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 (VIPs), Stochastic Extragradient (SEG) and Stochastic Gradient Descent Ascent (SGDA) have emerged as preeminent algorithms. Constant step-size versions of SEG/SGDA have gained popularity due to several appealing benefits, but their convergence behaviors are complicated even in rudimentary bilinear models. Our work elucidates the probabilistic behavior of these algorithms and their projected variants, for a wide range of monotone and non-monotone VIPs with potentially biased stochastic oracles. By recasting them as time-homogeneous Markov Chains, we establish geometric convergence to a unique invariant distribution and aymptotic normality. Specializing to min-max optimization, we characterize the relationship between the step-size and the induced bias with respect to the global solution, which in turns allows for bias refinement via the Richardson-Romberg scheme. Our theoretical analysis is corroborated by numerical experiments.

1 INTRODUCTION

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

In the setting with only noisy access to the underlying operator, various stochastic algorithms for VIP have been studied. Two prime examples are Stochastic Extragradient (SEG) (Juditsky et al., 2011) and Stochastic Gradient Descent Ascent (SGDA) methods (Nemirovski et al., 2009). Much progress has been made in recent years on understanding the convergence of SEG and SGDA, as well as stochastic gradient descent (SGD), a special case of SGDA. Classical results on these stochastic methods typically assume that a diminishing step-size is used, which allows for last-iterate almost sure convergence to the global solution (Mishchenko et al., 2020; Kannan and Shanbhag, 2019; Mertikopoulos and Zhou, 2019; Hsieh et al., 2020a; Gorbunov et al., 2022; Loizou et al., 2021; Yang et al., 2020; Beznosikov et al., 2023; Gorbunov et al., 2020).

In this paper, we focus on the constant step-size variants of SEG and SGDA. The use of constant step-sizes, which is popular in practice and performs well empirically, offers several major benefits: insensitivity to initial conditions, fast progress in early iterations, easy tuning with a single parameter, and low correlation between iterates facilitating statistical inference.

However, in theoretical study on stochastic VIP methods, their convergence properties have been widely acknowledged to be more delicate than their deterministic and loss minimization counterparts. In addition, the use

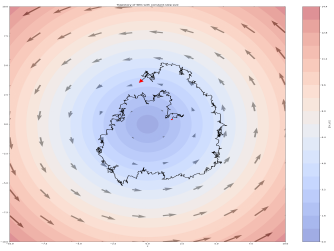


Figure 1: Non-convergence of constant step-size SEG in a quasi-bilinear game: $\min_x \max_y \epsilon x^2 + xy - \epsilon y^2$, with $\epsilon \approx 10^{-4}$.

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

of constant step-size SEG and SGDA present several challenges not present in the diminishing step-size case. Various non-convergent behaviors exist even in rudimentary bilinear models (Gidel et al., 2018; Mertikopoulos et al., 2019; Chavdarova et al., 2019; Daskalakis et al., 2018); see Figure 1 for an example. In particular, with a constant step-size, the iterates of SEG and SGDA do not converge to an exact VIP solution but rather fluctuate around the solution due to persistent stochastic noise. Existing results are typically in the form of an *upper bound* on the mean squared error or dual gap. These upper bounds typically conflate the deterministic (convergence) and stochastic (fluctuation) aspects of SEG/SGDA, and often fail to explain the benefits of constant stepsize.

In this work, we elucidate the fine-grained properties of SEG/SGDA with constant step-sizes. Rather than treating the stochastic fluctuation as a nuisance, we fully embrace the probabilistic nature of SEG/SGDA, by viewing them as time-homogeneous continuous state space Markov chains. We show that while the iterate does not converge, its distribution does. This perspective allows us to separately characterize the distributional convergence behavior and the properties of the limit distribution, as summarize below.

Our Contributions. For a class of constrained VIPs with *weak quasi strongly monotonicity*, which encompasses various non-monotone and non-convex problems, we establish the following results.

- We show that the iterates of SEG and SGDA form a Harris and positive recurrent Markov chain, which admits a unique invariant and limit distribution. Moreover, the distribution of the iterate, as well as any Lipschitz functional thereof, converge to the limit at a geometric rate.
- We derive an ergodic Law of Large Number (LLN) and a Central Limit Theorem (CLT) for the averaged iterate, establishing its asymptotic normality.
- We show that the induced bias—distance between the mean of the invariant distribution and the global VIP solution—is bounded by a linear function of the step-size and weak monotonicity parameter. Specializing to convex-concave min-max optimization, we quantify the bias w.r.t. 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. With this characterization, we apply the Richardson-Romberg refinement scheme to achieves an order-wise reduction of the bias.

In the above results, we quantify the dependence on the stepsize and the parameters of the VIP and stochastic oracle, highlighting the superior performance of SEG for smooth problems and the resilience of SGDA in nonsmooth settings. Moreover, our results apply to *projected* SEG/SGDA for constrained VIPs, and to potentially biased stochastic oracles.

Challenges and Techniques. By connecting SEG/SGDA to Markov chains, we leverage the powerful framework laid out in Meyn and Tweedie (2009); Douc et al. (2018) for convergence and ergodicity of stochastic processes. To this end, we establish several key properties of the associated Markov chain, including irreducibility, positive and Harris recurrence and the Foster-Lyapunov condition. These properties stipulate that the iterates will return to a "small set" infinitely thanks to a negative geometric drift of an appropriate potential function. These properties in turn ensure the existence and convergence to a unique invariant distribution, and the validity of limit theorems such as LLN and CLT.

While the above Markov chain framework provides a high level strategy, its implementation in stochastic VIPs is met with several challenges. VIPs are defined on subsets of $\mathbb{R}^d$, corresponding to a continuous, uncountable and multidimensional state space for the Markov chain, which requires more advanced machinery compared to finite state Markov chains. Moreover, unlike minimization problems, general VIPs lack a gradient field structure and a natural potential function. Therefore, analytic techniques for the former need not generalize to VIPs. This challenge is intensified in the analysis of SEG, which involves two interdependent random steps, necessitating more nuanced arguments compared to SGD(A). The analysis is further complicated by the use of additional projection steps in constrained VIPs, by the consideration of *quasi* monotonicity, and by the absence of cocercivity and unbiasedness of the noisy oracle in our setting.

Consequently, our analysis is considerably more delicate than the recent work in Dieuleveut et al. (2018); Yu et al. (2021), which also adopt the Markov chain perspective for (unconstrained) SGD. We compare with them in more details after presenting our assumptions and main results in Section 4.

Despite the discussed challenges, we manage to provide a unified, streamlined analysis of SEG, SGDA and their projected variants, in smooth and nonsmooth VIPs.

Other Related Work. There is an extensive body of work on the algorithms for VIPs, for both the deterministic setting (Gidel et al., 2018; Mokhtari et al., 2020; Diakonikolas et al., 2021) and the stochstic setting with SEG (Mertikopoulos and Zhou, 2019; Gorbunov et al., 2022) and SGDA (Lin et al., 2020; Beznosikov et al., 2023). More recent work considers extensions of constant stepsize SGD and other stochastic approximation algorithms (Bianchi et al., 2022; Durmus et al.,

2021; Huo et al., 2023). For concision, we refer to the Appendx for extended discussion of related work.

2 PROBLEM SETUP

We start by delineating the variational inequalities framework and the stochastic oracle setting.

2.1 Variational Inequalities

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a closed convex set and $V : \mathbb{R}^d \to \mathbb{R}^d$ be a single-valued operator. The corresponding VIP is

$$\text{Find } x^* \in \mathcal{X} : \langle V(x^*), x - x^* \rangle \geq 0 \ \ \forall x \in \mathcal{X}. \quad \text{(VI)}$$

The examples below showcase the applications of VIPs.

Example 2.1 (Solving nonlinear equations)**.** Solutions of (VI) with $\mathcal{X} = \mathbb{R}^d$ correspond to roots of the equation $V(x) = \mathbf{0}$. Examples include Navier-Stokes equations in computational dynamics (Hao, 2021).

Example 2.2 (Loss minimization)**.** For a C^1-smooth function $f : \mathcal{X} \to \mathbb{R}$, a solution x^* of (VI) with $V = \nabla f$ is a (KKT) critical point, where (i) $\langle \nabla f(x^*), x - x^* \rangle \geq 0, \forall x \in \mathcal{X}$ (constrained case) and (ii) $\nabla f(x^*) = 0$ if $\mathcal{X} = \mathbb{R}^d$ (unconstrained case). For convex f, x^* is a global minimizer. Loss minimization powers model training in machine learning (Lan, 2020).

Example 2.3 (Saddle-point problems)**.** For a (quasi) convex-concave function $\mathcal{L} : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$, a solution (x_1^*, x_2^*) of (VI) with $V = (\nabla_{x_1} \mathcal{L}, -\nabla_{x_2} \mathcal{L})$ and $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ is a saddle point of $\mathcal{L}$, that is,

$$\mathcal{L}(x_1^*, x_2) \leq \mathcal{L}(x_1^*, x_2^*) \leq \mathcal{L}(x_1, x_2^*), \ \forall (x_1, x_2) \in \mathcal{X}. \quad \text{(SP)}$$

Example 2.4 (Nash Equilibria)**.** Consider N players, each with a convex action set $\mathcal{X}_i \subset \mathbb{R}^{n_i}$ and a cost function $c_i : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} = \prod_i \mathcal{X}_i$. A joint action profile $x^* = (x_i^*)_{i=1}^N \in \mathcal{X}$ is Nash equilibrium (NE) if

$$c_i(x^*) \leq c_i(x_i; x_{-i}^*), \ \forall i, x_i \in \mathcal{X}_i. \quad \text{(NE)}$$

If c_i's are separately convex, then $V = (\nabla_{x_i} c_i(x))_{i=1}^N$ is monotone, and the solutions of (VI) and (NE) coincide.

The problems (SP) and (NE) are prominent in training GANs, actor-critic techniques, multi-agent reinforcement learning, and auction/bandit problems (Pfau and Vinyals, 2016; Zhang et al., 2021; Gidel et al., 2018; Daskalakis et al., 2018).

2.2 Assumptions and Stochastic Oracle

Our blanket assumptions in this study are the following:

Assumption 1. The solution set $\mathcal{X}^*$ of (VI) is nonempty and there exist $x^* \in \mathcal{X}^*, \mathrm{R} \in \mathbb{R}$ with $\|x^*\| \leq \mathrm{R}$.

Assumption 2. The operator V is λ-weak μ-quasi strongly monotone with parameters $\lambda \geq 0$, $\mu > 0$. That is, for all $x \in \mathbb{R}^d$:

$$\langle V(x), x - x^* \rangle \geq \mu \|x - x^*\|^2 - \lambda \text{ for some } x^* \in \mathcal{X}^*. \quad (1)$$

By letting $x = x^{*\prime} \in \mathcal{X}^*$ in (1), one can show that $\|x^* - x^{*\prime}\|^2 \leq \lambda/\mu$ for all $x^*, x^{*\prime} \in \mathcal{X}^*$, by using the fact that $\langle V(x^{*\prime}), x^* - x^{*\prime} \rangle \leq 0$. Hence the solution set $\mathcal{X}^*$ is contained in a ball of radius $\sqrt{\lambda/\mu}$, which vanishes if $\lambda = 0$. In the rest of the paper, x^* denotes an arbitrary fixed element of $\mathcal{X}^*$.

As an example of a function for which Assumption 2 is satisfied with $\lambda > 0$, one may consider the function $f(x, y) = (x^2 + 10\sin(x)) + xy - (y^2 - 10\cos(y))$. In this case, the assumption is satisfied for $(\mu, \lambda) = (1, 25)$.

Assumption 3. For different algorithms, we adopt the following reguarity conditions for V:

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

$$\|V(x') - V(x)\| \leq \ell \|x' - x\| \text{ for all } x, x' \in \mathbb{R}^d. \quad (2)$$

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

$$\|V(x)\| \leq L(1 + \|x\|) \text{ for all } x \in \mathbb{R}^d. \quad (3)$$

Our algorithms access V through a black-box stochastic oracle. When queried at $x_t \in \mathcal{X}$, the oracle returns

$$V_t = V(x_t) + U_t(x_t), \quad (4)$$

where $U_t(x_t)$ is additive noise. We impose the following assumption, which allows the noise to have a non-zero mean and a second moment with linear growth.

Assumption 4. $(U_t(\cdot))_{t \geq 0}$ is a sequence of i.i.d. random fields with the following properties:

- *Bounded Bias*: $\| \mathbb{E}[U_t(x) \,|\, \mathcal{F}_t] \| \leq b, \forall x, t$;
- *Second Moment*: $\mathbb{E}[\|U_t(x)\|^2 \,|\, \mathcal{F}_t] \leq \sigma^2 + \rho d^2(x, \mathcal{X}^*)$, where $d(x, S) = \inf_{y \in S} \|x - y\|$ and for some $\sigma, \rho > 0$.

Herein, $\mathcal{F}_t$ is the history (σ-algebra) generated by $x_1, \ldots, x_t$. Note that x_t is adapted to $\mathcal{F}_t$, but $U_t(x_t)$ is generated after x_t and thus not adapted to $\mathcal{F}_t$.

A few remarks are in order. When $\lambda = 0$, Assumption 2 has been considered in the VIPs literature under the names of quasi-strong monotonity (Loizou et al., 2021), strong stability condition (Mertikopoulos and Zhou, 2019) and strongly coherent VIPs (Song et al., 2020). This assumption is weaker than strong monotonicity/convexity, i.e., $\langle V(x), x - x' \rangle \geq \mu \|x - x'\|^2, \forall x, x'$. With $\lambda > 0$, Assumption 2 represents a further relaxation inspired by weakly convex optimization and dissipative dynamical systems (Erdogdu et al., 2018; Raginsky et al., 2017). This assumption emcompasses various

non-monotone games and problems frequently encountered in statistical learning (Tan and Vershynin, 2019), such as functions of the form $a_{\lambda,\mu}\|x\|^2 + b_{\lambda,\mu}\sin(\|x\|)$ and rescaled versions of the Rastrigin function.

Assumption 3 represents a well-established dichotomy on VIPs: we leverage (SEG) for its superior rates in smooth problems, whereas (SGDA) is employed in non-smooth settings. Note that unlike Dieuleveut et al. (2018), Assumption 3 is imposed on the true/expected operator V, not the stochastic oracle (4).

