SDEs for Minimax Optimization

Enea Monzio Compagnoni
Department of Mathematics
and Computer Science
University of Basel

Antonio Orvieto
ELLIS Institute Tübingen
MPI for Intelligent Systems
Tübingen AI Center

Hans Kersting
Yahoo! Research

Frank Norbert Proske
Department of Mathematics
University of Oslo

Aurelien Lucchi
Department of Mathematics
and Computer Science
University of Basel

Abstract

Minimax optimization problems have attracted a lot of attention over the past few years, with applications ranging from economics to machine learning. While advanced optimization methods exist for such problems, characterizing their dynamics in stochastic scenarios remains notably challenging. In this paper, we pioneer the use of stochastic differential equations (SDEs) to analyze and compare Minimax optimizers. Our SDE models for Stochastic Gradient Descent-Ascent, Stochastic Extragradient, and Stochastic Hamiltonian Gradient Descent are provable approximations of their algorithmic counterparts, clearly showcasing the interplay between hyperparameters, implicit regularization, and implicit curvature-induced noise. This perspective also allows for a unified and simplified analysis strategy based on the principles of Itô calculus. Finally, our approach facilitates the derivation of convergence conditions and closed-form solutions for the dynamics in simplified settings, unveiling further insights into the behavior of different optimizers.

1 INTRODUCTION

Minimax optimization plays a fundamental role in decision theory, game theory, and machine learning (Goodfellow et al., 2016). The problem it addresses is finding the solution of the following optimization problem:

$$\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\left[f(x,y):=\frac{1}{N}\sum_{i=1}^{N}f_i(x,y)\right], \qquad (1)$$

where $f, f_i : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ for $i = 1,\dots,N$. In machine learning, f is an empirical risk function where f_i is the contribution of the i-th data point of the training data. In this notation, $(x, y) \in \mathcal{X}\times\mathcal{Y}$ is a vector of trainable parameters and N is the size of the dataset. The goal is to find optimal saddle points $(\mathrm{x}^*, \mathrm{y}^*)$ such that

$$f\left(\mathrm{x}^*,\mathrm{y}\right)\le f\left(\mathrm{x}^*,\mathrm{y}^*\right)\le f\left(\mathrm{x},\mathrm{y}^*\right)\quad\forall\mathrm{x}\in\mathcal{X},\quad\forall\mathrm{y}\in\mathcal{Y}.$$

The most intuitive algorithm to solve Eq. (1) is Gradient Descent Ascent (GDA). However, its updates are computationally expensive for large datasets. Therefore, a common choice is to use *mini-batches* to approximate the gradients, which gives rise to Stochastic Gradient Descent Ascent (SGDA). Unfortunately, it is known that both GDA and SGDA do not converge on relatively simple landscapes such as $f(x, y) = xy$ for $(x, y) \in \mathbb{R}$. This led to the design of alternative optimizers such as Extragradient (EG) (Korpelevich, 1976) and Hamiltonian GD (Balduzzi et al., 2018). While these methods exhibit more favorable convergence guarantees compared to SGDA, they are relatively complex to study and some of their properties are still not well understood, especially in a stochastic setting.

In this paper, we leverage continuous-time models in the form of stochastic differential equations (SDEs) to study these minimax optimizers. SDEs have recently become popular in the *minimization* community: They provide a unified and simplified analysis strategy rooted in Itô calculus which facilitates the derivation of novel insights about the discrete algorithms, see e.g. (Su et al.,

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

2014; Li et al., 2017). It is worth mentioning that the interest in applying SDEs to minimax problems has been a topic of prior research discussions (Chavdarova et al., 2022). Following the framework of Li et al. (2017) for minimization, our work provides the first *formal* derivation — rooted in the theory of weak approximation (Mil'shtein, 1986) — of the SDEs of SGDA,

$$z_{k+1} = z_k - \eta F_{\gamma_k}(z_k), \tag{2}$$

SEG,

$$z_{k+1} = z_k - \eta F_{\gamma_k^1}(z_k - \rho F_{\gamma_k^2}(z_k)), \tag{3}$$

and SHGD

$$z_{k+1} = z_k - \eta \nabla \mathcal{H}_{\gamma_k^1, \gamma_k^2}(z_k), \tag{4}$$

where F is the drift field and $\mathcal{H}$ the Hamiltonian:

$$F_\gamma(z) = F_\gamma(x, y) := (\nabla_x f_\gamma(x, y), -\nabla_y f_\gamma(x, y)), \tag{5}$$

$$\mathcal{H}_{\gamma^1, \gamma^2}(z) := \frac{F_{\gamma^1}^\top(z) F_{\gamma^2}(z)}{2}. \tag{6}$$

Above, $\eta \in \mathbb{R}^{>0}$ is the stepsize and $\rho \in \mathbb{R}$ is the extra stepsize of SEG[1]. The mini-batches $\{\gamma_k^j\}$ are modelled as i.i.d. random variables uniformly distributed on $\{1, \cdots, N\}$, and of size $B \geq 1$.

Formally, these continuous-time models are weak approximations, i.e. approximations in distribution, of their respective discrete-time algorithms. We will exploit these models to derive novel insights into the convergence behavior, the effect of the noise and the curvature of the landscape, or the role of hyper-parameters such as the extra stepsize ρ appearing in SEG.

Contributions.

