi') - \text{Duality-Gap}(\theta, \phi) \geq -\text{Duality-Gap}(\theta, \phi) + \min_{\theta' \in \mathbb{R}^{d_1}} \max_{\phi' \in \mathbb{R}^{d_2}} V(\theta', \phi').$$

By applying the expectation with respect to the invariant distribution and Von-Neuman's minimax theorem we get the desired result in Eq. (14).

D.2 Bias Refinement in Quasi-Monotone Operators

Lemma D.1. *In the setting of Theorem 6 the moments Mom(k) $= \mathbb{E}[\|x_t - x^*\|^k]$ are bounded by a function of $f_k(\gamma)$ where γ is the step-size of* (SGDA) *for $k \in \{1, 2, 3, 4\}$.*

Proof.
Second moment. We start by analyzing the second moment

$$\begin{aligned}\|x_{t+1} - x^*\|^2 =& \|x_t - \gamma V(x_t) - \gamma U_t(x_t) - x^*\|^2\\ \leq& \|x_t - x^*\| - 2\gamma\langle V(x_t), x_t - x^*\rangle - 2\gamma\langle U_t(x_t), x_t - x^*\rangle\\ &+ 2\gamma^2\ell^2\|x_t - x^*\| + 2\gamma^2\|U_t(x_t)\|^2.\end{aligned}$$

We now apply the expectation and quasi strong monotonicity of the operator and get

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 \mid \mathcal{F}_t] \leq \|x_t - x^*\|^2(1 + 2\gamma^2\ell^2 - 2\gamma\mu) + 2\gamma^2\sigma^2.$$

By choosing $1 + 2\gamma^2\ell^2 - 2\gamma\mu < 1 - \gamma\mu$ equivalently $\gamma < \dfrac{\mu}{2\ell^2}$ we have

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x^*\|^2] &\leq \|x_0 - x^*\|^2(1 - \gamma\mu)^{t+1} + 2\gamma^2\sigma^2\sum_{k=0}^{t}(1 - \gamma\mu)^k\\ &\leq \|x_0 - x^*\|^2(1 - \gamma\mu)^{t+1} + \frac{2\gamma^2\sigma^2}{\gamma\mu}\\ &\leq \|x_0 - x^*\|^2(1 - \gamma\mu)^{t+1} + \frac{2\gamma\sigma^2}{\mu}.\end{aligned}$$

Thus if $x \sim \pi_\gamma$, where π_γ is the invariant distribution of the iterates of (SGDA) we have that

$$\int_{\mathbb{R}^d} \|x - x^*\|^2\, d(\pi(x)) \leq 2\frac{\sigma^2\gamma}{\mu}$$

since $\lim_{t\to\infty} x_t \sim \pi_\gamma$.
Fourth moment. For the fourth moment, similarly as before we have that

$$\begin{aligned}\|x_{t+1} - x^*\|^4 =& (\|x_{t+1} - x^*\|^2)^2\\ =& (\|x_t - x^*\|^2 - 2\gamma\langle V(x_t) + U_t(x_t), x_t - x^*\rangle + \gamma^2\|V(x_t) + U_t(x_t)\|^2)^2\\ =& \|x_t - x^*\|^4 + 4\gamma^2(\langle V(x_t) + U_t(x_t), x_t - x^*\rangle)^2 + \gamma^4\|V(x_t) + U_t(x_t)\|^4\\ &- 4\gamma\|x_t - x^*\|^2\langle V(x_t) + U_t(x_t), x_t - x^*\rangle\\ &- 4\gamma^3\|V(x_t) + U_t(x_t)\|^2\langle V(x_t) + U_t(x_t), x_t - x^*\rangle\\ &+ 2\gamma^2\|V(x_t) + U_t(x_t)\|^2\|x_t - x^*\|^2\\ \leq& \|x_t - x^*\|^4 + 4\gamma^2\|x_t - x^*\|^2(2\ell^2\|x_t - x^*\|^2 + 2\|U_t(x_t)\|^2) &\text{(D.2)}\\ &+ \gamma^4(8\ell^4\|x_t - x^*\|^4 + 8\|U_t(x_t)\|^4) &\text{(D.3)}\\ &- 4\gamma\mu\|x_t - x^*\|^4 - 4\gamma\|x_t - x^*\|^2\langle U_t(x_t), x_t - x^*\rangle &\text{(D.4)}\\ &+ 4\gamma^3(4\ell^3\|x_t - x^*\|^3 + 4\|U_t(x_t)\|^3)\|x_t - x^*\| &\text{(D.5)}\\ &+ 4\gamma^2(\ell^2\|x_t - x^*\|^4 + \|U_t(x_t)\|^2\|x_t - x^*\|^2), &\text{(D.6)}\end{aligned}$$