Assumption 4 is standard in the analysis of stochastic algorithms in VIPs and optimization (Nemirovski et al., 2009; Mertikopoulos and Zhou, 2019; Yang et al., 2020; Hsieh et al., 2019, 2020a). Such noisy access can emerge either explicitly due to privacy-induced noise (Song et al., 2013), or implicitly due to limited observability in games (Giannou et al., 2021a,b) or from model uncertainties in tasks like distributional robust optimization (DRO) (Rahimian and Mehrotra, 2019). When $\rho = 0$, Assumption 4 corresponds to the traditional assumption of bounded noise variance.

3 Algorithms

In this paper we focus on two of the most widely used algorithms for variational inequalities: SGDA and 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,

$$x_{t+1} = \Pi_{\mathcal{X}}(x_t - \gamma^{\text{SGDA}} V_t) \tag{SGDA}$$

where $\Pi_{\mathcal{X}}$ denotes the projection operator onto $\mathcal{X}$.

Stochastic Extra Gradient. Inspired by the *extragradient* (EG) algorithm proposed by Korpelevich (1976), *extra-point* schemes have been widely adopted for smooth VIPs. These algorithm incorporate an extra "look-ahead" step, denoted by $x_{t+1/2}$ below, to approximate the future value $V(x_{t+1})$ and enhance convergence. Here we examine its stochastic variant with a single constant step-size $\gamma^{\text{SEG}} \in (0, \infty)$, as follows:

$$\begin{aligned} x_{t+1/2} &= \Pi_{\mathcal{X}}(x_t - \gamma^{\text{SEG}} V_t), \\ x_{t+1} &= \Pi_{\mathcal{X}}(x_t - \gamma^{\text{SEG}} V_{t+1/2}), \end{aligned} \tag{SEG}$$

where $V_{t+1/2}$ is the output of the stochastic oracle queried at $x_{t+1/2}$, and $V_{t+1/2}$ is independent of V_t (this is termed the I-SEG model in Hsieh et al. 2020a).

We study the trajectories $(x_t)_{t\geq 0}$ of (SGDA) and (SEG) through the lens of Markov Chain theory. Observe that:

(i) The iterates $(x_t)_{t\geq 0}$ constitute respectively a Markov chain, with the post-update state x_{t+1} depending solely on the current state x_t.

(ii) The chain and its transition kernel is time-homogeneous, thanks to the use of constant step-size and 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.

By exploiting the specific structures of (SEG) and (SGDA), we study three fundamental properties of the induced Markov chain: *irreducibility*, *aperiodicity*, and *recurrence* (Meyn and Tweedie, 2009), which allows us to establish convergence and limit theorems that characterize the fine-grained behavior of SEG/SGDA.

3.1 Preliminary Convergence Results

We derive an initial convergence result, which takes the form of "geometric convergence up to a constant factor" and an associated descent inequality. This preliminary result serves as the first step for proving our main results on distributional convergence of the Markov chain, enabling a unified analysis for both methods.

Classical work of Nesterov et al. (2018) and Tseng (1995) shows that when the operator V is Lipschitz and strongly monotone, the noiseless versions of (SGDA) and (SEG) converge exponentially to the solution set $\mathcal{X}^*$. In our stochastic setting with the relaxed weakly quasi strong monotonicity Assumption 2, we derive similar convergence results up to an additive constant that depends on the stepsize γ, the noise variance σ^2 and the shift λ of weakly quasi-monotonicity.

Theorem 1. *Under Assumptions 1–4, consider (SGDA) and (SEG) with step-sizes* $\gamma^{\text{SGDA}} = \mathcal{O}(\frac{\mu}{L^2+\rho})$, $\gamma^{\text{SEG}} = \mathcal{O}(\frac{1}{\ell} \wedge \frac{\mu}{\rho})$ *respectively, and let* $(x_t)_{t\geq 0}$ *be the generated iterations. There exists a pair of constants* $c_1^{Alg} \in (0,1)$ *and* $c_2^{Alg} \in (0, +\infty)$ *that depend on the choice of step-sizes and model's parameters such that*

$$\mathbb{E}[\|x_{t+1} - x^*\|^2] \leq \left(1 - c_1^{Alg}\right)^t \|x_0 - x^*\|^2 + c_2^{Alg}, \tag{5}$$

for any initial point $x_0 \in \mathcal{X}$ *and* $Alg \in \{\text{SGDA,SEG}\}$.

The contraction and bias constants $c_1^{\text{Alg}}, c_2^{\text{Alg}}$ above will also play a role in our main results in Section 4 to follow. Our proofs give explicit formulas for these constants. In particular, assuming the stepsize choices in Theorem 1 and ignoring universal constants, we have $c_1^{\text{SEG}} \asymp c_1^{\text{SGDA}} \asymp \gamma\mu$, $c_2^{\text{SEG}} \asymp \gamma\frac{\sigma^2}{\mu} + \frac{\lambda\mu+b^2}{\mu^2}$, and $c_2^{\text{SGDA}} \asymp \gamma\frac{\sigma^2+L^2(1+\mathrm{R}^2)}{\mu} + \frac{\lambda\mu+b^2}{\mu^2}$. A salient feature is that the biases c_2^{SEG} and c_2^{SGDA} are proportional to the stepsize γ when $\lambda = b = 0$ (quasi-strong monotonicity and unbiased oracle). Moreover, compared to (SGDA), (SEG) can use a larger stepsize in the smooth setting (hence better c_1^{SEG} by a factor of the condition number ℓ/μ) and achieve a smaller bias c_2^{SEG}. We also note that these expressions are consistent with existing results

in the noiseless setting. See Appendix for additional discussion on these points.

Notably, besides I-SEG assumption of independent noise per step, we opt for a single step-size scheme, differing from intricate double step-size approaches of Loizou et al. (2020); Hsieh et al. (2020b). Interestingly, under Assumption 2, even such a simplified scheme suffices to vanish the bias c_2^{SEG} when γ approaches 0.

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

Corollary 1. *Under the conditions of Theorem 1, there exist constants* $\widehat{c_1^{Alg}} \in (0,1)$ *and* $\widehat{c_2^{Alg}} \in (0,\infty)$ *with* $Alg \in \{\text{SGDA,SEG}\}$ *such that the function* $\mathcal{E}(x_t, x^*) := \|x_t - x^*\|^2 + 1$ *satisfies*

$$\mathbb{E}[\mathcal{E}(x_{t+1}, x^*) \,|\, \mathcal{F}_t] \leq \widehat{c_1^{Alg}}\mathcal{E}(x_t, x^*) + \widehat{c_2^{Alg}}. \tag{6}$$

The function $\mathcal{E}$ is often called an energy, potential or Lyapunov function. The above results apply to an arbitrary fixed $x^* \in \mathcal{X}^*$. In the sequel, we omit the reference to x^* and simply write the function as $\mathcal{E}(x_t)$.

Our subsequent Markov chain analysis centers around three types of recurrence properties: *(null)-recurrence*, *Harris recurrence*, and *positive recurrence*, which stipulate, respectively, that the chain revisits regions of the state space, doing so infinitely often with probability 1, and with finite expected return time; we defer their formal definitions to the supplement. For our continuous, uncountable and potentially unbounded state spaces, these properties are substantially more nuanced and challenging to establish than for finite state spaces.

To streamline the analysis, we make use of a commonly accepted regularity assumption on the noise:

Assumption 5. For all $t \geq 0$ and each $x \in \mathcal{X}$, the distribution of the random variable $U_t(x)$ can be decomposed as $\nu_1 + \nu_2$, where ν_1 is a probability measure that has a probability density function (pdf), denoted as $\text{pdf}_{U_t(x)}$, which satisfies $\inf_{x\in C} \text{pdf}_{U_t(x)}(t) > 0$ for all bounded sets $C \subseteq \mathcal{X}$.

Note that the measure induced by the random variable $U_t(x)$ need not to have a density function, since ν_2 is an arbitrary measure and we only require ν_1 to have a density function.

A similar assumption is used in Yu et al. (2021) for SGD. Assumption 5 is relatively weak, satisfied by Gaussian and other continuous random fields supported on $\mathbb{R}^d$. In fact, one can always satisfy this assumption by adding (arbitrarily) small continuous noise to the stochastic oracle (4)—itself a common practice for inducing privacy—without affecting subsequent quantitative bounds. As shall become clear in the analysis, Assumption 5 ensures *ψ-irreducibility* of the continuous space Markov chains of SEG and SGDA. In return, we do not require the noisy oracle $V_t(\cdot) = V(\cdot) + U_t(\cdot)$ to be *almost surely co-cooercive*, namely, $\ell\langle V_t(x) - V_t(x'), x - x'\rangle \geq \|V_t(x) - V_t(x')\|^2, \forall x, x'$, which is a strong assumption needed in the prior work Dieuleveut et al. (2018).

4 MAIN RESULTS

We summarize the main result of this section as follows:

Informal Theorem. *Under Assumptions 1–5, the iterates of* (SGDA) *and* (SEG) *are strongly aperiodic, positive and Harris recurrent continuous-state Markov Chains. Each chain converges a unique stationary distribution regardless of initialization, and the averaged iterates satisfy a Law of Large Numbers and an ergodic Central Limit Theorem.*

4.1 Minorization, Drift and Recurrence

In this subsection, we establish *(i)* the *Minorization Condition* and *(ii)* the *Geometric Drift Property* for our methods. These properties serve an important role in proving *Harris and positive* recurrence, respectively.

Lemma 1. *Let the Assumptions 1–5 be satisfied for* (SGDA) *and* (SEG). *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*

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

If the set C encompassed the entire space $\mathcal{X}$, the minorization condition (MC) would indicate that every subset of $\mathcal{X}$ is reachable from any state. This together with standard coupling arguments would imply geometric convergence of the distribution of x_t towards a unique distribution. In fact, we do not need $C = \mathcal{X}$, a restricted condition when $\mathcal{X}$ is unbounded. Rather, a subset C that satisfies (MC), known as a "small/petite" set, can still ensure geometric convergence thanks to the following Foster-Lyapunov drift property.

Corollary 2. *Under the setting of Lemma 1, the function* $\mathcal{E} : \mathcal{X} \to \mathbb{R}$ *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*

$$\Delta\mathcal{E}(x) \leq -\beta\mathcal{E}(x) + b\mathbb{1}_C(x), x \in \mathcal{X}, \tag{FL}$$

[†]Eq. (FL) is popularized by Meyn and Tweedie (2009) as the (V4) geometric drift property.

where $\Delta\mathcal{E}(x) = \int_{y\in\mathcal{X}} P(x,dy)\mathcal{E}(y) - \mathcal{E}(x)$.

A negative r.h.s. of the Foster-Lyapunov inequality (FL) ensures that the energy function $\mathcal{E}$ decreases exponentially as the Markov chain transitions from states outside the set C. Consequently, the chain quickly forgets its initial state and returns to C, exhibiting stationary behavior around regions with small values of the energy function $\mathcal{E}(\cdot)$.

We highlight that the projection operator $\Pi_{\mathcal{X}}$ and its interplay with the double steps in (SEG) present a major hurdle in proving Lemma 1 and Corollary 2.

Equipped with the minorization and geometric drift properties, we are ready to establish the necessary conditions for the ergodicity of (SGDA) and (SEG):

Lemma 2. *The Markov chains* $(x_t)_{t\geq 0}$ *corresponding to* (SGDA) *and* (SEG) *satisfy following:*

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

4.2 Convergence and Limit Theorems

Our first main result is about the invariant measure:

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

1. *The iterates* $(x_t)_{t\geq 0}$ *admit a unique invariant distribution* $\pi_\gamma^{Alg} \in \mathcal{P}_2(\mathcal{X})$*, where* $\mathcal{P}_2(\mathcal{X})$ *is the set of distributions on* $\mathcal{X}$ *with bounded second moments.*
2. *For any test function* $\phi : \mathcal{X} \to \mathbb{R}$ *of* L_ϕ*-linear growth and any initialization* $x_0 \in \mathcal{X}$*, there exist constants* $\tau_{\phi,\gamma}^{Alg} \in (0,1)$ *and* $\zeta_{\phi,x_0,\gamma}^{Alg} \in (0,\infty)$ *such that:*
$$\left|\mathbb{E}_{x_t}[\phi(x_t)] - \mathbb{E}_{x\sim\pi_\gamma^{Alg}}[\phi(x)]\right| \leq \zeta_{\phi,x_0,\gamma}^{Alg}(\tau_{\phi,\gamma}^{Alg})^t. \quad (7)$$
Hence, (SGDA) *and* (SEG) *converges geometrically under the total variation distance to* π_γ^{Alg}*.*
3. *For each* ℓ_ϕ*-Lipschitz test function* ϕ*, it holds that*
$$|\mathbb{E}_{x\sim\pi_\gamma^{Alg}}[\phi(x)] - \phi(x^*)| \leq \ell_\phi\sqrt{D^{Alg}}, \quad (8)$$
for some constant $D^{Alg} \propto c_2^{Alg}$*.*

Theorem 2 is established using generalized ergodic theorems for Markov chains satisfying (MC) and (FL). The theorem asserts geometric convergence of constant step-size (SGDA) and (SEG) to unique stationary distributions and provides bounds for mean of the limit distribution relative to the VIP solution x^*. These results hold even for non-smooth and non-convex VIPs.

Following the influential work of Polyak and Juditcky (Polyak, 1990), we next study Asymptotic Normality of the two algorithms. To the best of our knowledge, such a result is the first of its kind for stochastic methods within the variational inequalities framework, especially for extrapolation techniques like (SEG). To streamline our discussion, let us introduce a notation.

Definition 1. For a given function ϕ, denote the average of ϕ evaluated over iterate of our methods, known as the Césaro mean (Hardy and Series, 1992), by $\overline{S_T}(\phi) := \frac{1}{T}S_T(\phi) := \frac{1}{T}\sum_{t=0}^{T}\phi(x_t)$.

We begin with a Law of Large Numbers (LLN) for (SGDA) and (SEG), established using the analogue of the Birkhoff–Khinchin ergodic theorem for continuous state space Markov Chains.