- We provide the *first formal* derivation of the SDE models of popular minimax optimizers. Then, we use them to make the following additional contributions:
  1. **Moderate Exploration regime.** If $\rho = \mathcal{O}(\eta)$, we show that SEG essentially behaves like SGDA;
  2. **Aggressive Exploration regime** (Hsieh et al., 2020). For $\rho = \mathcal{O}(\sqrt{\eta})$, the dynamics of SEG can be interpreted as that of SGDA on an **implicitly** regularized vector field with additional *implicit curvature-induced* noise;
  3. SHGD uses **explicit** curvature-based information. Thus, it has an *explicit curvature-induced* noise;
  4. We characterize the evolution of the Hamiltonian under the dynamics of SEG and SHGD;
  5. We use the latter to derive convergence conditions for SEG and SHGD on a wide class of functions.
- For Bilinear Games with different noise structures:
  1. We explicitly solve the differential equation of the Hamiltonian, thus elucidating the interplay of all hyperparameters in determining the speed of convergence (or divergence) of these methods;
  2. We provide necessary and sufficient conditions for stepsize schedulers to recover convergence.
- We explicitly solve the SDEs for some *Quadratic* Games, meaning that we derive the *first* closed-form formula for the dynamics of SEG and SHGD on these landscapes. This allows for a 1-to-1 comparison of the two optimizers, particularly of their first and second moments. One key takeaway of this comparison is that selecting ρ is a matter of trade-off between the speed of convergence and asymptotic optimality: Our formulas show how a suitable choice of ρ allows SEG to match (or outperform) SHGD w.r.t. convergence speed but negatively impacts its optimality (the iterates converge to a larger neighborhood of the optimum). Interestingly, the curvature determines whether SEG or SHGD is faster at converging as well as more suboptimal. Importantly, we provide the first experimental and theoretical evidence that **negative** ρ might be advisable for certain landscapes.
- Finally, we present extensive experiments on various relevant minimax problems: these are meant to verify that each formula derived from our SDEs correctly describes the behavior of the respective discrete-time algorithms. Figure 1 offers a preliminary glimpse at the accuracy of the SDEs approximations.

[1]We also support the cases where the stepsizes and extra steps depend on time, e.g. η_k and ρ_k, as well as depend on the coordinates, e.g. $\eta = (\eta_1, \cdots, \eta_d)$.

2 RELATED WORKS

We start by discussing existing continuous-time analysis for minimax optimization and related applications. For related works regarding SGDA, SEG, SHGD, and bilinear games, we refer the reader to Appendix A.

ODE Approximations and Applications. Several works use *continuous-time models* to describe the dynamics of minimax optimizers. First, Ryu et al. (2019) informally derived ODEs to study Stochastic Gradient Methods with Optimism and Anchoring. Then, Lu (2022) formally showed that different saddle-point optimizers yield the same ODE and derived High-Resolution (Ordinary) Differential Equations (HRDEs) to provide convergence conditions on a wide class of problems. Similarly, Chavdarova et al. (2023) derived HRDEs as well and established the convergence of certain methods in continuous time on bilinear games. Finally, Hsieh et al. (2021) modeled a wide class of zeroth- and first-order minimax algorithms with ODEs and proved that they may be subject to inescapable convergence failures, meaning that they could get attracted by spurious attractors. Unfortunately, their approach based on Robbins–Monro templates cannot

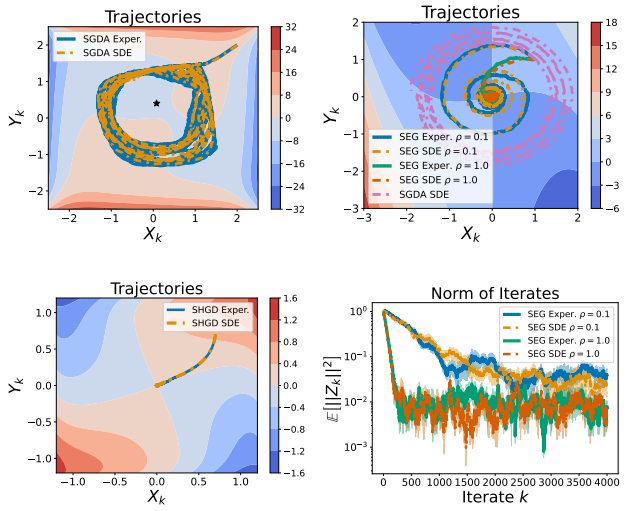

Figure 1: Empirical validation of Theorem 3.4 and 3.6: The trajectories of the simulated SDEs match those of the respective algorithms averaged over 5 runs - That of SGDA gets trapped in limit cycles as well (Top Left); That of SHGD converges to the optimum of a highly nonlinear landscape (Bottom Left); The SDE of SGDA would not be a good model for SEG (Top Right); The SDEs and the optimizers move along the trajectory at the same speed (Bottom Right). For a description of the landscapes and of the simulation settings for the SDEs, see Appendix G.

handle ergodic averages, second-order methods, adaptive methods, and constant stepsizes.