where we used in the second summand Eq. (D.2) the Cauchy-Schwarz inequality, Lipschitz continuity of the operator and the identinty $\|x + y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$. For the third one Eq. (D.3) we used the identity $\|x + y\|^4 \leq 8\|x\|^4 + 8\|y\|^4$, Lipschitzness of the operator. For the fourth one Eq. (D.4) we used the quasi strong monotonicity of the operator. For the firth one Eq. (D.5) we used Cauchy-Schwarz inequality and the identity $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ and Lipschitzness of the operator. Thus in the right-hand side of the above inequality we have constant terms, the $\|x_t - x^*\|^4$, $\|x_t - x^*\|^2$ and $\|x_t - x^*\|$. Specifically, by rearranging we get

$$\begin{aligned}\|x_{t+1} - x^*\|^4 \leq &\|x_t - x^*\|^4(1 + 8\gamma^2\ell^2 + 8\gamma^4\ell^4 - 4\gamma\mu + 16\gamma^3\ell^3 + 4\gamma^2\ell^2) \\ &+ \|x_t - x^*\|^2(12\gamma^2\|U_t(x_t)\|^2) \\ &+ \|x_t - x^*\|(16\gamma^3\|U_t(x_t)\|^3 - 4\|x_t - x^*\|^2\langle U_t(x_t), x_t - x^*\rangle \\ &+ 8\gamma^4\|U_t(x_t)\|^4.\end{aligned}$$

Applying the expectation given the filtration $\mathcal{F}_t$ and setting $\bar{\ell} = \max\{\ell^2, \ell^3, \ell^4\}$ we have

$$\begin{aligned}\mathbb{E}[\|x_t - x^*\|^4 \mid \mathcal{F}_t] \leq &\mathbb{E}[\|x_{t+1} - x^*\|^4 \mid \mathcal{F}_t](1 + 16\bar{\ell}(\gamma^2 + \gamma^3 + \gamma^4) - 4\gamma\mu) \\ &+ \mathbb{E}[\|x_t - x^*\|^2 \mid \mathcal{F}_t](12\gamma^2\sigma^2) \\ &+ \mathbb{E}[\|x_t - x^*\| \mid \mathcal{F}_t](16\gamma^3\delta_{\text{KYRT}}{}^3) + 8\gamma^4\delta_{\text{KYRT}}{}^4.\end{aligned}$$

By choosing step-size such that

$$\begin{cases} \gamma < 1 & \text{for simplicity} \\ 16\bar{\ell}(\gamma^2 + \gamma^3 + \gamma^4) - 4\gamma\mu < -2\gamma\mu & \end{cases}$$

we have that

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x^*\|^4 \mid \mathcal{F}_t](2\gamma\mu) \leq &\mathbb{E}[\|x_t - x^*\|^2 \mid \mathcal{F}_t](12\gamma^2\sigma^2) \\ &+ \mathbb{E}[\|x_t - x^*\| \mid \mathcal{F}_t](16\gamma^3\delta_{\text{KYRT}}{}^3) + 8\gamma^4\delta_{\text{KYRT}}{}^4.\end{aligned}$$