Theorem 3. *Let the Assumptions 1–5 hold. Then for the choice of step-sizes specified in Theorem 2 and any function* ϕ *satisfying* $\pi_\gamma^{Alg}(|\phi|) < \infty$*, where* $\pi_\gamma^{Alg}(|\phi|) = \mathbb{E}_{x\sim\pi_\gamma^{Alg}}[|\phi(x)|]$*, it holds that for* $Alg \in \{\text{SGDA,SEG}\}$ *:*
$$\lim_{T\to\infty}\overline{S_T}(\phi) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}\phi(x_t) = \pi_\gamma^{Alg}(\phi) \quad a.s.$$
(Law of Large Numbers for (SGDA) & (SEG))

The next result is a central limit theorem (CLT) for (SGDA) and (SEG), establishing the asymptotic normality of their averaged iterates. This result provides theoretical justifications for constructing confidence intervals in VIPs and min-max games, surpassing the sole dependence on empirical evidence in the prior work Antonakopoulos et al. (2021); Hsieh et al. (2020a).

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 for* $Alg \in \{\text{SGDA,SEG}\}$*:*
$$\sqrt{T}\cdot\left(\overline{S_T}(\phi) - \pi_\gamma^{Alg}(\phi)\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2_{\pi_\gamma^{Alg}}(\phi)\right),$$
(Central Limit Theorem for (SGDA) & (SEG))
where $\pi_\gamma^{Alg}(\phi) = \mathbb{E}_{x\sim\pi_\gamma^{Alg}}[\phi(x)]$, $\sigma^2_{\pi_\gamma^{Alg}}(\phi) = \lim_{T\to\infty}\frac{1}{T}\mathbb{E}_{\pi_\gamma^{Alg}}[S_T^2(\phi - \pi_\gamma^{Alg}(\phi))]$*, and* $\mathbb{E}_{\pi_\gamma^{Alg}}$ *denotes that the initial distribution of the Markov chain is* π_γ^{Alg}*.*

Remark 1. We compare our results with the recent work in Dieuleveut et al. (2018); Yu et al. (2021), which view constant step-size SGD as Markov chains. Both of them consider only unconstrained minimization problems. Our work studies constrained VIPs and the projected verison of SGDA, as well as projected SEG, a more complicated, extrapolation-based algorithm. Moreover, we allow for a biased stochastic oracle.

The Markov chain analysis in Dieuleveut et al. (2018) uses coupling and convergence in *Wasserstein* distance. This approach requires (exact-)strong convexity and

smoothness/cocoercivity of the noise. Our work is instead based on irreducibility, positive/Harris recurrence and convergence in total variation distance, leveraging Assumption 5 and quasi-strong monotonicity.

Yu et al. (2021) also uses an irreducibility and recurrence based approach, but focuses exclusively on vanilla SGD in the nonsmooth case. Our results provide a unified treatment of the smooth setting of SEG and the nonsmooth setting of SGDA, and at the same time are strong enough to differentiate performance of SEG and SGDA in the smooth case. Compared to Yu et al. (2021), we also provide a more refined characterization of the bias for SGDA (and hence SGD; see Section 5.2).

5 APPLICATIONS

In this section, we discuss the applications of our main results 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.

5.1 Min-Max Convex-Concave Games

We consider a specific class of operators as follows

Assumption 6. The operator V is monotone:

$$\langle V(x) - V(x'), x - x' \rangle \geq 0 \text{ for all } x, x' \in \mathcal{X}. \quad (9)$$

Note that Assumption 2 and 6 together are weaker than strong monotonicity. Also define the restricted merit function $\mathrm{Gap}_V(x) := \sup_{x^* \in \mathcal{X}^*} \langle V(x), x - x^* \rangle$.

Theorem 5. *Let Assumptions 1–6 hold with $b = 0$. Then the iterates of (SGDA), (SEG), when run with the step-sizes given in Theorem 1, admit a stationary distribution π_γ^{Alg} such that for $Alg \in \{\text{SGDA,SEG}\}$:*

$$\mathbb{E}_{x \sim \pi_\gamma^{Alg}}[\mathrm{Gap}_V(x)] \leq c\gamma^{Alg}, \quad (10)$$

where $c \in \mathbb{R}$ depends on the parameters of the problem.

Consider the subcase of convex-concave min-max games (Example 2.3) with objective $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$, a convex function in the first argument and concave in the second one. With $x = (\theta, \phi)$ and $V = (\nabla_\theta f, -\nabla_\phi f)$, the aforementioned $\mathrm{Gap}_V(x)$ upper bounds the standard notion of *duality gap*:

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

also known as *primal-dual optimality gap* or *Nash gap*.

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 $\pi_\gamma^{\mathrm{Alg}}$ of the iterates of (SGDA) and (SEG), we have

$$|\mathbb{E}_{(\theta,\phi) \sim \pi_\gamma^{\mathrm{Alg}}}[f(\theta, \phi)] - \mathrm{val}^*| \leq c\gamma^{\mathrm{Alg}}. \quad (11)$$

From (10) and (11), we see that (SGDA) and (SEG) converge to val^*—the unique value at a Nash Equilibrium—within an expected error that is proportional to the stepsize γ^{Alg}, where the error is measured by the duality gap or the difference in the game value.

5.2 Bias Refinement in Quasi-Strong Case

Here we focus on the class of quasi-monotone operators (i.e., $\lambda = 0$ in Assumption 2), which encompasses a variety of non-monotone and non-convex problems. We provide a refined analysis of the stationary distribution induced by (SGDA) under the following smoothness and regularity assumptions for the operator and noise.

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 $i = 1, \ldots, 4$). Furthermore, the noise has bounded kyrtosis, meaning that $\mathbb{E}[\|U_t(x)\|^4] < \delta^4_{\mathrm{KYRT}}$ 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 $i \in \{1, 2, 3\}$.

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 (Gautschi, 2011) to construct a new estimate provably closer to the optimal solution. Our result is a strict generalization of Dieuleveut et al. (2018), which requires co-coersive noisy oracles.

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*

$$\mathbb{E}_{x \sim \pi_\gamma}[x] - x^* = \gamma \Delta(x^*) + \mathcal{O}(\gamma^2), \quad (12)$$

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

Notably, Eq. (12) is an *equality* (up to a second order term). In the above setting, this equality gives a more precise characterization of the bias than the *upper bound* (C.6) applied to $\phi(x) = x$.

As an immediate implication of the refined characterization above, one can use the following RR refinement scheme to obtain a better estimate of x^*. Consider running two (SGDA) recursions with step-sizes 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 π_γ and $\pi_{2\gamma}$ be the resulting unique stationary distributions. By our LLN result (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. Eq. (12) implies that

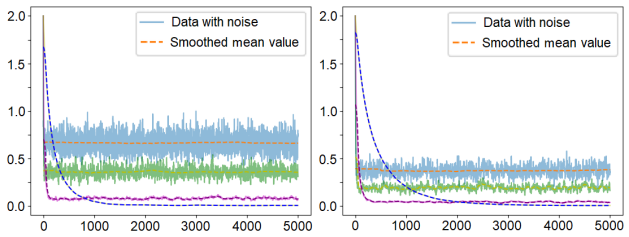
$$\big(\mathbb{E}_{x\sim\pi_\gamma}[2x] - \mathbb{E}_{y\sim\pi_{2\gamma}}[y]\big) - x^* = \mathcal{O}(\gamma^2).$$

Therefore, the RR-refined average 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.

6 EXPERIMENTS

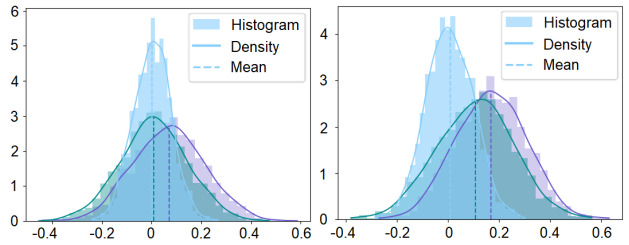
We conduct a series of experiments to empirically validate our results. We focus on a strongly convex-concave game with two players. See the appendix for the details of the game and experiment setup.

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



(a) SGDA. **(b)** SEG.

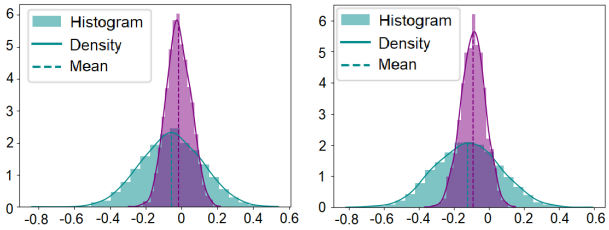
Figure 2: Convergence and bias under different step-sizes for SGDA and SEG.



(a) SGDA **(b)** SEG

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

The second set of experiments examines the central limit theorem (CLT). We use as test function the value of the game, which is zero, 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 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.



(a) SGDA **(b)** SEG

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

Lastly, we investigate the effect of the RR refinement scheme discussed in Section 5.2. 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 the error for the RR refinement scheme. 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.

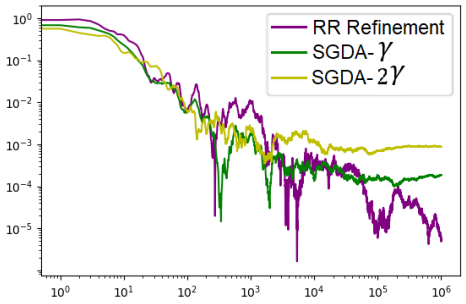


Figure 5: Errors of SGDA and RR refinement.

7 CONCLUDING REMARKS

In this work, we delve into the probabilistic structures inherent in the Stochastic Extragradient and Stochastic Gradient Descent Ascent algorithms, widely used in constrained min-max optimization and variational inequalities problems. By treating constant step-size, projected variants of SEG/SGDA as time-homogeneous Markov Chains, we establish geometric ergodicity, 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.

Our work points to several future directions. Of immediate interests is extension to broader operator families, alternative noise models and other step-size schemes, which may involve time-inhomogeneous Markov chains with sub-geometric convergence. Investigating other optimization algorithms, such as Optimistic Gradient Descent Ascent, which requires higher-order Markov analysis, is another promising line of research. It will also be fruitful to study applications of our results, particularly the use of large step-sizes and iterate averaging, in statistical inference, adversarial training, and robust machine learning.

Acknowledgments

We would like to thank all the reviewers for their helpful feedback. Q. Xie is supported in part by NSF grant CNS-1955997. Y. Chen is supported in part by NSF grants CCF-1704828 and CCF-2233152. This project started when E. V. V. Gkaragkounis, Q. Xie and Y. Chen were attending the Data-Driven Decision Processes program of the Simons Institute for the Theory of Computing. Finally, E. V. V. Gkaragkounis is grateful for his financial support by the FODSI-Simons Fellowship and Pancretan Association of America.

References

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

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

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

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

B. Can, M. Gurbuzbalaban, and N. S. Aybat. A variance-reduced stochastic accelerated primal dual algorithm. *arXiv preprint arXiv:2202.09688*, 2022.

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

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

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

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

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains, 2018.

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

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

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

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

A. Giannou, E.-V. Vlatakis-Gkaragkounis, and P. Mertikopoulos. Survival of the strictest: Stable and unstable equilibria under regularized learning with partial information. In *Annual Conference Computational Learning Theory*, 2021a. URL https://api.semanticscholar.org/CorpusID:231816176.

A. Giannou, E.-V. Vlatakis-Gkaragkounis, and P. Mertikopoulos. On the rate of convergence of regularized learning in games: From bandits and uncertainty to optimism and beyond. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 22655–22666. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/bf40f0ab4e5e63171dd16036913ae828-Paper.pdf.

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

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

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

W. Hao. 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.

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

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

Y.-G. Hsieh, F. Iutzeler, J. Malick, and P. Mertikopoulos. 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*, 2020a.

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

D. L. Huo, Y. Chen, and Q. Xie. Bias and extrapolation in markovian linear stochastic approximation with constant stepsizes. *ACM SIGMETRICS*, 2023. URL https://arxiv.org/abs/2210.00953.

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

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

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

G. Lan. *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.

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

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

N. Loizou, H. Berard, G. Gidel, I. Mitliagkas, and S. Lacoste-Julien. 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.

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

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

S. P. Meyn and R. L. Tweedie. *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.

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

A. Mokhtari, A. Ozdaglar, and S. Pattathil. 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*, pages 1497–1507. PMLR, 2020.

J.-M. Morvan. *Generalized Curvatures*, volume 2 of *Geometry and Computing*. Springer, Berlin, Heidelberg, 2008.

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

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

Y. Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

C. H. Papadimitriou, E. Vlatakis-Gkaragkounis, and M. Zampetakis. 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.

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

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

M. Raginsky, A. Rakhlin, and M. Telgarsky. Nonconvex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pages 1674–1703. PMLR, 2017.

H. Rahimian and S. Mehrotra. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

C. Song, Z. Zhou, Y. Zhou, Y. Jiang, and Y. Ma. Optimistic dual extrapolation for coherent non-monotone

variational inequalities. *Advances in Neural Information Processing Systems*, 33:14303–14314, 2020.

S. Song, K. Chaudhuri, and A. D. Sarwate. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pages 245–248. IEEE, 2013.

Y. S. Tan and R. Vershynin. 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.

P. Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1):237–252, 1995. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(94)00094-H. URL https://www.sciencedirect.com/science/article/pii/037704279400094H. Proceedings of the International Meeting on Linear/Nonlinear Iterative Methods and Verification of Solution.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2009. ISBN 978-3-540-71050-9. doi: 10.1007/978-3-540-71050-9. URL https://link.springer.com/book/10.1007/978-3-540-71050-9.

Wikipedia. Spherical measure, 2023. URL https://en.wikipedia.org/wiki/Spherical_measure.

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

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

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] Sections 2,3 contains our setting and Assumptions. See also section 3 and 5.1 for assumptions related to specific results.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] See our main results.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] Included in the supplemental material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] Sections 2, 3 and 5.1.
   (b) Complete proofs of all theoretical results. [Yes] See supplemental material.
   (c) Clear explanations of any assumptions. [Yes] See sections 2,3,5.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] Supplemental material.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] Supplemental material.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] Supplemental material.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] Supplemental material.
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes] Supplemental material.

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

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 $\mathcal{X}$, $\mathcal{B}(\mathcal{X})$ associated either with (SGDA) or (SEG) to be[1]

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

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

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

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

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

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

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

for any n and all $A_i \in \mathcal{B}(\mathcal{X})$.