SDE Approximations and Applications. (Li et al., 2017) first proposed a formal theoretical framework to derive SDEs to appropriately capture the intrinsic stochasticity of stochastic optimizers. These SDEs can be understood as weak approximations of stochastic gradient algorithms (See Definition 3.2). SDEs open to a variety of concrete applications that include *stochastic optimal control* to select the stepsize (Li et al., 2017, 2019) or the batch size (Zhao et al., 2022) and *scaling rules* (Malladi et al., 2022) to adjust the optimization hyperparameters w.r.t. the batch size. Additionally, SDEs give access to the fine-grained structure of the interaction between stochasticity and curvature. For example, the study of *escape times* of SGD from minima of different sharpness (Xie et al., 2021), the factors influencing the minima found by SGD Jastrzebski et al. (2018), the convergence bounds for mini-batch SGD and SVRG derived in Orvieto and Lucchi (2019), and the fundamental interplay between noise and curvature of the landscape for SAM (Compagnoni et al., 2023). For more references, see (Kushner and Yin, 2003; Ljung et al., 2012; Chen et al., 2015; Mandt et al., 2015; Chaudhari and Soatto, 2018; Zhu et al., 2019; He et al., 2018; An et al., 2020). A gentle introduction to SDEs is provided in Appendix B.2

3 RESULTS & INSIGHTS: THE SDEs

This section provides the general formulations of the SDEs of SGDA (Theorem 3.3), SEG (Theorem 3.4), and SHGD (Theorem 3.6). Due to the technical nature of the analysis, we refer the reader to the appendix for the complete formal statements and proofs.

Assumption 3.1. We assume that

1. $\nabla f, \nabla f_i$ satisfy a Lipschitz condition: $\exists L > 0$ s.t. $|\nabla f(u) - \nabla f(v)| + \sum_{i=1}^{N} |\nabla f_i(u) - \nabla f_i(v)| \leq L|u - v|$;
2. f, f_i and their partial derivatives up to order 7 have polynomial growth;
3. $\nabla f, \nabla f_i$ satisfy a linear growth condition: $\exists M > 0$ s.t. $|\nabla f(z)| + \sum_{i=1}^{N} |\nabla f_i(z)| \leq M(1 + |z|)$.

Definition 3.2 (Weak Approximation)**.** A continuous-time stochastic process $\{Z_t\}_{t\in[0,T]}$ is an order α weak approximation (or α-order SDE) of a discrete stochastic process $\{z_k\}_{k=0}^{\lfloor T/\eta \rfloor}$ if for every polynomial growth function g, there exists a positive constant C, independent of the stepsize η, such that $\max_{k=0,\dots,\lfloor T/\eta \rfloor} |\mathbb{E}g(z_k) - \mathbb{E}g(Z_{k\eta})| \leq C\eta^{\alpha}$.

This definition comes from the field of numerical analysis of SDEs, see Mil'shtein (1986). When $g(z) = \|z\|^j$, the bound restricts the disparity between the j-th moments of the discrete and the continuous process.

3.1 SGDA SDE

Theorem 3.3 (SGDA SDE - Informal Statement of Theorem C.5)**.** *Under sufficient regularity conditions, the solution of the following SDE is an order* 1 *weak approximation of the discrete update of SGDA* (2)*:*

$$dZ_t = -F(Z_t)\,dt + \sqrt{\eta \Sigma(Z_t)}dW_t, \tag{7}$$

where $\Sigma(z)$ *is the noise covariance*

$$\Sigma(z) = \mathbb{E}[\xi_\gamma(z)\xi_\gamma(z)^\top], \tag{8}$$

and $\xi_\gamma(z) := F(z) - F_\gamma(z)$ *the noise in the sample* F_γ*.*

3.2 SEG SDE

A notable characteristic of SEG is the inclusion of the variable ρ that controls the magnitude of the extra step. This variable plays an important role in the derivation of the SDE of SEG and one has to differentiate between two different regimes:

1. When $\rho \sim \eta$, the SDE of SEG is the same as SGDA, which is consistent with the literature on ODEs (Chavdarova et al., 2023; Lu, 2022). The formal proof is given in Theorem C.12.

2. However, if the extra stepsize ρ is sizeably larger than η (Hsieh et al., 2020) (i.e. $\rho = \mathcal{O}(\sqrt{\eta})$), SEG enters a more exploratory regime, for which the SDE becomes distinct from the first regime.

Before presenting our main result, we introduce some notation. Let $\boldsymbol{\gamma} := (\gamma^1, \gamma^2)$, $\bar{F}_{\boldsymbol{\gamma}}(z) := \nabla F_{\gamma^1}(z) F_{\gamma^2}(z)$, and $\bar{F}(z) := \mathbb{E}[\bar{F}_{\boldsymbol{\gamma}}(z)]$ be its expectation. We denote the noise in $\bar{F}$ as $\bar{\xi}_{\boldsymbol{\gamma}}(z) := \bar{F}_{\boldsymbol{\gamma}}(z) - \bar{F}(z)$ and consider the mixed (non-symmetric) covariance matrix $\bar{\Sigma}(z) = \mathbb{E}[\xi_{\gamma^1}(z)\bar{\xi}_{\boldsymbol{\gamma}}(z)^\top]$.

Theorem 3.4 (Informal Statement of Theorem C.8)**.** *Let*

$$F^{SEG}(z) := F(z) - \rho \bar{F}(z), \tag{9}$$

$$\Sigma^{SEG}(z) := \Sigma(z) + \rho \left[\bar{\Sigma}(z) + \bar{\Sigma}(z)^\top\right]. \tag{10}$$

Under sufficient regularity conditions and $\rho = \mathcal{O}(\sqrt{\eta})$, the solution of the following SDE is the order 1 weak approximation of the discrete update of SEG

$$dZ_t = -F^{SEG}(Z_t)\, dt + \sqrt{\eta \Sigma^{SEG}(Z_t)} dW_t. \tag{11}$$

Proof. To prove that the SDE is a weak approximation of SEG as per Definition 3.2, we prove that the first and second moments of its discretization match those of SEG up to an error of order η and η^2, respectively. □

For didactic reasons, we now present Corollary 3.5, a consequence of Theorem 3.4 that provides a more interpretable SDE for SEG which we will use to establish a comparison with SGDA (Eq.(7)) and SHGD (Eq.(16)).