Now consider $x \sim \pi_\gamma$ and let $\mathbb{E}_{x\sim\pi_\gamma}[\|x - x^*\|^k] = \mathrm{Mom}(k)$. Notice that the first moment is also bounded by $\mathcal{O}(\sqrt{\gamma/\mu})$ since from Eq. (B.3) and Lipschitzness of the operator we have

$$\|x_{t+1} - x^*\|^2 \leq (1 - 2\mu\gamma + \gamma^2\ell^2)\|x_t - x^*\|^2 + \|U_t(x_t)\|^2$$

Thus, combining all these we have

$$\mathrm{Mom}(4)2\mu\gamma \leq \mathrm{Mom}(2)\,\mathcal{O}(\gamma^2) + \mathrm{Mom}(1)\,\mathcal{O}(\gamma^3) + \mathcal{O}(\gamma^4).$$

equivalently

$$\mathrm{Mom}(4) \leq \mathrm{Mom}(2)\,\mathcal{O}(\gamma/\mu) + \mathrm{Mom}(1)\,\mathcal{O}(\gamma^2/\mu) + \mathcal{O}(\gamma^3/\mu).$$

But $\mathrm{Mom}(2) \leq \mathcal{O}(\gamma/\mu)$ and $\mathrm{Mom}(1) \leq \mathcal{O}(\sqrt{\gamma/\mu})$, thus

$$\mathrm{Mom}(4) \leq \mathcal{O}(\gamma^2/\mu^2),$$

which implies that there exists $c \leq c_0 \max\{\delta_{\text{KYRT}}{}^3, \delta_{\text{KYRT}}{}^4, \sigma, \sigma^2\}$ such that

$$\mathrm{Mom}(4) \leq c\gamma^2/\mu^2.$$

■

Theorem D.2. *[Restated Theorem 6] Suppose Assumptions 1–5 and 7 hold. There exists a threshold θ such that if $\gamma \in (0, \theta)$,* (SGDA) *admits unique stationary distribution π, that depends on the choice of step-size, and*

$$\mathbb{E}_{x\sim\pi}[x] - x^* = \gamma\Delta(x^*) + \mathcal{O}(\gamma^2), \tag{D.7}$$

where $\Delta(x^*)$ is a vector independent of the choice of step-size γ.

Proof. Let $\bar{x} = \int_{\mathbb{R}^d} x\pi_\gamma(x)\,dx = \mathbb{E}_{x\sim\pi_\gamma}[x]$ and let $\gamma < \min(\gamma_{\text{thresh}}^{\text{D.1}}, \gamma_{\text{thresh}}^{\text{C.1}}) := \theta'$ such that Lemma D.1 and Theorem C.1 hold. Assume that we run (SGDA) $(x_t)_{t\geq 0}$ and $x_0 \sim \pi_\gamma$; since the algorithm is initialized with the invariant distribution, then all the iterations inevitably follow the invariant distribution. We start by applying Taylor expansion, on the operator, of second and third order around the solution x^*

$$V(x) = \nabla V(x^*) \odot [x - x^*] + \frac{1}{2}\nabla^2 V(x^*) \odot [x - x^*]^2 + \mathrm{Res}_3(x), \tag{A}$$

$$V(x) = \nabla V(x^*) \odot [x - x^*] + \mathrm{Res}_2(x), \tag{B}$$

where $\mathrm{Res}_2(x), \mathrm{Res}_3(x)$ are the corresponding residuals of the Taylor expansion for which it holds that $\sup_{x\in\mathbb{R}^d}\{\|\mathrm{Res}_3(x)\|/\|x-x^*\|^3\} < \infty$ and $\sup_{x\in\mathbb{R}^d}\{\|\mathrm{Res}_2(x)\|/\|x-x^*\|^2\} < \infty$. Notice also that

$$\int_{x\in\mathbb{R}^d} \mathrm{Res}_3(x)\pi_\gamma(x)\,dx < c_3\int_{x\in\mathbb{R}^d}\|x-x^*\|^3\pi_\gamma(x)\,dx \leq c_3\mathrm{Mom}(3) \leq \mathcal{O}(\gamma^{3/2}), \tag{C}$$

$$\int_{x\in\mathbb{R}^d} \mathrm{Res}_2(x)\pi_\gamma(x)\,dx \leq c_2\int_{x\in\mathbb{R}^d}\|x-x^*\|^2\pi_\gamma(x)\,dx \leq c_2\mathrm{Mom}(2) \leq \mathcal{O}(\gamma). \tag{D}$$