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 $\mathcal{B}(\mathcal{X})$ such that for all $x \in \mathcal{X}$ 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 Meyn and Tweedie (2009). Consider a $\psi-$irreducible Markov chain, we use $\mathcal{B}^+(\mathcal{X})$ to denote the set of sets $A \in \mathcal{B}(\mathcal{X})$ 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 $\mathcal{B}^+(\mathcal{X})$ is recurrent then we call Φ recurrent.
- **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 $A \in \mathcal{B}^+(\mathcal{X})$ is positive then Φ is called positive recurrent.
- **Harris recurrent:** The set A is called Harris recurrent if $\mathbb{P}(\eta_A = \infty \,|\, \Phi_0 = x) = 1$ for all $x \in A$. If every set $A \in \mathcal{B}^+(\mathcal{X})$ is Harris recurrent, then Φ is called Harris recurrent.

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

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 $\mathcal{B}(\mathcal{X})$ satisfying $\nu_1(A) > 0$, and for all $x \in A$ and $S \in \mathcal{B}(\mathcal{X})$,

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

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 $\nu(C) = 1$:

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

If C was the entire $\mathcal{X}$, 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 $C = \mathcal{X}$ is usually impossible. The set where such a condition holds is called "small".

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

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

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

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

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 $C \in \mathcal{B}(\mathcal{X})$ ν_a-petite if the sampled chain satisfies the bound

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

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

Proposition A.1 (Proposition 5.5.3 in Meyn and Tweedie (2009))**.** *If a set $C \in \mathcal{B}(\mathcal{X})$ 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 Meyn and Tweedie (2009).

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 $f : \mathcal{X} \to [1, \infty]$, a measurable set C, and constants $\beta > 0$, $b < \infty$ such that

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

where $\Delta f(x) = \int_{y \in \mathcal{X}} P(z, dy) f(y) - f(x)$.

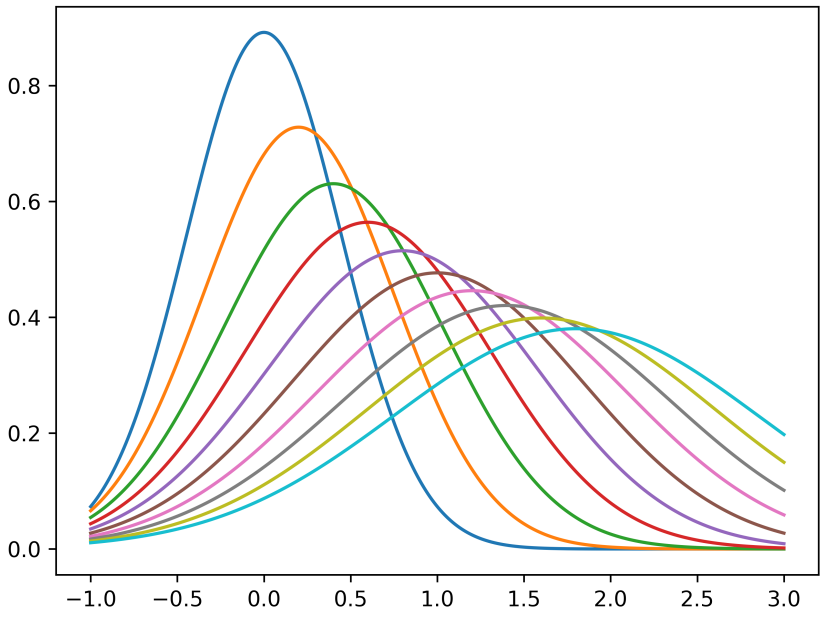


Figure 6: Example of transition kernel $P(x, C)$ for $x \in \mathcal{X}$

B Omitted Proofs of Section 3

We start with a basic lemma for projections:

Lemma B.1. *Let x, y be vectors in $\mathbb{R}^d$ and $\mathcal{X}$ be an arbitrary compact convex set. Then it holds that*

$$y = \Pi_{\mathcal{X}}(x) \Leftrightarrow x \in y + \mathcal{N}_{\mathcal{X}}(y) \equiv y + \{z \in \mathbb{R}^d : \langle z, k - y \rangle \leq 0 \quad \forall k \in \mathcal{X}\}$$

where $\mathcal{N}_{\mathcal{X}}(y)$ corresponds to the normal cone of the convex set $\mathcal{X}$ at the point y.

Proof. Indeed, let's define the projection of a point over an arbitrary set $\mathcal{X}$ as an unconstrained optimization problem of a lower semi-continuous extended convex function. More precisely, it holds that

$$y = \Pi_{\mathcal{X}}(x) = \arg\min_{z \in \mathcal{X}} \{\|x - z\|_2^2/2\} = \arg\min_{z \in \mathbb{R}^d} \{\|x - z\|_2^2/2 + \mathbf{1}_{\mathcal{X}}(z)\}$$

From the generalized Fermat's theorem[2] and the fact that every stationary point for a convex function corresponds to a global minimizer, it holds that:

$$y = \Pi_{\mathcal{X}}(x) \Leftrightarrow 0 \in \partial\{\|x - z\|_2^2/2 + \mathbf{1}_{\mathcal{X}}(z)\}(y) \Leftrightarrow 0 \in y - x + \partial \mathbf{1}_{\mathcal{X}}(y) \Leftrightarrow x \in y + \mathcal{N}_{\mathcal{X}}(y)$$

[2]Generalized Fermat's theorem Statement (See Theorem 8.15 Rockafellar and Wets (2009)): If a function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is nondifferential, convex, proper and it has a local minimum at $\bar{x}$, then $0 \in \partial f(\bar{x})$.

where we used the fact from subdifferential calculus that $\mathcal{N}_{\mathcal{X}}(y) = \partial \mathbf{1}_{\mathcal{X}}(y)$. ■

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

Lemma B.2. *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:

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

By this definition we get that

$$\begin{aligned} P(x, B) &= \mathbb{P}(x_{t+1} \in B | x_t = x) \\ &= \mathbb{P}\big(\Pi_{\mathcal{X}}(x_t - \gamma^{\text{SGDA}} V_t) \in B | x_t = x\big) \\ &= \mathbb{P}\Big(x_t - \gamma^{\text{SGDA}}(V(x_t) + U_t(x_t)) \in \hat{B} | x_t = x\Big) \\ &= \mathbb{P}\Big(x - \gamma^{\text{SGDA}}(V(x) + U_t(x_t)(x)) \in \hat{B}\Big) \\ &= \mathbb{P}\left(U(x) \in (\frac{x}{\gamma^{\text{SGDA}}} - V(x)) + (-\frac{1}{\gamma^{\text{SGDA}}}\hat{B})\right), \end{aligned}$$

where $\hat{B} = \bigcup_{\chi \in B}(\chi + \mathcal{N}_{\mathcal{X}}(\chi))$ and $(U_t(x_t)(x))_{t \geq 0} \sim U(x)$, 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

$$\begin{aligned} x_{t+1} &= \Pi_{\mathcal{X}}[x_t - \gamma^{\text{SEG}} V_{t+1/2}] \\ &= \Pi_{\mathcal{X}}\left[x_t - \gamma^{\text{SEG}}\left(V(x_{t+1/2}) + U_{t+1/2}(x_{t+1/2})\right)\right] \\ &= \Pi_{\mathcal{X}}\left[x_t - \gamma^{\text{SEG}}\begin{bmatrix} V(\Pi_{\mathcal{X}}\left[x_t - \gamma^{\text{SEG}}\left\{V(x_t) + U_t(x_t)\right\}\right]) \\ + \\ U_{t+1/2}(\Pi_{\mathcal{X}}\left[x_t - \gamma^{\text{SEG}}\left\{V(x_t) + U_t(x_t)\right\}\right]) \end{bmatrix}\right]. \end{aligned} \tag{SEG}$$

Thus, if we call $V(x) = v_x$ and $\hat{B} = \bigcup_{\chi \in B}(\chi + \mathcal{N}_{\mathcal{X}}(\chi))$ for the transition kernel we get that

$$\begin{aligned} P(x, B) =& \mathbb{P}(x_{t+1} \in B | x_t = x) \\ =& \mathbb{P}(x - \gamma^{\text{SEG}}\left\{V(\Pi_{\mathcal{X}}\left[x - \gamma^{\text{SEG}}\left\{v_x + U^{(i)}(x)\right\}\right]) + U^{(ii)}(\Pi_{\mathcal{X}}\left[x - \gamma^{\text{SEG}}\left\{v_x + U^{(i)}(x)\right\}\right])\right\} \in \hat{B}) \\ =& \int_{\xi \in \mathbb{R}^d} \mathbf{1}\{\tilde{x} = \Pi_{\mathcal{X}}[x - \gamma^{\text{SEG}}\{v_x + \xi\}]\} \text{pdf}_{U^{(i)}(x)}(\xi)\, \mathbb{P}\left(U^{(ii)}(\tilde{x}) \in (\frac{x}{\gamma^{\text{SEG}}} - V(\tilde{x}) + (-\frac{1}{\gamma^{\text{SEG}}}\hat{B})\right) \text{d}\xi \end{aligned}$$

where $U_t(x_t)(x) \sim \text{law}(U^{(i)}(x))$, $U_{t+1/2}(x) \sim \text{law}(U^{(ii)}(x))$ and $U^{(i)}(x) \perp U^{(ii)}(x)$, identically distributed. So again, $P(x, B)$ is 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

$$\|x + y\|^2 \leq 2(\|x\|^2 + \|y\|^2) \quad \& \quad \|x + y + z\|^2 \leq 3(\|x\|^2 + \|y\|^2 + \|z\|^2). \tag{B.1}$$

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

Lemma B.3. *Suppose that Assumptions 1–4 hold then the iterations $(x_t)_{t\geq 0}$ of (SGDA), if the step-size is $\gamma < \frac{\mu/8}{L^2+\rho}$, satisfy:*

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

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

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|\Pi_{\mathcal{X}}[x_t - \gamma V_t] - \Pi_{\mathcal{X}}[x^*]\|^2 \leq \|x_t - \gamma V_t - x^*\|^2 \\
&\leq \|x_t - x^*\|^2 - 2\gamma\langle V_t, x_t - x^*\rangle + \gamma^2\|V_t\|^2 \\
&\leq \|x_t - x^*\|^2 - 2\gamma\langle V(x_t), x_t - x^*\rangle - 2\gamma\langle U_t(x_t), x_t - x^*\rangle + \gamma^2\|V_t\|^2 \\
&\leq \|x_t - x^*\|^2 - 2\gamma\langle V(x_t), x_t - x^*\rangle - 2\gamma\langle U_t(x_t), x_t - x^*\rangle + 2\gamma^2\|V(x_t)\|^2 + 2\gamma^2\|U_t(x_t)\|^2 \\
&\leq \|x_t - x^*\|^2 - 2\gamma(\mu\|x_t - x^*\|^2 - \lambda) - 2\gamma\langle U_t(x_t), x_t - x^*\rangle + 2\gamma^2\|V(x_t)\|^2 + 2\gamma^2\|U_t(x_t)\|^2 \\
&\leq (1-2\gamma\mu)\|x_t - x^*\|^2 + 2\gamma\lambda - 2\gamma\langle U_t(x_t), x_t - x^*\rangle + 4\gamma^2(L^2((1+\mathrm{R})^2 + \|x_t - x^*\|^2) + \tfrac{\|U_t(x_t)\|^2}{2}) \\
&\leq (1-2\gamma\mu + 4\gamma^2L^2)\|x_t - x^*\|^2 + 2\gamma\lambda - 2\gamma\langle U_t(x_t), x_t - x^*\rangle + 4\gamma^2(L^2(1+\mathrm{R})^2 + \tfrac{\|U_t(x_t)\|^2}{2}).
\end{aligned} \tag{B.2}$$

In the initial inequality, we employ the Lipschitz property of the projection operator. The subsequent inequality stems from the squared expansion. For the third and fourth inequalities, we invoke the definition of the noisy operator model along with Fact 1. The fifth inequality draws upon Assumption 6, highlighting monotonicity. The sixth leans on the linear growth as detailed in Assumption 3, in conjunction with Fact 1 and Assumption 1. Finally, the last inequality is derived using the Cauchy-Schwarz inequality. By taking the expectation condition on the filtration $\mathcal{F}_t$, we derive the following bound:

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 \,|\, \mathcal{F}_t] \leq (1-2\gamma\mu + 4\gamma^2L^2)\|x_t - x^*\|^2 + 2\gamma\lambda + 2\gamma b\|x_t - x^*\| + 4\gamma^2(L^2(1+\mathrm{R})^2 + \tfrac{\sigma^2 + \rho\|x_t - x^*\|^2}{2}), \tag{B.3}$$

where we have used that x_t is $\mathcal{F}_t-$measurable and by Assumption 4 we have that

$$\begin{cases}
\mathbb{E}[\|U_t(x_t)\|^2 \,|\, \mathcal{F}_t] \leq \sigma^2 + \rho\|x_t - x^*\|^2 \\
-2\gamma\langle \mathbb{E}[U_t(x_t) \,|\, \mathcal{F}_t], x_t - x^*\rangle \leq 2\gamma\|\mathbb{E}[U_t(x_t) \,|\, \mathcal{F}_t]\|\|x_t - x^*\| \leq 2b\|x_t - x^*\|
\end{cases} \tag{B.4}$$

While employing Jensen's inequality could merge the bounds, we opt to separate variance and bias. This distinction emphasizes the terms that would be redundant if the stochastic oracle were unbiased. Re-ordering the terms we get that:

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 \,|\, \mathcal{F}_t] \leq (1-2\gamma\mu + 4\gamma^2L^2 + 2\gamma^2\rho)\|x_t - x^*\|^2 + 2\gamma\lambda + 2\gamma b\|x_t - x^*\| + 4\gamma^2(L^2(1+\mathrm{R})^2 + \sigma^2/2) \tag{B.5}$$

Finally, for the bias term we will use the fundamental inequality : $2\langle x, y\rangle \leq \|x\|^2 + \|y\|^2$

$$2\gamma b\|x_t - x^*\| \leq \gamma\frac{b^2}{\mu} + \gamma\mu\|x_t - x^*\|^2$$

Thus combining (B.5) we get the following final bound:

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 \,|\, \mathcal{F}_t] \leq \underbrace{(1-\gamma\mu + 4\gamma^2L^2 + 2\gamma^2\rho)}_{\mathrm{Term}_0}\|x_t - x^*\|^2 + \gamma\cdot\underbrace{2(\lambda + \tfrac{b^2}{\mu})}_{\mathrm{Term}_1} + \gamma^2\cdot\underbrace{4(L^2(1+\mathrm{R})^2 + \sigma^2/2)}_{\mathrm{Term}_2} \tag{B.6}$$

Thus if we request $\gamma^{\mathrm{SGDA}} := \gamma \leq \dfrac{\mu/2}{4L^2 + 2\rho} \Rightarrow \begin{cases} \mathrm{Term}_0 \leq 1 \\ \gamma^2(4L^2 + 2\rho) \leq \dfrac{\mu\gamma}{2} \end{cases}$ which implies $\mathrm{Term}_0 \leq (1 - \dfrac{\mu\gamma}{2})$. Therefore, we conclude that

$$\mathbb{E}[\|x_t - x^*\|^2 \,|\, \mathcal{F}_t] \leq (1 - \tfrac{\gamma\mu}{2})^t\|x_0 - x^*\|^2 + \cancel{\gamma}\cdot\frac{2(\lambda + \frac{b^2}{\mu})}{\cancel{\gamma}\mu/2} + \cancel{\gamma^2}\cdot\frac{4(L^2(1+\mathrm{R})^2 + \sigma^2/2)}{\cancel{\gamma}\mu/2} \tag{B.7}$$

which proves the claim of the lemma for

$$c := c_1^{\mathrm{SGDA}} = \Theta\left(\gamma\mu\right) \text{ and } c' := c_2^{\mathrm{SGDA}} = \Theta\left(\frac{\lambda\mu + b^2}{\mu^2} + \gamma\frac{\sigma^2 + L^2(1+\mathrm{R}^2)}{\mu}\right)$$

■

We proceed on proving a similar lemma for the case of (SEG).