Corollary 3.5 (Informal Statement of Corollary C.10)**.** *Under the assumptions of Theorem 3.4, that $\gamma^1 = \gamma^2 = \gamma$, and that the stochastic gradients are $\nabla_x f_\gamma(z) = \nabla_x f(z) + U^x$ and $\nabla_y f_\gamma(z) = \nabla_y f(z) + U^y$ such that U^x and U^y are independent noises that do not depend on z, the following SDE provides a 1 weak approximation of the discrete update of SEG*

$$\begin{aligned} dZ_t = &- \left(F(Z_t) - \rho \nabla F(Z_t) F(Z_t)\right) dt \\ &+ \left(\mathbf{I}_{2d} - \rho \nabla F(Z_t)\right) \sqrt{\eta \Sigma} dW_t. \end{aligned} \tag{12}$$

If instead (under the same assumptions) γ^1 and γ^2 are uncorrelated, the SDE has a drift regularization term but no variance regularization:

$$dZ_t = -\left(F(Z_t) - \rho \nabla F(Z_t) F(Z_t)\right) dt + \sqrt{\eta \Sigma} dW_t.$$

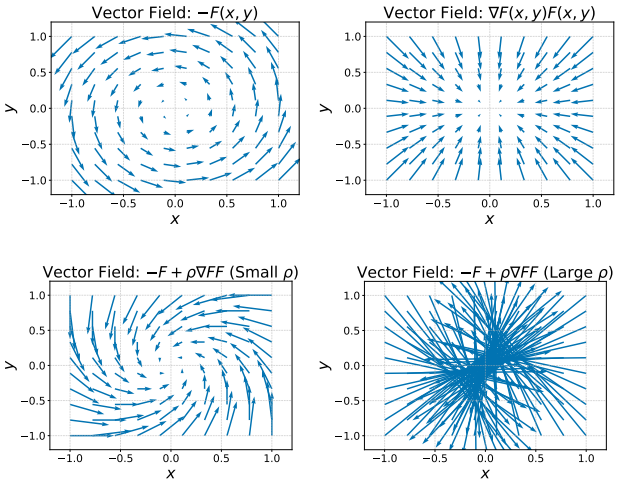

Figure 2: Graphical representation of the *implicit regularization* of the vector field of SEG for $f(x, y) = xy$: $-F$ spins the dynamics in a circle (Top Left); $+\nabla F F$ pulls it towards 0 (Top Right); If ρ is small, $-F + \rho \nabla F F$ combines the two fields and spirals towards the origin (Bottom Left); If ρ is large, $-F + \rho \nabla F F$ is a chaotic field that makes the dynamics diverge (Bottom Right).

Proof. The noise assumption implies $\nabla F_{\gamma^1}(z) = \nabla F(z)$. Therefore $\bar{F}(z) := \nabla F(z) F(z)$. Next, note that $\bar{\xi}_{\boldsymbol{\gamma}}(z) := \nabla F(z) \xi_{\gamma^2}(z)$ and therefore $\bar{\Sigma}(z) = \mathbb{E}[\xi_{\gamma^1}(z) \xi_{\gamma^2}^\top(z) \nabla F(z)^\top] = \Sigma \nabla F(z)^\top$ if $\gamma^1 = \gamma^2$ and is zero otherwise. Next, note that $\Sigma^{\text{SEG}}(z) := \Sigma + \rho \Sigma \nabla F(z)^\top + \rho \nabla F(z) \Sigma = (\mathbf{I}_{2d} + \rho \nabla F(z)) \Sigma (\mathbf{I}_{2d} + \rho \nabla F(z)^\top) + \mathcal{O}(\rho^2)$. Since terms of order ρ^2 have a vanishing influence, this proves the result. □

3.3 SHGD SDE

Theorem 3.6 (SHGD SDE - Informal Statement of Theorem C.14)**.** *Let $\mathcal{H}_{\boldsymbol{\gamma}} := \mathcal{H}_{\gamma^1, \gamma^2}$ and let us define*

$$F^{SHGD}(z) := \nabla \mathbb{E}\left[\mathcal{H}_{\boldsymbol{\gamma}}(z)\right], \tag{13}$$

$$\Sigma^{SHGD}(z) := \mathbb{E}\left[\hat{\xi}_{\boldsymbol{\gamma}}(z) \hat{\xi}_{\boldsymbol{\gamma}}(z)^\top\right], \tag{14}$$

where $\hat{\xi}_{\boldsymbol{\gamma}}(z) = F^{SHGD}(z) - \nabla \mathcal{H}_{\boldsymbol{\gamma}}(z)$. Under sufficient regularity conditions, the solution of the following SDE is the order 1 weak approximation of the discrete update of SHGD (4):

$$dZ_t = -F^{SHGD}(Z_t)\, dt + \sqrt{\eta \Sigma^{SHGD}(Z_t)} dW_t. \tag{15}$$

Once again, we provide a more interpretable SDE under additional assumptions:

Corollary 3.7 (Informal Statement of Corollary C.16)**.** *Under the assumptions of Theorem 3.6, that $\gamma^1 = \gamma^2 = \gamma$, and that the stochastic gradients are $\nabla_x f_\gamma(z) = \nabla_x f(z) + U^x$ and $\nabla_y f_\gamma(z) = \nabla_y f(z) + U^y$ such that*

U^x and U^y are independent noises that do not depend on z, the SDE is

$$dZ_t = -\nabla \mathcal{H}\left(Z_t\right) dt + \sqrt{\eta}\nabla^2 f\left(Z_t\right)\sqrt{\Sigma} dW_t. \quad (16)$$