Additionally, by definition of (SGDA) we get that $x_1 = x_0 - \gamma V(x_0) - \gamma U_0(x_0)$. Since $x_0 \sim \pi_\gamma$ we have that $x_1 \sim \pi_\gamma$ and thus we have

$$\mathbb{E}_{x_1\sim\pi_\gamma}[x_1] = \mathbb{E}_{x_0\sim\pi_\gamma}[x_0] - \gamma\,\mathbb{E}_{x_0\sim\pi_\gamma}[V(x_0)] - \gamma\,\mathbb{E}_{x_0\sim\pi_\gamma}[U_0(x_0)],$$

which implies that

$$\mathbb{E}_{x\sim\pi_\gamma}[V(x)] = 0. \tag{E}$$

With these equations at hand, we proceed and take the expectation of (A) with respect to the invariant distribution, combining also (C) and (E) and we get

$$\nabla V(x^*) \odot [\bar{x} - x^*] + \frac{1}{2}\int_{x\in\mathbb{R}^d} \nabla^2 V(x^*) \odot [x - x^*]^2\pi_\gamma(x)\,dx = \mathcal{O}(\gamma^{3/2}). \tag{D.8}$$

Again we focus on the first update of (SGDA) and we have

$$\begin{aligned}
x_1 &= x_0 - \gamma V(x_0) - \gamma U_0(x_0)\\
x_1 - x^* &= x_0 - x^* - \gamma\left(\nabla V(x^*) \odot [x_0 - x^*] + \mathrm{Res}_2(x_0)\right) - \gamma U_0(x_0)\\
x_1 - x^* &= (I - \gamma(V(x^*)) \odot [x_0 - x^*] - \gamma\mathrm{Res}_2(x_0) - \gamma U_0(x_0).
\end{aligned}$$

We now compute $[x_1 - x^*]^2 = (x_1 - x^*)(x_1 - x^*)^\top$ and apply the expectation with respect to the invariant distribution and the noise and we have

$$\begin{aligned}
\mathbb{E}_{x\sim\pi_\gamma}[[x - x^*]^2] = (I - \gamma\nabla V(x^*)) \odot \mathbb{E}_{x\sim\pi_\gamma}[(x - x^*)^2] \odot (I - \gamma\nabla V(x^*)) + \gamma^2\,\mathbb{E}_{x_0\sim\pi_\gamma}[[U_0(x_0)]^2]\\
+\mathcal{O}\left(\underbrace{\gamma\int_{x\in\mathbb{R}^d}\mathrm{Res}_3(x) \odot (I - \gamma(V(x^*)) \odot [x_0 - x^*]\pi_\gamma(x)\,dx + \gamma^2 + \cdots}_{\gamma^{5/2}}\right).
\end{aligned}$$

This leads to

$$\mathbb{E}_{x\sim\pi_\gamma}[[x - x^*]^2] = \gamma Q(x^*)\,\mathbb{E}_{x_0\sim\pi_\gamma}[[U_0(x_0)]^2] + \mathcal{O}(\gamma^{3/2}),$$

where $Q(x^*) := (\nabla V(x^*) \odot I + I \odot \nabla V(x^*) - \gamma\nabla V(x^*) \odot \nabla V(x^*))^{-1}$, which is invertible since

$$\nabla V(x^*) \odot I + I \odot \nabla V(x^*) - \gamma\nabla V(x^*) \odot \nabla V(x^*) = \nabla V(x^*) \odot M(x^*) + M(x^*) \odot \nabla V(x^*),$$

where $M(x^*) := I - \gamma/2\,\nabla V(x^*)$. By quasi-monotonicity around x^* and by choosing $\gamma < \min(2L, \theta') := \theta$ we get that the tensor $Q(\gamma^*)$ is positive definite tensor.