Lemma B.4. *Suppose that Assumptions 1–4 hold then the iterations $(x_t)_{t\geq 0}$ of (SEG), if the step-size $\gamma \leq \frac{1}{18}(\frac{1}{\ell} \wedge \frac{\mu}{\rho})$, satisfy:*

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 \,|\, \mathcal{F}_t] \leq (1-c)^t \|x_0 - x^*\|^2 + c' \tag{B.8}$$

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 using the 1-co-coerciveness of the projection operator:

$$\|\Pi_{\mathcal{X}}(w) - \Pi_{\mathcal{X}}(w^*)\|^2 \leq \langle w - w^*, \Pi_{\mathcal{X}}(w) - \Pi_{\mathcal{X}}(w^*)\rangle$$

Thus we have

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= \|\Pi_{\mathcal{X}}\left(x_t - \gamma V_{t+1/2}\right) - \Pi_{\mathcal{X}}(x^*)\|^2 && \text{(B.9)}\\
&\leq \langle x_{t+1} - x^*, x_t - \gamma V_{t+1/2} - x^*\rangle && \text{(B.10)}\\
&= \frac{1}{2}\|x_{t+1} - x^*\|^2 + \frac{1}{2}\|x_t - x^*\|^2 - \frac{1}{2}\|x_t - x_{t+1}\|^2 - \gamma\langle x_{t+1} - x^*, V_{t+1/2}\rangle && \text{(B.11)}
\end{aligned}$$

We will analyze the last term:

$$-\gamma\langle x_{t+1} - x^*, V_{t+1/2}\rangle = \mathrm{Term_A} + \mathrm{Term_B} + \mathrm{Term_C} \tag{B.12}$$

where we define the following terms:

$$\begin{cases}
\mathrm{Term_A} = -\gamma\langle x_{t+1} - x_{t+1/2}, V_t\rangle\\
\mathrm{Term_B} = \gamma\langle x_{t+1} - x_{t+1/2}, V_t - V_{t+1/2}\rangle\\
\mathrm{Term_C} = -\gamma\langle x_{t+1/2} - x^*, V_{t+1/2}\rangle
\end{cases}$$

For the $\mathrm{Term_A}$, we first recall that by optimality conditions we have that

$$x_{t+1/2} := \Pi_{\mathcal{X}}(x - \gamma V_t) \Rightarrow \langle x_{t+1/2} - (x_t - \gamma V_t), x - x_{t+1/2}\rangle \geq 0 \quad \forall x \in \mathcal{X}$$

and by setting $x = x_{t+1}$, we get

$$\langle x_{t+1} - (x_t - \gamma V_t), x_{t+1} - x_{t+1/2}\rangle \geq 0 \tag{B.13}$$

Thus, by reordering of the terms we derive that

$$\begin{aligned}
-\gamma\langle V_t, x_{t+1} - x_{t+1/2}\rangle &\leq \langle x_{t+1} - x_t, x_{t+1} - x_{t+1/2}\rangle\\
&= \frac{1}{2}\|x_{t+1} - x_t\|^2 - \frac{1}{2}\|x_t - x_{t+1/2}\|^2 - \frac{1}{2}\|x_{t+1} - x_{t+1/2}\|^2
\end{aligned}$$

Therefore, it holds that:

$$\mathrm{Term_A} \leq \frac{1}{2}\left(\|x_{t+1} - x_t\|^2 - \|x_t - x_{t+1/2}\|^2 - \|x_{t+1} - x_{t+1/2}\|^2\right) \tag{B.14}$$

For the $\mathrm{Term_B}$, we start by using Cauchy-Schwarz:

$$\mathrm{Term_B} = \gamma\langle x_{t+1} - x_{t+1/2}, V_t - V_{t+1/2}\rangle \leq \gamma\|x_{t+1} - x_{t+1/2}\|\|V_t - V_{t+1/2}\| \tag{B.15}$$

Then, by applying triangle inequality and Assumption 3 and the definition of stochastic oracle:

$$\|V_{t+1/2} - V_t\| \leq \|U_{t+1/2}(x_{t+1/2})\| + \|U_t(x_t)\| + \ell\|x_t - x_{t+1/2}\|$$

Leveraging the standard inequality $\langle w, z\rangle \leq \frac{1}{2}\|x\|^2 + \frac{1}{2}\|z\|^2$, we get that:

$$\gamma\|x_{t+1} - x_{t+1/2}\|\|V_{t+1/2} - V_t\| \leq \frac{1}{2}\|x_{t+1} - x_{t+1/2}\|^2 + \frac{1}{2}\gamma^2\|V_t - V_{t+1/2}\|^2$$

Thus, it holds that:

$$\text{Term}_\text{B} \leq \frac{1}{2}\|x_{t+1} - x_{t+1/2}\|^2 + \frac{3}{2}\gamma^2\left(\|U_{t+1/2}(x_{t+1/2})\|^2 + \|U_{t+1/2}(x_{t+1/2})\|^2 + \ell^2\|x_t - x_{t+1/2}\|^2\right) \tag{B.16}$$

Finally for Term_C, using the definition of stochastic oracle for operator g and Assumption 2, we get

$$\begin{aligned}\text{Term}_\text{C} &= -\gamma\langle x_{t+1/2} - x^*, V_{t+1/2}\rangle = -\gamma\langle x_{t+1/2} - x_1, V(x_{t+1/2})\rangle - \gamma\langle x_{t+1/2} - x_1, U\left(x_{t+1/2}\right)\rangle \\ &\leq -\gamma\mu\|x_{t+1/2} - x^*\|_2^2 + \gamma\lambda - \gamma\langle x_{t+1/2} - x^*, U_{t+1/2}\left(x_{t+1/2}\right)\rangle\end{aligned}$$

By combining the bounds for all terms, we get:

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &= \frac{1}{2}\|x_{t+1} - x^*\|^2 + \frac{1}{2}\|x_t - x^*\|^2 - \frac{1}{2}\|x_t - x_{t+1}\|^2 + \text{Term}_\text{A} + \text{Term}_\text{B} + \text{Term}_\text{C} && \text{(B.17)} \\ &\leq \frac{1}{2}\|x_{t+1} - x^*\|^2 + \frac{1}{2}\|x_t - x^*\|^2 + \frac{1}{2}(3\gamma^2\ell^2 - 1)\|x_t - x_{t+1/2}\|^2 \\ &+ \frac{3}{2}\gamma^2\left(\|U_{t+1/2}(x_{t+1/2})\|^2 + \|U_t(x_t)\|^2\right) \\ &- \gamma\mu\|x_{t+1/2} - x^*\|_2^2 + \gamma\lambda - \gamma\langle x_{t+1/2} - x^*, U_{t+1/2}\left(x_{t+1/2}\right)\rangle && \text{(B.18)}\end{aligned}$$

Rewriting the bound, we get that

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &\leq \|x_t - x^*\|^2 + (3\gamma^2\ell^2 - 1)\|x_t - x_{t+1/2}\|^2 + 3\gamma^2\left(\|U_{t+1/2}(x_{t+1/2})\|^2 + \|U_t(x_t)\|^2\right) \\ &- 2\gamma\mu\|x_{t+1/2} - x^*\|^2 + 2\gamma\lambda - 2\gamma\langle x_{t+1/2} - x^*, U\left(x_{t+1/2}\right)\rangle && \text{(B.19)} \\ &\leq (1 - 2\gamma\mu)\|x_t - x^*\|^2 + (3\gamma^2\ell^2 + 4\gamma\mu - 1)\|x_t - x_{t+1/2}\|^2 + 2\gamma\lambda \\ &+ 3\gamma^2\left(\|U_{t+1/2}(x_{t+1/2})\|^2 + \|U_t(x_t)\|^2\right) + 2\gamma\|x_{t+1/2} - x^*\|\|U\left(x_{t+1/2}\right)\| && \text{(B.20)} \\ &\leq (1 - 2\gamma\mu)\|x_t - x^*\|^2 + (3\gamma^2\ell^2 + 4\gamma\mu - 1)\|x_t - x_{t+1/2}\|^2 + 2\gamma\lambda \\ &+ 3\gamma^2\left(\|U_{t+1/2}(x_{t+1/2})\|^2 + \|U_t(x_t)\|^2\right) - 2\gamma\langle x_{t+1/2} - x^*, U_{t+1/2}\left(x_{t+1/2}\right)\rangle && \text{(B.21)}\end{aligned}$$

By taking the expectation condition on the filtration $\mathcal{F}_t$ and tower property ($\mathbb{E}[\cdot \,|\, \mathcal{F}_t] = \mathbb{E}[\mathbb{E}[\cdot \,|\, \mathcal{F}_{t+1/2}] \,|\, \mathcal{F}_t]$) and using Assumption 4, we derive the following bound:

$$\begin{aligned}\mathbb{E}[\|x_{t+1} - x^*\|^2 \,|\, \mathcal{F}_t] &\leq (1 - 2\gamma\mu)\|x_t - x^*\|^2 + (3\gamma^2\ell^2 + 4\gamma\mu - 1)\,\mathbb{E}[\|x_t - x_{t+1/2}\|^2 \,|\, \mathcal{F}_t] + 2\gamma\lambda \\ &+ 3\gamma^2\left(\mathbb{E}[\sigma^2 + \rho\|x_{t+1/2} - x^*\|^2 \,|\, \mathcal{F}_t] + \sigma^2 + \rho\|x_t - x^*\|^2\right) \\ &- 2\gamma\,\mathbb{E}[\langle x_{t+1/2} - x^*, \mathbb{E}[U_{t+1/2}\left(x_{t+1/2}\right) \,|\, \mathcal{F}_{t+1/2}]\rangle \,|\, \mathcal{F}_t] && \text{(B.22)} \\ &\leq (1 - 2\gamma\mu)\|x_t - x^*\|^2 + (3\gamma^2\ell^2 + 4\gamma\mu - 1)\,\mathbb{E}[\|x_t - x_{t+1/2}\|^2 \,|\, \mathcal{F}_t] + 2\gamma\lambda \\ &+ 3\gamma^2\left(\mathbb{E}[\sigma^2 + 2\rho\|x_t - x^*\|^2 + 2\rho\|x_{t+1/2} - x_t\|^2 \,|\, \mathcal{F}_t] + \sigma^2 + \rho\|x_t - x^*\|^2\right) \\ &+ 2\gamma\,\mathbb{E}[\|x_{t+1/2} - x^*\| \,|\, \mathcal{F}_t] \cdot b && \text{(B.23)} \\ &\leq (1 - 2\gamma\mu + 9\gamma^2\rho)\|x_t - x^*\|^2 + (3\gamma^2(\ell^2 + 2\rho) + 4\gamma\mu - 1)\,\mathbb{E}[\|x_t - x_{t+1/2}\|^2 \,|\, \mathcal{F}_t] \\ &+ 2\gamma\lambda + 6\gamma^2\sigma^2 + \gamma\,\mathbb{E}[2\tfrac{\mu}{2}\|x_t - x^*\|^2 + 2\tfrac{\mu}{2}\|x_{t+1/2} - x_t\|^2 + \tfrac{2}{\mu}b^2 \,|\, \mathcal{F}_t] && \text{(B.24)} \\ &\leq \underbrace{(1 - \gamma\mu + 9\gamma^2\rho)}_{\text{Term}_0}\|x_t - x^*\|^2 + \underbrace{(3\gamma^2(\ell^2 + 2\rho) + 5\gamma\mu - 1)}_{\text{Term}_1}\,\mathbb{E}[\|x_t - x_{t+1/2}\|^2 \,|\, \mathcal{F}_t] \\ &+ 2\gamma\lambda + 6\gamma^2\sigma^2 + \gamma\tfrac{2}{\mu}b^2 && \text{(B.25)}\end{aligned}$$