If instead γ^1 and γ^2 are uncorrelated, the SDE is the same but with less variance:

$$dZ_t = -\nabla \mathcal{H}\left(Z_t\right) dt + \sqrt{\frac{\eta}{2}}\nabla^2 f\left(Z_t\right)\sqrt{\Sigma} dW_t. \quad (17)$$

Empirical Validation Figure 1 shows the empirical validation of Theorem 3.4 and Theorem 3.6: The top left shows that the SDE of SGDA matches the algorithm and is also attracted to a limit cycle. The bottom left shows that the SDE of SHGD matches the empirical optimization of a highly nonlinear landscape. The top right shows that the SDE of SEG matches its discrete-time counterpart for different values of ρ. Also, the SDE of SGDA is not a good model to describe the dynamics of SEG. The bottom right shows the evolution of the norm of the iterates in time: We understand that the SDE $Z_{k\eta}$ and optimizer z_k move at the exact same speed along the trajectory — This justifies their use as investigation tools. Figure 9 shows that if $\rho = \mathcal{O}(\eta)$ or smaller, the SDE of SGDA models the dynamics of SEG accurately. However, if $\rho = \mathcal{O}(\sqrt{\eta})$ or larger, the SDE of SGDA no longer does so while the SDE of SEG does. All experiments are averaged over 5 runs and additional details are in Appendix G.

3.4 Comparisons

There are three notable observations we immediately derive from the SDEs presented above:

1. Let us use $\tilde{\nabla} := (\nabla_x, -\nabla_y)$. Then, one can see that the drift term F is simply equal to $F = \tilde{\nabla} f$ for SGDA, while SEG *implicitly* introduces an additional regularizer such that

   $$F^{\text{SEG}} = \tilde{\nabla}\left[f + \frac{\rho}{2}\left[\|\nabla_y f\|_2^2 - \|\nabla_x f\|_2^2\right]\right].$$

   Therefore, the dynamics of SEG is equivalent to that of SGDA on an **implicitly** regularized vector field. Figure 2 illustrates this phenomenon.

2. The presence of $\rho\nabla F$ in the diffusion term of SEG shows that the extra step **implicitly** adds (on top of that of SGDA) a noise component that depends on the curvature of the landscape.

3. SHGD is a second-order method that **explicitly** optimizes the Hamiltonian which by definition uses curvature-based information. The SDE in Eq.(16) shows how this results in $\nabla^2 f$ directly affecting its noise structure.

4 CONVERGENCE CONDITIONS

In this section, we derive the ODE that characterizes the evolution of the expected Hamiltonian H_t along the dynamics of SEG and SHGD. We use it to derive convergence conditions on a wide class of functions. Then, we focus on Bilinear Games where we can explicitly solve the ODE which allows us to single out the role of each ingredient of the dynamics. Finally, we provide sufficient conditions to craft stepsize schedulers that induce convergence.

4.1 SHGD

We begin by introducing an auxiliary result that elucidates the evolution of the Hamiltonian. Here we denote by Z_t the stochastic process that defines the evolution of SHGD. We also define $H_t := \mathbb{E}_{\boldsymbol{\gamma}}[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)]$ and $\Sigma_t^{\text{SHGD}} := \Sigma^{\text{SHGD}}(Z_t)$. Then,

$$\mathbb{E}\left[\dot{H}_t\right] = -\mathbb{E}\left[\|\nabla H_t\|^2\right] + \frac{\eta}{2}\text{Tr}\left(\mathbb{E}\left[\Sigma_t^{\text{SHGD}}\nabla^2 H_t\right]\right).$$

We observe that:

1. $-\mathbb{E}\left[\|\nabla H_t\|^2\right]$ comes from the drift of the SDE and is pulling the dynamics towards regions with zero energy;
2. $\frac{\eta}{2}\text{Tr}\left(\mathbb{E}\left[\Sigma_t^{\text{SHGD}}\nabla^2 H_t\right]\right)$ is induced by the diffusion term and has an adversarial effect;
3. Convergence can only be achieved if the pulling force is stronger than the repulsive one, even at vanishing energies.

We formalize this in Theorem 4.1 and Corollary 4.2.

Theorem 4.1 (SHGD General Convergence)**.** *Consider the solution Z_t of the SHGD SDE with $\gamma^1 \neq \gamma^2$. Let $v_t := \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}\left[\|\nabla \mathcal{H}_t\left(Z_t\right) - \nabla \mathcal{H}_{\boldsymbol{\gamma}}\left(Z_t\right)\|_2^2\right]\right]$ measure the error in $\nabla\mathcal{H}$, in expectation over the whole randomness up to time t. Suppose that:*

1. *The smallest eigenvalue (in absolute value) μ of $\nabla^2 f(z)$ is non-zero;*
2. *$\|\nabla^2 H(z)\|_{op} < \mathcal{L}_{\mathcal{T}}$, for all $z \in \mathbb{R}^{2d}$.*

Then,

$$\mathbb{E}\left[H_t\right] \leq e^{-2\mu^2 t}\left[H_0 + \frac{\eta\mathcal{L}_{\mathcal{T}}}{2}\int_0^t v_s e^{2\mu^2 s} ds\right]. \quad (18)$$

Proof. We derive the SDE of H_t via Itô's Lemma and take its expectation to obtain the ODE of $\mathbb{E}\left[H_t\right]$. Then, we use the assumptions to derive a bound on it. □