By applying a second-order Taylor expansion about x^* in $Op(x) := [U_t(x)]^2$, and utilizing the same reasoning as above in combination with the differentiability of the noise tensor (see Assumption 7), we derive the following:

$$\mathbb{E}_{x\sim\pi_\gamma}[[U_t(x)]^2] = [U_t(x^*)]^2 + \mathcal{O}(\gamma) \tag{D.9}$$

$$\mathbb{E}_{x\sim\pi_\gamma}[[U_t(x)]^2 \odot [x - x^*]] = [U_t(x^*)]^2 \odot [\mathbb{E}_{x\sim\pi}[x] - x^*] + \mathcal{O}(\gamma). \tag{D.10}$$

Combining (D.8),(D.2),(D.9), we get that

$$\bar{x} - x^* = -\frac{1}{2}[\nabla V(x^*)]^{-1} \odot \nabla^2 V(x^*) \odot \left(\gamma Q(x^*)\, \mathbb{E}_{x_0\sim\pi_\gamma}[[U_0(x_0)]^2] + \mathcal{O}(\gamma^{3/2})\right) + \mathcal{O}(\gamma^{3/2}),$$

which implies that

$$\bar{x} - x^* = -\frac{1}{2}[\nabla V(x^*)]^{-1} \odot \nabla^2 V(x^*) \odot \left(\gamma Q(x^*) \odot \{[U_t(x^*)]^2 + \mathcal{O}(\gamma)\} + \mathcal{O}(\gamma^{3/2})\right) + \mathcal{O}(\gamma^{3/2}),$$

or equivalently

$$\bar{x} - x^* = \gamma\Delta(x^*) + \mathcal{O}(\gamma^{3/2}).$$

The rest of the proof has the goal to improve the last term the order to $\mathcal{O}(\gamma^2)$.

1. We have seen that via (D.2),(D.9),: $\mathbb{E}_{x\sim\pi_\gamma}[[x - x^*]^2] = \gamma Q(x^*) \odot [U_t(x^*)] + \gamma^2 Q(x^*) + o(\gamma^2)$

2. With similar calculations we can prove that: $\mathbb{E}_{x\sim\pi_\gamma}[[x - x^*]^3] = \gamma^2 B(x^*) + o(\gamma^2)$

Using 4-th order taylor again we get the following equality

$$\begin{aligned} x_1 - x^* &= x_0 - x^* \\ &\quad - \gamma\big(\nabla V(x^*) \odot [x - x^*] + \frac{1}{2!}\nabla^2 V(x^*) \odot [x - x^*]^2 \\ &\qquad + \frac{1}{3!}\nabla^3 V(x^*) \odot [x - x^*]^2 + \mathrm{Res}_4(x)\big) \\ &\quad - \gamma U_0(x_0) \end{aligned}$$

Applying expectation in the above equality and combining the bounds (1.) and (2.), we have that

$$\left\{\begin{array}{c} \nabla V(x^*) \odot [\bar{x} - x^*] + \frac{1}{2}\nabla^2 V(x^*) \odot \mathbb{E}_{x\sim\pi_\gamma}[[x - x^*]^2] \\ + \\ \frac{1}{3!}\nabla^3 V(x^*) \odot \mathbb{E}_{x\sim\pi_\gamma}[[x - x^*]^3] + \mathbb{E}_{x\sim\pi_\gamma}[\mathrm{Res}_4(x)] \end{array}\right\} = 0 \tag{D.11}$$

By applying the fourth-moment bound for $\mathbb{E}_{x\sim\pi_\gamma}[\mathrm{Res}_4(x)] = \mathcal{O}(\gamma^2)$ we get the promised result. ■