Thus by inverting the requirements and recalling $\sqrt{1+x}-1 \geq \sqrt{x}/3$ for $x \geq 1$, we get:

$$\left\{\begin{matrix} 1-\gamma\mu+9\gamma^2\rho \leq 1 \\ 3\gamma^2(\ell^2+2\rho)+5\gamma\mu-1 \leq 0 \end{matrix}\right\} \Leftarrow \left\{\begin{matrix} \gamma \leq \frac{\mu}{9\rho} \\ \gamma \leq \frac{-5\mu+\sqrt{25\mu^2+12(\ell^2+2\rho)}}{6(\ell^2+2\rho)} = \frac{5\mu}{6(\ell^2+2\rho)} \cdot \left(\sqrt{1+12\frac{(\ell^2+2\rho)}{25\mu^2}}-1\right) \end{matrix}\right\}$$
$$\Leftarrow \left\{\begin{matrix} \gamma \leq \frac{\mu}{9\rho} \\ \gamma \leq \frac{\sqrt{12}}{18\sqrt{(\ell^2+2\rho)}} \end{matrix}\right\} \Leftarrow \left\{\begin{matrix} \gamma \leq \frac{\mu}{9\rho} \\ \gamma \leq \frac{1}{9\ell\sqrt{1+2\frac{\rho}{\ell^2}}} \end{matrix}\right\} \Leftarrow \left\{\gamma \leq \tfrac{1}{9}(\tfrac{\mu}{\rho} \wedge \tfrac{1}{\ell}).\right\} \tag{B.26}$$

Thus if we request $\gamma^{\text{SEG}} := \gamma \leq \frac{1}{18}(\frac{\mu}{\rho} \wedge \frac{1}{\ell}) \Rightarrow \left\{\begin{matrix} \text{Term}_0 \leq 1 \\ 9\gamma^2\rho \leq \frac{\mu\gamma}{2} \\ \text{Term}_1 \leq 0 \end{matrix}\right\}$ which implies $\left\{\begin{matrix} \text{Term}_0 \leq (1-\frac{\mu\gamma}{2}) \\ \text{Term}_1 \leq 0 \end{matrix}\right\}$. Therefore, we derive the following bound:

$$\mathbb{E}[\|x_{t+1}-x^*\|^2 \,|\, \mathcal{F}_t] \leq (1-\gamma\mu+9\gamma^2\rho)\|x_t-x^*\|^2 + 2\gamma\lambda + 6\gamma^2\sigma^2 + \gamma\tfrac{2}{\mu}b^2 \tag{B.27}$$

Hence, we conclude that

$$\mathbb{E}[\|x_t-x^*\|^2 \,|\, \mathcal{F}_t] \leq (1-\tfrac{\gamma\mu}{2})^t\|x_0-x^*\|^2 + \cancel{\gamma} \cdot \frac{2(\lambda+\frac{b^2}{\mu})}{\cancel{\gamma}\mu/2} + \gamma^{\cancel{2}} \cdot \frac{6\sigma^2}{\cancel{\gamma}\mu/2} \tag{B.28}$$

which proves the claim of the lemma for

$$c := c_1^{\text{SEG}} = \Theta\left(\gamma\mu\right) \text{ and } c' := c_2^{\text{SEG}} = \Theta\left(\frac{\lambda\mu+b^2}{\mu^2} + \gamma\frac{\sigma^2}{\mu}\right)$$

■

B.3 Proof of Theorem 1

Theorem B.1 (Restated Theorem 1). *Under Assumptions 1–4, consider (SGDA) and (SEG) with step-sizes $\gamma^{\text{SGDA}} = \mathcal{O}(\frac{\mu}{L^2+\rho})$, $\gamma^{\text{SEG}} = \mathcal{O}(\frac{1}{\ell} \wedge \frac{\mu}{\rho})$ respectively, and let $(x_t)_{t\geq 0}$ be the generated iterations. There exists a pair of constants $c_1^{Alg} \in (0,1)$ and $c_2^{Alg} \in (0,+\infty)$ that depend on the choice of step-sizes and model's parameters such that*

$$\mathbb{E}[\|x_{t+1}-x^*\|^2] \leq \left(1-c_1^{Alg}\right)^t\|x_0-x^*\|^2 + c_2^{Alg}, \tag{B.29}$$

for any initial point $x_0 \in \mathcal{X}$ and $Alg \in \{\text{SGDA,SEG}\}$.

Proof. Proof follows by combining Lemma B.3 and B.4. ■

B.4 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, there exist constants $\widehat{c_1^{Alg}} \in (0,1)$ and $\widehat{c_2^{Alg}} \in (0,\infty)$ with $Alg \in \{\text{SGDA,SEG}\}$ such that the function $\mathcal{E}(x_t,x^*) := \|x_t-x^*\|^2+1$ satisfies*

$$\mathbb{E}[\mathcal{E}(x_{t+1},x^*) \,|\, \mathcal{F}_t] \leq \widehat{c_1^{Alg}}\mathcal{E}(x_t,x^*) + \widehat{c_2^{Alg}}. \tag{B.30}$$

Proof. For (SGDA), by (B.6) in the proof of Lemma B.3, we have

$$\mathbb{E}[\|x_{t+1}-x^*\|^2 \,|\, \mathcal{F}_t] \leq (1-\gamma\mu+4\gamma^2L^2+2\gamma^2\rho)\|x_t-x^*\|^2 + \gamma\cdot 2(\lambda+\tfrac{b^2}{\mu}) + \gamma^2\cdot 4(L^2(1+\mathrm{R})^2+\sigma^2/2)$$

Following $\gamma \leq \gamma^{\text{SGDA}} \leq \frac{\mu/8}{L^2+\rho}$

$$\mathbb{E}[\|x_{t+1}-x^*\|^2+1 \,|\, \mathcal{F}_t] \leq (1-\gamma\mu/2)(\|x_t-x^*\|^2+1) + 2\gamma\cdot(\lambda+\tfrac{b^2}{\mu}+\tfrac{\mu}{4}) + \gamma^2\cdot 4(L^2(1+\mathrm{R})^2+\sigma^2/2).$$

Let $\widehat{c_1^{\mathrm{SGDA}}} = (1-\gamma\mu/2)$ and $\widehat{c_2^{\mathrm{SGDA}}} = \gamma \cdot 2(\lambda + \frac{b^2}{\mu} + \frac{\mu}{4}) + \gamma^2 \cdot 4(L^2(1+\mathrm{R})^2 + \sigma^2/2)$.

For (SEG), by (B.27) in the proof of Lemma B.4, we have

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 \,|\, \mathcal{F}_t] \leq (1 - \gamma\mu + 9\gamma^2\rho)\|x_t - x^*\|^2 + 2\gamma\lambda + 6\gamma^2\sigma^2 + \gamma\tfrac{2}{\mu}b^2.$$

Following $\gamma \leq \gamma^{\mathrm{SEG}} \leq \frac{1}{18}(\frac{\mu}{\rho} \wedge \frac{1}{\ell})$

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 + 1 \,|\, \mathcal{F}_t] \leq (1 - \gamma\mu/2)(\|x_t - x^*\|^2 + 1) + 2\gamma\lambda + 6\gamma^2\sigma^2 + \gamma\tfrac{2}{\mu}b^2.$$

Let $\widehat{c_1^{\mathrm{SEG}}} = (1-\gamma\mu/2)$ and $\widehat{c_2^{\mathrm{SEG}}} = 2\gamma(\lambda + \frac{b^2}{\mu} + \frac{\mu}{4}) + 6\gamma^2\sigma^2$.
which conclude our proof. ■

C Omitted Proofs of Section 4

C.1 Clarification about Borel Algebra in Constrained sets

For the following section and establishing Lemma 1, it is critical to clairify the defintion of the Borel σ-algebra over unconstrained and constrained space $\mathcal{X}$. More formally, we have that:

Definition of Borel σ-Algebra: Given a topological space X, the Borel σ-algebra, denoted as $\mathcal{B}(X)$, is the σ-algebra generated by the open sets of X. This means that $\mathcal{B}(X)$ is the smallest σ-algebra that contains all the open sets of X.

Unconstrained Case: $\mathbb{R}^d$. When X corresponds to $\mathbb{R}^d$, the d-dimensional Euclidean space, the Borel σ-algebra $\mathcal{B}(\mathbb{R}^d)$ is generated by the open sets in $\mathbb{R}^d$. This includes all open intervals, open balls, and any set that can be obtained from open sets through countable unions, intersections, and complements.

Constrained Case: Subsets of $\mathbb{R}^d$. When considering a subset $\mathcal{X} \subseteq \mathbb{R}^d$, we can define the Borel σ-algebra on $\mathcal{X}$ in the following two common ways:

1. **Relative Topology**: $\mathcal{B}(\mathcal{X})$ is generated by the open sets of $\mathcal{X}$ with respect to the relative topology induced by $\mathbb{R}^d$. That is, a set U is open in $\mathcal{X}$ if there exists an open set V in $\mathbb{R}^d$ such that $U = V \cap \mathcal{X}$.

2. **Subspace Borel σ-Algebra**: You can also consider the Borel σ-algebra on $\mathcal{X}$ as the collection of sets $\{A \cap \mathcal{X} \mid A \in \mathcal{B}(\mathbb{R}^d)\}$. This is the Borel σ-algebra on $\mathcal{X}$ as a subspace of $\mathbb{R}^d$.

Both approaches yield a Borel σ-algebra on $\mathcal{X}$ that allows for the definition of measures and the integration of functions over $\mathcal{X}$. While results hold true for both methodologies, without loss of generality, we will adopt the second approach for reasons of mathematical convention.

Also we equip $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ with the natural Lebesgue measure for the underlying subspace $\mathcal{X}$. See further details for case of $\mathbb{S}^{n-1}$ Sphere Wikipedia (2023) or case of simplex or the general Riemannian manifold case (Morvan, 2008, Chapter 5.2).

Additionally, we recall the following crucial dimensionality fact:

Fact 2. Let $\mathcal{X}$ be a convex subset of $\mathbb{R}^d$, and for each $x \in \mathcal{X}$, let $\mathcal{N}_{\mathcal{X}}(x)$ be the normal cone to $\mathcal{X}$ at x. Then, the union of $x + \mathcal{N}_{\mathcal{X}}(x)$ over all $x \in \mathcal{X}$ spans $\mathbb{R}^d$.

C.2 Minorization Condition and Geometric Drift Property

Lemma C.1 (Restated Lemma 1)**.** *Let the Assumptions 1–5 be satisfied for (SGDA) and (SEG). 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*

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

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

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 $C(r) := \{x \in \mathcal{X} | \mathcal{E}(x) \leq r\}$ are either empty or small for all $r > 0$. Indeed assume that $C(r) = \{x \in \mathcal{X} | \mathcal{E}(x) \leq r\}$ is non-empty ($r > 1$), then the sublevel sets correspond to some ball $\mathbb{B}(x^*, \sqrt{r-1})$ for $r > 1$ intersected with $\mathcal{X}$. We will prove that such a set $C(r) := \mathcal{X} \cap \mathbb{B}(x^*, \sqrt{r-1})$ for $r > 1$ is actually ν_1-small for $m = 1$ (see Definition A.6).

SGDA: We recall the definition for (SGDA) kernel for an arbitrary set A of $\mathcal{B}(\mathcal{X})$:

$$\Pr[x_{t+1} \in A | x_t = x] = P_{(\text{SGDA})}(x, A) = \mathbb{P}\left(U(x) \in (\frac{x}{\gamma^{\text{SGDA}}} - V(x)) + (-\frac{1}{\gamma^{\text{SGDA}}}\hat{A})\right),$$

where $A = \mathcal{X} \cap \tilde{A}$ for some $\tilde{A} \in \mathcal{B}(\mathbb{R}^d)$, $\hat{A} = \bigcup_{\chi \in A} (\chi + \mathcal{N}_{\mathcal{X}}(\chi))$ and $(U_t(x_t)(x))_{t\geq 0} \sim U(x)$, since we assume i.i.d noise random fields. Thus, we have

$$P(x, A) = \int_{\alpha \in \hat{A}} \mathrm{pdf}_{U(x)}(\frac{x-\alpha}{\gamma} - V(x))\mathrm{d}\alpha \geq \int_{\alpha \in \hat{A}} \inf_{x \in C(r)} \mathrm{pdf}_{U(x)}(\frac{x-\alpha}{\gamma} - V(x))\mathrm{d}\alpha \tag{C.1}$$

$$\geq \int_{\alpha \in \hat{A}} \inf_{x \in C(r)} \mathrm{pdf}_{U(x)}(\frac{x-\alpha}{\gamma} - V(x))\mathrm{d}\alpha \tag{C.2}$$

$$= \int_{\alpha \in \bigcup_{\chi \in A}(\chi + \mathcal{N}_{\mathcal{X}}(\chi))} \inf_{x \in C(r)} \mathrm{pdf}_{U(x)}(\frac{x-\alpha}{\gamma} - V(x))\mathrm{d}\alpha := \nu_r^{\mathrm{SGDA}}(A). \tag{C.3}$$

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

$$\nu_r^{\mathrm{SGDA}}(C(r)) = \int_{x' \in (\bigcup_{\chi \in C(r)}(\chi + \mathcal{N}_{\mathcal{X}}(\chi)))} \inf_{x \in C(r)} \mathrm{pdf}_{U(x)}\left(\frac{x-x'}{\gamma} - V(x)\right)\mathrm{d}x' > 0,$$

which follows from Assumption 5 and the fact that $\bigcup_{\chi \in C(r)}(\chi + \mathcal{N}_{\mathcal{X}}(\chi))$ has positive Lebesgue measure thanks to Fact 2 and $C(r) = \mathcal{X} \cap \mathbb{B}(x^*, \sqrt{r-1})$ and non-empty.

We now fix $r = r_0 > 1$ and proceed in proving the minorization property. Consider the measure

$$\tilde{\nu}_{r_0}^{\mathrm{SGDA}}(X) = \mathbb{1}(X \subseteq C(r_0)) \frac{\nu_{r_0}^{\mathrm{SGDA}}(X)}{\nu_{r_0}^{\mathrm{SGDA}}(C(r_0))} \text{ for all } X \in \mathcal{B}(\mathcal{X}).$$

- It is easy to verify that $\tilde{\nu}_{r_0}^{\mathrm{SGDA}}(C(r_0)) = 1$ and $\tilde{\nu}_{r_0}^{\mathrm{SGDA}}(C(r_0)^c) = 0$.
- Additionally, if $\{x \notin C(r_0) \text{ or } A \nsubseteq C(r_0)\}$ we have that $P(x, A) \geq \delta\, \mathbb{1}_{C(r_0)}(x) \tilde{\nu}_{r_0}^{\mathrm{SGDA}}(A) = 0$.
- Also, if $\{x \in C(r_0) \text{ and } A \subseteq C(r_0)\}$ we have $P(x, A) \geq \nu_{r_0}^{\mathrm{SGDA}}(A) = \delta\, \mathbb{1}_{C(r_0)}(x) \tilde{\nu}_{r_0}^{\mathrm{SGDA}}(A)$, where $\delta = \nu_{r_0}^{\mathrm{SGDA}}(C(r_0)) > 0$ and thus the proof is completed.