Corollary 4.2. *Under the assumptions of Theorem 4.1, if for* $\mathcal{L}_{\mathcal{V}} > 0$

$$v_t \leq \mathcal{L}_{\mathcal{V}} \mathbb{E}\left[H_t\right], \tag{19}$$

the solution is more explicit:

$$\mathbb{E}\left[H_t\right] \leq H_0 e^{\left(-2\mu^2 + \eta \mathcal{L}_{\mathcal{V}} \mathcal{L}_{\mathcal{T}}\right)t}. \tag{20}$$

If instead

$$v_t \leq \mathcal{L}_{\mathcal{V}}, \tag{21}$$

we have

$$\mathbb{E}\left[H_t\right] \leq H_0 e^{-2\mu^2 t} + \left(1 - e^{-2\mu^2 t}\right) \frac{\eta \mathcal{L}_{\mathcal{V}} \mathcal{L}_{\mathcal{T}}}{2\mu^2}. \tag{22}$$

Discussion about Assumptions Note that:

1. (Loizou et al., 2020) which first proposed SHGD assumed independent mini-batches;
2. Lipschitzianity on $\nabla \mathcal{H}_{\gamma^1, \gamma^2}$ and Error Bound on F imply Eq. (19);
3. Bounded variance on $\nabla \mathcal{H}_{\gamma^1, \gamma^2}$ implies Eq. (21).

Concrete Examples We analyze Bilinear Games for which we can provide explicit formulas for the results presented above. We focus on: $f(x, y) = x^\top \mathbb{E}_\xi\left[\mathbf{\Lambda}_\xi\right] y$ and $f(x, y) = x^\top \mathbf{\Lambda} y$ where $\mathbf{\Lambda}$ and $\mathbf{\Lambda}_\xi$ are square, diagonal, and positive semidefinite matrices.

Proposition 4.3. *For* $f(x, y) = x^\top \mathbb{E}_\xi\left[\mathbf{\Lambda}_\xi\right] y$*, Eq.* (19) *holds and we have*

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} = \sum_{i=1}^{d} \frac{\|Z_0^i\|^2}{2} e^{-\left(2\lambda_i^2 - \eta\sigma_i^2\left(2\lambda_i^2 + \sigma_i^2\right)\right)t}. \tag{23}$$

In particular, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{=} 0$ *if* $\eta < \frac{2\lambda_i^2}{\sigma_i^2\left(2\lambda_1^2 + \sigma_i^2\right)}, \quad \forall i.$

In this case, the noise structure is such that v_t scales like $\mathbb{E}\left[H_t\right]$. Thus, $\mathbb{E}\left[H_t\right]$ exponentially decays to 0.

Proposition 4.4. *For* $f(x, y) = x^\top \mathbf{\Lambda} y$ *and covariance noise* $\Sigma := \operatorname{diag}(\sigma_1, \cdots, \sigma_d)$*, Eq. 21 holds and*

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} = \sum_{i=1}^{d} \frac{\|Z_0^i\|^2}{2} e^{-2\lambda_i^2 t} + \frac{\eta\sigma_i^2}{2}\left(1 - e^{-2\lambda_i^2 t}\right), \tag{24}$$

which implies that $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{=} \frac{\eta}{2}\sum_{i=1}^{d}\sigma_i^2 > 0.$

In this case, v_t is bounded, meaning that $\mathbb{E}\left[H_t\right]$ reaches an asymptotic suboptimality exponentially fast.

Now we provide sufficient and necessary conditions to craft stepsize schedulers that recover convergence.

Proposition 4.5. *Under the assumptions of Prop. 4.4, for any stepsize scheduler* η_t*,* $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2}$ *is equal to*

$$\sum_{i=1}^{d} e^{-2\lambda_i^2 \int_0^t \eta_s ds}\left(\frac{\|Z_0^i\|^2}{2} + \eta\sigma_i^2\lambda_i^2 \int_0^t e^{2\lambda_i^2 \int_0^s \eta_r dr}\eta_s^2 ds\right).$$

Therefore,

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{\to} 0 \iff \int_0^\infty \eta_s ds = \infty \text{ and } \lim_{t\to\infty}\eta_t = 0. \tag{25}$$

Among other possible choices of η_t*,*

$$\eta_t = \frac{1}{(t+1)^\gamma} \implies \frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \to 0, \text{ for } \gamma \in \{0.5, 1\}.$$

4.2 SEG

Let Z_t be the solution of the SEG SDE, $\Sigma_t^{\text{SEG}} = \Sigma^{\text{SEG}}(Z_t)$, $H_t = \mathbb{E}_{\boldsymbol{\gamma}}[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)]$, and $F_t^{\text{SEG}} = F^{\text{SEG}}(Z_t)$. Then,

$$\mathbb{E}\left[\dot{H}_t\right] = -\mathbb{E}\left[\nabla H_t^\top F_t^{\text{SEG}}\right] + \frac{\eta}{2}\operatorname{Tr}\left(\mathbb{E}\left[\Sigma_t^{\text{SEG}}\nabla^2 H_t\right]\right).$$

Once again, we have to study how the pulling and repulsive forces balance each other in order to dissect the convergence behavior of SEG.

Theorem 4.6 (SEG General Convergence)**.** *Consider the solution* Z_t *of the SEG SDE with* $\gamma^1 \neq \gamma^2$*. Let* $v_t := \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}\left[\|F^{SEG}(Z_t) - F_{\boldsymbol{\gamma}}^{SEG}(Z_t)\|_2^2\right]\right]$ *measure the error in* F^{SEG}*, in expectation over the whole randomness up to time* t*. Suppose that:*

1. *The smallest eigenvalue (in absolute value)* μ_ρ *of* $\mathbf{M}$ *is non-zero, where* $\mathbf{M} = \operatorname{diag}(\mathbf{M}_{1,1}, \mathbf{M}_{2,2})$*, with* $\mathbf{M}_{1,1} := \nabla^2 f_{xx} + \rho\left(\nabla^2 f_{xy}\nabla^2 f_{xy}^T - \nabla^2 f_{xx}^2\right)$*, and* $\mathbf{M}_{2,2} := -\nabla^2 f_{yy} + \rho\left(\nabla^2 f_{xy}\nabla^2 f_{xy}^T - \nabla^2 f_{yy}^2\right)$*;*
2. $\|\nabla^2 H(z)\|_{op} < \mathcal{L}_{\mathcal{T}}$*, for all* $z \in \mathbb{R}^{2d}$*.*

Then,

$$\mathbb{E}\left[H_t\right] \leq e^{-2\mu_\rho^2 t}\left[H_0 + \frac{\eta\mathcal{L}_{\mathcal{T}}}{2}\int_0^t v_s e^{2\mu_\rho^2 s} ds\right]. \tag{26}$$

Corollary 4.7. *The very same result as Corollary 4.2 holds where we substitute* μ *with* μ_ρ*.*

Discussion about Assumptions Note that:

1. Independent mini-batches are used in Indep.-Sample SEG (Du et al., 2022; Gorbunov et al., 2022), and are not a necessary condition;
2. κ_1-Lipschitzianity on F_{γ^1}, κ_2-Lipschitzianity on $\nabla F_{\gamma^1} F_{\gamma^2}$, and β-Error Bound on F, imply $v_t \leq \beta^2(\kappa_1^2 + \rho^2\kappa_2^2)\mathbb{E}\left[H_t\right]$;
3. σ_1-Bounded variance on F_{γ^1} and σ_2-Bounded variance on $\nabla F_{\gamma^1} F_{\gamma^2}$, imply $v_t \leq \sigma_1^2 + \rho^2\sigma_2^2$.

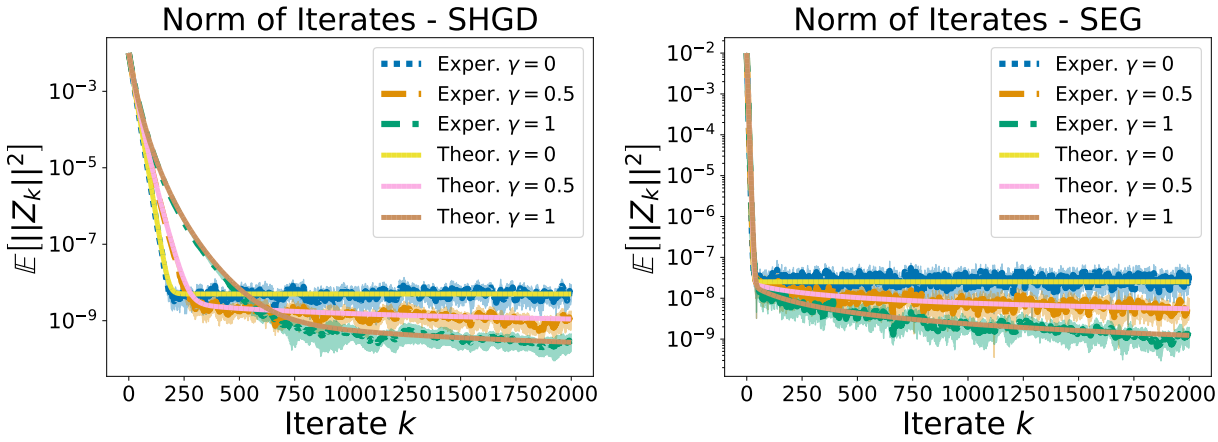

Figure 3: Empirical validation of Prop. 4.4 and Prop. 4.5 (Left); Prop. 4.9 and Prop. 4.10 (Right): The dynamics of $\mathbb{E}\left[\|Z_t\|^2\right]$ averaged across 5 runs perfectly matches that prescribed by our results for all schedulers. Both for SEG and SHGD, $\eta = 0.01$, while $\rho = 1$.

Interestingly, we notice that increasing ρ might be detrimental as it could possibly lead to divergence.

Concrete Examples

Proposition 4.8. *For $f(x,y) = x^\top \mathbb{E}_\xi\left[\mathbf{\Lambda}_\xi\right] y$, we have*

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} = \sum_{i=1}^{d} \frac{\|Z_0^i\|^2}{2} e^{-\left(2\rho\lambda_i^2 - \eta\sigma_i^2\left(1+\rho^2\left(2\lambda_i^2+\sigma_i^2\right)\right)\right)t}. \tag{27}$$

In particular, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{=} 0$ if $, \forall i \in \{i, \cdots, d\}$

$$2\rho\lambda_i^2 - \eta\sigma_i^2\left(1+\rho^2\left(2\lambda_i^2+\sigma_i^2\right)\right) > 0. \tag{28}$$

Proposition 4.9. *For $f(x,y) = x^\top \mathbf{\Lambda} y$ and covariance noise $\Sigma := \mathrm{diag}(\sigma_1, \cdots, \sigma_d)$, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2}$ is equal to*

$$\sum_{i=1}^{d} \frac{\|Z_0^i\|^2}{2} e^{-2\rho\lambda_i^2 t} + \frac{\eta\sigma_i^2}{2}\frac{1+\rho^2\lambda_i^2}{\rho\lambda_i^2}\left(1 - e^{-2\rho\lambda_i^2 t}\right), \tag{29}$$

which implies that

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{=} \frac{\eta}{2}\sum_{i=1}^{d} \sigma_i^2 \frac{1+\rho^2\lambda_i^2}{\rho\lambda_i^2} > 0. \tag{30}$$

We derive necessary and sufficient conditions for stepsize schedulers to remediate the convergence deficiency.