SEG: We continue with the proof when (SEG) is run. Similarly as before we have for an arbitrary set $A \in \mathcal{B}(\mathcal{X})$:

$$\begin{aligned}\Pr[x_{t+1} \in A | x_t = x] &= P_{(\mathrm{SEG})}(x, A) \\ &= \int_{\xi \in \mathbb{R}^d} \mathrm{pdf}_{U^{(i)}(x)}(\xi)\, \mathbb{P}\left(U^{(ii)}(\tilde{x}(x,\xi)) \in (\frac{x}{\gamma} - V(\tilde{x}(x,\xi)) + (-\frac{1}{\gamma}\hat{A})\right)\mathrm{d}\xi\end{aligned}$$

where we denote *(a)* $A = \mathcal{X} \cap \tilde{A}$ for some arbitrary $\tilde{A} \in \mathcal{B}(\mathbb{R}^d)$, *(b)* $\hat{A} = \bigcup_{\chi \in A}(\chi + \mathcal{N}_{\mathcal{X}}(\chi))$ as previously, *(c)* $V(x) = v_x$, *(d)* $U_t(x_t)(x) \sim \mathrm{law}(U^{(i)}(x))$, $U_{t+1/2}(x) \sim \mathrm{law}(U^{(ii)}(x))$ & $U^{(i)}(x) \perp U^{(ii)}(x)$, identically distributed and *(e)* $\tilde{x}(x, \xi) = \Pi_{\mathcal{X}}[x - \gamma\{v_x + \xi\}]$. Therefore, we get

$$\begin{aligned}\Pr[x_{t+1} \in A | x_t = x] &= P_{(\mathrm{SEG})}(x, A) \\ &= \int_{\xi \in \mathbb{R}^d} \mathrm{pdf}_{U^{(i)}(x)}(\xi)\, \mathbb{P}\left(U^{(ii)}(\tilde{x}(x,\xi)) \in (\frac{x}{\gamma} - V(\tilde{x}(x,\xi)) + (-\frac{1}{\gamma}\hat{A})\right)\mathrm{d}\xi \\ &= \int_{\alpha \in \hat{A}} \int_{\xi \in \mathbb{R}^d} \mathrm{pdf}_{U^{(i)}(x)}(\xi) \mathrm{pdf}_{U^{(ii)}(\tilde{x}(x,\xi))}\left(\frac{x-\alpha}{\gamma} - V(\tilde{x}(x,\xi))\right) \\ &\geq \int_{\alpha \in \hat{A}} \int_{\xi \in \mathbb{B}(0,1)} \mathrm{pdf}_{U^{(i)}(x)}(\xi) \mathrm{pdf}_{U^{(ii)}(\tilde{x}(x,\xi))}\left(\frac{x-\alpha}{\gamma} - V(\tilde{x}(x,\xi))\right)\mathrm{d}\xi\mathrm{d}\alpha\end{aligned}$$

Notice that since $x \in C(r)$ and $\xi \in \mathbb{B}(0,1)$, we have that $\tilde{x}(x,\xi) \in C'(r) := \Big(C(r) + (-\gamma)\cdot(V(C(r)) + \mathbb{B}(0,1))\Big) \cap \mathcal{X}$. Thus, under Assumption 5, we have that

1. $\mathrm{pdf}_{U^{(i)}(x)}(t) \geq \inf_{x \in C(r)} \mathrm{pdf}_{U^{(i)}(x)}(t) > 0$ for all $t \in \mathbb{R}^d$.
2. $\mathrm{pdf}_{U^{(ii)}(\tilde{x}(x,\xi))}(t) \geq \inf_{\rho \in C'(r)} \mathrm{pdf}_{U^{(ii)}(\rho)}(t) > 0$ for all $t \in \mathbb{R}^d$.

Hence, we can define the following measure for any set $A \in \mathcal{B}(\mathcal{X})$:

$$\nu_r^{\mathrm{SEG}}(A) := \int_{\alpha \in \hat{A}} \int_{\xi \in \mathbb{B}(0,1)} \inf_{x \in C(r)} \mathrm{pdf}_{U^{(i)}(x)}(\xi) \inf_{\rho \in C'(r)} \mathrm{pdf}_{U^{(ii)}(\rho)}\left(\frac{x-\alpha}{\gamma} - V(\tilde{x}(x,\xi))\right)\mathrm{d}\xi\mathrm{d}\alpha.$$

Notice that the measure is non-trivial since we have that $\nu_r^{\mathrm{SEG}}(C(r)) > 0$ which follows from Assumption 5 and the fact that $\bigcup_{\chi \in C(r)}(\chi + \mathcal{N}_{\mathcal{X}}(\chi))$ has positive Lebesgue measure thanks to Fact 2 and $C(r) = \mathcal{X} \cap \mathbb{B}(x^*, \sqrt{r-1})$ and non-empty.

As in the case of SGDA we define for some fixed $r_0 > 1$

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

Thus, we have that

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

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 Meyn and Tweedie (2009) 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*

$$\Delta f_i(x) \leq -\beta f_i(x) + b\,\mathbb{1}_C(x), x \in \mathbb{R}^d, \tag{C.4}$$

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

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

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 for any $Alg \in \{SGDA, SEG\}$

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

with $\widehat{c_1^{\mathrm{Alg}}} \in (0, 1)$ and $\widehat{c_2^{\mathrm{Alg}}} \in (0, +\infty)$.

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

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

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.3 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 $\psi-$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

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

where $\tilde{A}\subseteq\mathcal{B}(\mathbb{R}^d)$, $A=\tilde{A}\cap\mathcal{X}\neq\emptyset\in\mathcal{B}(\mathcal{X})$ and $\hat{A}=\bigcup_{\chi\in A}(\chi+\mathcal{N}_{\mathcal{X}}(\chi))$.

Under the "naturally" induced Lebesgue measure over $\mathcal{X}$, it holds that for every $A=\tilde{A}\cap\mathcal{X}\subseteq\mathcal{B}(\mathcal{X})$ with $\psi(A)>0$, we have that $\exists\varepsilon>0$ such that $\mathbb{B}(a_0,\varepsilon)\subseteq\tilde{A}$ for some $a_0\in A$ such that $\mathbb{B}(a_0,\varepsilon)\cap\mathcal{X}\neq\emptyset$ and $\psi(\underbrace{\bigcup_{\chi\in\mathbb{B}(a_0,\varepsilon)\cap\mathcal{X}}(\chi+\mathcal{N}_{\mathcal{X}}(\chi))}_{\vec{A}_\epsilon})>0$,. Thus,

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

Similarly, for the case of (SEG) and by repeating the same argument for some non-zero σ-finite measure φ in $\mathcal{B}(\mathcal{X})$ algebra, we have for $\tilde{x}(x,\xi)=\Pi_{\mathcal{X}}[x-\gamma\{V(x)+\xi\}]$ that

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

where $C=(\mathbb{B}(x,1)-\gamma V(\mathbb{B}(x,1))-\gamma\mathbb{B}(0,1))\cap\mathcal{X}$. 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 Meyn and Tweedie (2009) 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 Meyn and Tweedie (2009). 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 for Alg* $\in \{\text{SGDA,SEG}\}$*:*

1. *The iterates* $(x_t)_{t\geq 0}$ *admit a unique invariant distribution* $\pi_\gamma^{Alg} \in \mathcal{P}_2(\mathcal{X})$*, where* $\mathcal{P}_2(\mathcal{X})$ *is the set of distributions on* $\mathcal{X}$ *with bounded second moments.*
2. *For any test function* $\phi : \mathcal{X} \to \mathbb{R}$ *of* L_ϕ*-linear growth and any initialization* $x_0 \in \mathcal{X}$*, there exist constants* $\tau_{\phi,\gamma}^{Alg} \in (0,1)$ *and* $\zeta_{\phi,x_0,\gamma}^{Alg} \in (0,\infty)$ *such that:*

$$\left|\mathbb{E}_{x_t}[\phi(x_t)] - \mathbb{E}_{x\sim\pi_\gamma^{Alg}}[\phi(x)]\right| \leq \zeta_{\phi,x_0,\gamma}^{Alg}(\tau_{\phi,\gamma}^{Alg})^t. \tag{C.5}$$

   Hence, (SGDA) and (SEG) converges geometrically under the total variation distance to π_γ^{Alg}*.*
3. *For each* ℓ_ϕ*-Lipschitz test function* ϕ*, it holds that*

$$|\mathbb{E}_{x\sim\pi_\gamma^{Alg}}[\phi(x)] - \phi(x^*)| \leq \ell_\phi\sqrt{D^{Alg}}, \tag{C.6}$$

   for some constant $D^{Alg} \propto c_2^{Alg}$*.*

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 Meyn and Tweedie (2009)) the measure is unique and finite. Additionally assume that $x_0 \sim \pi_\gamma^{\text{Alg}}$. Then by the invariance property $(x_t)_{t\geq 0} \sim \pi_\gamma^{\text{Alg}}$. Using Corollary 1 for some arbitrary fixed $x^* \in \mathcal{X}^*$, there exist two corresponding constants $(\widehat{c_1^{\text{Alg}}}, \widehat{c_2^{\text{Alg}}})$ such that $\widehat{c_1^{\text{Alg}}} \in (0,1)$ and $\widehat{c_2^{\text{Alg}}} \in (0,\infty)$ that satisfy

$$\mathbb{E}[\|x_{t+1} - x^*\|^2 + 1 \,|\, \mathcal{F}_t] \leq \widehat{c_1^{\text{Alg}}}(\|x_t - x^*\|^2 + 1) + \widehat{c_2^{\text{Alg}}} \Rightarrow \tag{C.7}$$

$$\mathbb{E}_{x\sim\pi_\gamma^{\text{Alg}}}[\|x - x^*\|^2] \leq \frac{\widehat{c_1^{\text{Alg}}} + \widehat{c_2^{\text{Alg}}} - 1}{1 - \widehat{c_1^{\text{Alg}}}} \tag{C.8}$$

At the same time, by construction it holds that

$$(\widehat{c_1^{\text{Alg}}}, \widehat{c_2^{\text{Alg}}}) := (1 - c_1^{\text{Alg}}, (1 + c_2^{\text{Alg}})c_1^{\text{Alg}})$$

which implies that:

$$\mathbb{E}_{x\sim\pi_\gamma^{\text{Alg}}}[\|x - x^*\|^2] \leq \frac{(1 - c_1^{\text{Alg}}) + (1 + c_2^{\text{Alg}})c_1^{\text{Alg}} - 1}{1 - (1 - c_1^{\text{Alg}})} = \frac{c_2^{\text{Alg}}c_1^{\text{Alg}}}{c_1^{\text{Alg}}} = c_2^{\text{Alg}} \tag{C.9}$$

Since $\|x^*\| \leq \text{R}$, the above inequality implies that $\pi_\gamma^{\text{Alg}} \in \mathcal{P}_2(\mathbb{R}^d)$. Applying Jensen's inequality, we derive the necessary bound to conclude the first part.

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 Meyn and Tweedie (2009))

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

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 $c' \geq 1$, $\mathcal{E}''(x) := c'\mathcal{E}'(x)$ 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 Meyn and Tweedie (2009)). 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

$$|P^k\phi(x_0) - \mathbb{E}_{x\sim\pi_\gamma^{\text{Alg}}}[\phi(x)]| \leq R_{\ell_\phi,\gamma}r_{\ell_\phi,\gamma}^k|\mathcal{E}''(x_0)|,$$

thus by setting $\zeta^{\text{Alg}}_{\phi,x_0,\gamma} := R_{\ell_\phi,\gamma}|\mathcal{E}''(x_0)|$ and $\tau^{\text{Alg}}_{\phi,\gamma} := r_{\ell_\phi,\gamma}$ 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

$$\sup_{|\phi|\le 1} |P^k\phi(x_0) - \mathbb{E}_{x\sim\pi^{\text{Alg}}_\gamma}[\phi(x)]| \le R_\gamma r^k_\gamma |\mathcal{E}''(x_0)|,$$

which implies the geometric convergence under total variation distance via the dual representation of Radon metric for bounded initial conditions (Villani, 2009).

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

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

where $D^{\{\text{SGDA,SEG}\}} \propto c_2^{\text{Alg}}$ by (C.8) and (C.9).

■

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

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

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*

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

and that

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

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

Proof. According to Theorem 17.0.1 in Meyn and Tweedie (2009), 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 Meyn and Tweedie (2009) and (ii) $|\phi(x)|^2 \le \mathcal{E}^*(x)$.

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

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*

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

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.3 (cf. Eq. (B.3) and Assumption 3) we have that

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

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

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

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

For the case of (SEG), it easy to see that $\mathrm{Gap}_V(x) \le \ell\|x_t - x^*\|^2$. 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 $V = (\nabla_\theta f, -\nabla_\phi f)$:

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

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

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

Additionally, it is easy to see that

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

and

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

By weak duality we have that $\max_{\phi'\in\mathbb{R}^{d_2}}\min_{\theta'\in\mathbb{R}^{d_1}} f(\theta',\phi') \le \min_{\theta'\in\mathbb{R}^{d_1}}\max_{\phi'\in\mathbb{R}^{d_2}} f(\theta',\phi')$, thus we have that

$$|f(\theta,\phi) - \min_{\theta'\in\mathbb{R}^{d_1}}\max_{\phi'\in\mathbb{R}^{d_2}} f(\theta',\phi')| \le \text{Duality-Gap}_f(\theta,\phi)$$

And the result follows.

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

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

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

$$\|x_{t+1} - x^*\|^4 \leq \|x_t - x^*\|^4(1 + 8\gamma^2\ell^2 + 8\gamma^4\ell^4 - 4\gamma\mu + 16\gamma^3\ell^3 + 4\gamma^2\ell^2)$$

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

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

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

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

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

Now consider $x \sim \pi_\gamma$ and let $\mathbb{E}_{x\sim\pi_\gamma}[\|x - x^*\|^k] = \text{Mom}(k)$. Notice that the first moment is also bounded by $\mathcal{O}(\sqrt{\gamma/\mu})$ since from **??** 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

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

equivalently

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

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

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

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

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

■

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 $\gamma < \min(\gamma^{\text{D.1}}_{\text{thresh}}, \gamma^{\text{C.1}}_{\text{thresh}}) := \theta'$ 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^*

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

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

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

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

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

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

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

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. ■

E Experiment Details of Section 6

We provide the details for the experiments presented in Section 6. We have adapted the code of the repository of Hsieh et al. (2020a).

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 $\sigma = 0.5$.

For the experiments on the RR refinement scheme (Fig. 5), we consider 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

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

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

F 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 Nemirovski (2004) study a variant of Extra Gradient algorithm (Korpelevich, 1976) and establishes optimal convergence rates for ergodic average, and the work Gidel et al. (2018); Mokhtari et al. (2020) studies proximal point algorithm with geometric convergence results.

Most related to us are works for the stochastic setting, for which SEG (Juditsky et al., 2011) and SGDA (Nemirovski et al., 2009) are two of the most prominent algorithms. Non-convergent phenomena are observed even in unconstrained bilinear games (Gidel et al., 2018; Mertikopoulos et al., 2019; Chavdarova et al., 2019; Daskalakis et al., 2018; Hsieh et al., 2020a). 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 Juditsky et al. (2011) provided the first analysis of SEG for monotone VIPs. Subsequent work has extended these results to other settings (Mishchenko et al., 2020; Kannan and Shanbhag, 2019; Mertikopoulos and Zhou, 2019; Hsieh et al., 2020a; Beznosikov et al., 2020; Gorbunov et al., 2022). A parallel line of work studies SGDA and its variants under different scenarios (Nemirovski et al., 2009; Loizou et al., 2021; Yang et al., 2020; Lin et al., 2020). Recently Beznosikov et al. (2023) 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 Gorbunov et al. (2022) for SEG and Beznosikov et al. (2023) 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 (Diakonikolas et al., 2021; Daskalakis et al., 2021; Papadimitriou et al., 2022). Similar conditions have been considered in prior work to establish the convergence guarantee of various algorithms (Hsieh et al., 2020a; Gorbunov et al., 2022; Yang et al., 2020; Song et al., 2020; Loizou et al., 2021).

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 Dieuleveut et al. (2018) studies SGD for smooth and strongly convex functions. Extensions to non-convex functions are considered in Yu et al. (2021), which establishes a central limit theorem that is similar in spirit to our results. Another extension is considered in the work in Can et al. (2022), which studies an accelerated version of constant stepsize SGDA for the unconstrained strongly convex strongly concave saddle point problem, which is a special case of the weakly quasi strongly monotone VIPs considered in our work; they do not consider SEG, constrained problems or biased stochastic oracles, and they do not establish an CLT. More recently, Bianchi et al. (2022) studies SGD for non-smooth non-convex functions. The work Durmus et al. (2021) considers constant step-size SA on Riemannian manifolds and studies the limiting behavior as the step-size approaches zero. The work Huo et al. (2023) considers linear SA with Markovian noise; see the references therein for other recent results on SA. We mention that Dieuleveut et al. (2018), Can et al. (2022) and Huo et al. (2023) examine a form of the Richardson-Romberg bias refinement scheme, which we also consider in this paper.

Analysis for Constant and Diminishing Step-sizes. While our consideration of constant stepsizes aligns well with the pragmatic choices of many practitioners, the theoretical literature (especially in bandits, online convex optimization/learning, and game dynamics) often considers a diminishing stepsize sequence γ_t. In particular, the analysis of many stochastic methods (such as SGD/AdaSGD/Stoch MirrorProx) typically involves an error term $\frac{\sum \gamma_t^2 \sigma_t^2}{\sum \gamma_t}$. By selecting γ_t to satisfy $\sum \gamma_t \to \infty, \sum \gamma_t^2 < \infty$, this term can be nullified. This general idea is well-established in the literature

With a constant stepsize, this error term remains, leading additional bias terms in the convergence analysis and final guarantee. While Richardson's bias reduction scheme has a simple implementation, its correctness relies on the machinery of establishing a limiting distribution and a precise characterization of the bias. A mere upper bound on the mean-squared error—the typical product of existing work on constant stepsize—conflates bias

with variance and is thus insufficient for this purpose. Overcoming the above challenges is a main goal of this paper.

Below we provide additional discussion on the two papers Dieuleveut et al. (2018) and Yu et al. (2021), which also consider the Markov chain perspective for constant-stepsize SGD.

F.1 Comparison with Dieuleveut et al. (2018)

Dieuleveut et al. (2018) considers SGD for unconstrained smooth and strongly convex optimization and views the iterates of SGD as a Markov chain. The approach taken by Dieuleveut et al. (2018) is based on coupling and convergence in *Wasserstein* distance. Our work, on the other hand, is based on irreducibility (implied by Assumption 5) and positive/Harris recurrence, which entail convergence in *total variation* distance. These two approaches to Markov chain analysis are complementary and have their own merits:

- While Dieuleveut et al. (2018) does not impose Assumption 5/irreducibility on the noise, it requires co-coercivity of the noisy gradient oracle. Their results hold for strongly convex and smooth minimization. Our work does not rely on co-coercivity of the noise and our results hold for nonsmooth problems and quasi-strongly monotone problems as well.
- In the proof of Proposition 2 in Dieuleveut et al. (2018), the Wasserstein distance is bounded for two arbitrary initializations. Specifically, it is necessary to bound the term $\langle \nabla f(\theta_1) - \nabla f(\theta_2), \theta_1 - \theta_2 \rangle$ - or in the notation we follow in this work the term $\langle V(x_1) - V(x_2), x_1 - x_2 \rangle$ - by a negative drift, $-\mu\|\theta_1 - \theta_2\|^2$ Quasi-monotonicity, as assumed in our paper, provides such a negative drift but, crucially, only relative to the optimum θ^* (i.e., for $\theta_2 = \theta^*$). One may try to add and subtract terms:

$$\langle f(\theta_1) - \nabla f(\theta^*), \theta_1 - \theta^* \rangle + \langle \nabla f(\theta_2) - \nabla f(\theta^*), \theta_2 - \theta^* \rangle \tag{F.1}$$
$$+ \langle \nabla f(\theta_1) - \nabla f(\theta^*), \theta^* - \theta_2 \rangle + \langle \nabla f(\theta_2) - \nabla f(\theta^*), \theta^* - \theta_1 \rangle. \tag{F.2}$$

  While first two terms can contribute to a desired drift, it is not straightforward to control the last two terms (referred to as the cross terms) under the assumption of quasi-monotonicity alone.

  This highlights the differences between quasi-strong-monotonicity and (exact-)strong-monotonicity: the latter is stronger and more restrictive, and would imply easier control over these cross terms that would appear in a coupling/Wasserstein-based analysis. This issue is critical even if one restricts attention to the smooth gradient case, thereby abandoning a unified approach for both non-smooth and smooth cases.
- Establishing a Central Limit Theorem (CLT) is a key product of our work. While it is possible to establish CLT under the Wasserstein distance framework, additional work is required. For example, in the absence of irreducibility or Assumption 5, further Lipschitz-type restriction must be placed on the test function ϕ for which one hopes to prove a CLT (e.g., see Douc et al. (2018)); our CLT only requires linear growth for ϕ. Further complications may arise when the problem is non-smooth. We note that Dieuleveut et al. (2018) does not give an explicit sufficient condition for a CLT.

In additional to the above differences, our work applies to the unconstrained case and the projected versions of SGDA/SEG, and allows for a potentially biased stochastic oracle.

F.2 Comparison with Yu et al. (2021)

Yu et al. (2021) considers SGD for unconstrained nonconvex optimization and views the iterates of SGD as a Markov chain. Similarly to us and different from Dieuleveut et al. (2018), their Markov analysis is based on irreducibility and convergence in total variational distance.

In addition to the more general VIPs setting that we consider, the main differences between Yu et al. (2021) and our work include: (i) we consider the more complicated SEG algorithm and expose its advantage in the smooth case; (ii) we provide refine analysis for the bias of SGDA; (iii) our work applies to the unconstrained case and the projected versions of SGDA/SEG, and allows for a potentially biased stochastic oracle. We elaborate below.

We begin by highlighting that one of our main contributions is a unified treatment of the smooth setting of SEG and the nonsmooth setting of SGDA. Moreover, while being unified, our analysis is strong enough to differentiate

performance of SEG and SGDA in the smooth case (see the in the next section). In comparison, Yu et al. (2021) focuses exclusively on unconstrained SGD in the nonsmooth case.

For SEG, our analysis for the miniorization condition (Lemma 1) is different from Yu et al. (2021), due to the more complicated form of the SEG update (involving an additional extrapolation step) and its interplay with the projection step, as compared to vanilla SGD. We also depart from Yu et al. (2021) by presenting results for minmax games for SEG.

For SGDA, we direct attention to our Section 5. In Section 5.1, we present bounds on the duality gap and game value of convex-concave minmax games. These results, as well as their analysis, are certainly absent in Yu et al. (2021), which only considers the minimization setting. In Section 5.2, we discuss Richardson-Romberg bias refinement. Establishing this result requires a more refined characterization of the bias of the limit distribution. In particular, for the bias vector we prove an expression that is an *equality* (up to higer order terms of γ; our analysis can be generalized to establish a more precise, higher order equality). Such an equality allows us to show that the Richardson-Romberg scheme can exactly cancel out the first-order term in the bias. In comparison, Yu et al. (2021) only provides *upper bounds* on the norm of the bias, which is insufficient for bias refinement.

G Additional Discussion on SEG vs. SGDA

In this section, we provide a comparison of SEG and SGDA in the context of our preliminary convergence result in Theorem 1. This theorem states that the mean squared error of the algorithm $Alg \in \{SGDA, SEG\}$ converges geometrically with a contraction factor $1 - c_1^{\mathrm{Alg}}$ up to a bias term c_2^{Alg}. It is preferable to have a larger c_1^{Alg} and a smaller c_2^{Alg}.

Assuming the stepsize choices in Theorem 1 and ignoring universal constants, our proofs give the following explicit expressions:

$$\begin{aligned} c_1^{\mathrm{SEG}} \asymp c_1^{\mathrm{SGDA}} &\asymp \gamma\mu, \\ c_2^{\mathrm{SEG}} &\asymp \gamma\frac{\sigma^2}{\mu} + \frac{\lambda\mu + b^2}{\mu^2}, \\ c_2^{\mathrm{SGDA}} &\asymp \gamma\frac{\sigma^2 + L^2(1+\mathrm{R}^2)}{\mu} + \frac{\lambda\mu + b^2}{\mu^2}. \end{aligned}$$

Here and in what follows, we write $a \asymp b$ to mean equality up to a universal multiplicative constant, i.e., $a = \Theta(b)$.

Below we provide a detailed discussion on these constants and compare them between SEG and SGDA, showing that the former has better constants in the smooth case. To make the comparison, we recall that γ is the stepsize, μ the strong monotonicity parameter, λ the weak monotonicity parameter, σ^2 the noise variance (we assume $\kappa = 0$ for simplicity), b the noise bias, and R the norm of x^*. For SEG, V is assumed to be ℓ-Lipschitz; for SGDA, V is assumed to have L-linear growth.

The contraction factor $1 - c_1$ (a larger c_1 is better). We assume V is ℓ-Lipschitz, which implies ℓ-linear growth, hence the assumptions for both SEG and SGDA are satisfied.

- For SEG, we have $c_1 \asymp \gamma\mu$. Under the stepsize choice $\gamma \asymp \frac{1}{\ell}$ (which is orderwise optimal under the assumption of Theorem 1), we have $c_1 \asymp \frac{\mu}{\ell}$, which is inversely proportional to the condition number $\frac{\ell}{\mu}$.
- For SGDA, we also have $c_1 \asymp \gamma\mu$. Under the stepsize choice $\gamma \asymp \frac{\mu}{\ell^2}$ (which is orderwise optimal under the assumption of Theorem 1), we have $c_1 \asymp \frac{\mu^2}{\ell^2}$, which is inversely proportional to the **square** of the condition number $\frac{\ell}{\mu}$. Therefore, SEG has a more favorable dependence on the condition number than SGDA.

The bias term c_2 (the smaller, the better).

- For SEG:
  1. For SEG, c_2 vanishes in the ideal case $\lambda = \sigma = b = 0$ (i.e., strongly monotone, unbiased and noiseless oracle).
  2. When $\lambda = b = 0$ (strongly monotone and unbiased oracle) but $\sigma > 0$, the bias c_2 diminishes to 0 as the stepsize γ diminishes to 0. Note that in this strongly monotone case, the solution x^* is unique.
  3. When $\lambda > 0$ (weakly strongly monotone), the bias c_2 has a persistent term $\frac{\lambda}{\mu}$, which does not scale with γ or σ. In this case, there may be multiple solutions x^*, and $\sqrt{\frac{\lambda}{\mu}}$ is precisely (an upper bound of) the radius of the solution set $\mathcal{X}^*$; see the Remark after Assumption 2. Therefore, our result implies convergence up to this radius, as should be expected.

  We emphasize that Points 2 and 3 above do *not* contradict with each other.
- For SGDA:
  1. For SGDA, c_2 is larger than the c_2 of SEG due to the additional term $\gamma\frac{L^2(1+\mathrm{R}^2)}{\mu}$. In particular, c_2 for SGDA remains nonzero even when $\lambda = \sigma = b = 0$.
  2. When $\lambda = b = 0$ (strongly monotone and unbiased oracle), the bias c_2 is proportional to the stepsize γ. If γ goes to 0, then the bias goes to zero.
  3. When $\lambda > 0$, we have convergence up to the radius of the solution set, similarly to SEG.

  Note that the above results for SGDA are consistent with known results for SGD for nonsmooth optimization.

We mention in passing that bounds for SEG often depend on certain variance parameters, whereas SGDA often depends on the second moment, a larger quantity. This phenomenon can be seen in the above discussion as well as in other regimes (e.g., for SEG/SGDA with diminishing stepsize). It reflects the fact that SEG can better leverage gradient smoothness, thanks to its extrapolation step.