Proposition 4.10. *Under the assumptions of Prop. 4.9, for any stepsize scheduler η_t and ρ_t, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2}$ is equal to*

$$\sum_{i=1}^{d} e^{-2\lambda_i^2\rho\int_0^t \eta_s\rho_s ds}\left(\frac{\|Z_0^i\|^2}{2}\right. \tag{31}$$
$$\left. + \eta\sigma_i^2 \int_0^t e^{2\lambda_i^2\rho\int_0^s \eta_r\rho_r dr}\eta_s^2(1+\lambda_i^2\rho^2\rho_s^2) ds\right).$$

Therefore, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{\to} 0$ if and only if

$$\int_0^\infty \eta_s\rho_s ds = \infty \text{ and } \lim_{t\to\infty}\eta_t\rho_t = \lim_{t\to\infty}\frac{\eta_t}{\rho_t} = 0. \tag{32}$$

In particular, when $\rho_t = 1$,

$$\eta_t = \frac{1}{(t+1)^\gamma} \implies \frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \to 0, \text{ for } \gamma \in \{0.5, 1\}.$$

The left of Figure 3 shows the empirical validation of Prop. 4.4 and Prop. 4.5 while its right side shows that of Prop. 4.9 and Prop. 4.10. Figure 7 shows the same for Prop. 4.3 and Prop. 4.8. More details are available in Appendix G.

Conclusion:

1. If the uncertainty v_t is well behaved as in Prop. 4.3 and Prop. 4.8, the Hamiltonian decays exponentially to 0;
2. When v_t is constant as in Prop. 4.4 and Prop. 4.9, both algorithms exponentially reach a level of suboptimality that depends on the curvature of the landscape (and on ρ for SEG);
3. Prop. 4.5 and Prop. 4.10 provide a recipe to craft schedulers that recover convergence. We provide examples of such necessary and sufficient conditions;
4. Eq. (27) and Eq. (29) clearly show that large ρ speeds up the convergence. However, this might violate Eq. (28) and increase the suboptimality in Eq. (30).

5 QUADRATIC GAMES: EXACT DYNAMICS EXPRESSION

In this section, we derive the exact solution to the SDEs of SEG and SHGD for the Quadratic Games $f(x,y) = \frac{x^\top \mathbf{A} x}{2} + x^\top \mathbf{\Lambda} y - \frac{y^\top \mathbf{A} y}{2}$ where $\mathbf{\Lambda}$ and $\mathbf{A}$ are square, diagonal and positive semidefinite matrices. We notice that if $\mathbf{A} = \mathbf{0}$, these are classic Bilinear Games.

5.1 Exact Dynamics - SEG

Theorem 5.1 (Exact Dynamics of SEG)**.** *Under the assumptions of Corollary 3.5, we take the covariance of the noise on the gradients to be $\sigma^2 \mathbf{I}_d$ and have that*

$$Z_t = \tilde{\mathbf{E}}(t)\tilde{\mathbf{R}}(t)\left(z + \sqrt{\eta}\sigma \int_0^t \tilde{\mathbf{E}}(-s)\tilde{\mathbf{R}}(-s)\mathbf{M} dW_s\right), \tag{33}$$

$$\tilde{\mathbf{E}}(t) = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix}, \tilde{\mathbf{R}}(t) = \begin{bmatrix} \mathbf{C}(t) & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{C}(t) \end{bmatrix},$$

and $\mathbf{M} = \begin{bmatrix} \mathbf{I}_d - \rho\mathbf{A} & -\rho\mathbf{\Lambda} \\ \rho\mathbf{\Lambda} & \mathbf{I}_d - \rho\mathbf{A} \end{bmatrix}$, *where*

$$\mathbf{E}(t) := \operatorname{diag}\left(e^{\rho\left(a_1^2-\lambda_1^2\right)t - a_1 t}, \cdots, e^{\rho\left(a_d^2-\lambda_d^2\right)t - a_d t}\right), \tag{34}$$

$$\mathbf{C}(t) := \operatorname{diag}\left(\cos\left(\hat{\lambda}_1 t\right), \cdots, \cos\left(\hat{\lambda}_d t\right)\right), \tag{35}$$

$$\mathbf{S}(t) := \operatorname{diag}\left(\sin\left(\hat{\lambda}_1 t\right), \cdots, \sin\left(\hat{\lambda}_d t\right)\right), \tag{36}$$

and $\hat{\lambda}_i := \lambda_i(1 - 2\rho a_i)$. *If* $\rho\left(a_i^2 - \lambda_i^2\right) - a_i < 0$*:*

1. $\mathbb{E}\left[Z_t\right] = \tilde{\mathbf{E}}(t)\tilde{\mathbf{R}}(t) z \overset{t\to\infty}{=} 0$;
2. *The covariance matrix of Z_t is equal to*

$$\frac{\eta\sigma^2}{2}\begin{bmatrix} \mathbf{I}_d - \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d - \mathbf{E}(2t) \end{bmatrix} \bar{\Sigma} \overset{t\to\infty}{=} \frac{\eta\sigma^2}{2}\bar{\Sigma} \tag{37}$$

where $\bar{\Sigma} := \operatorname{diag}(\mathbf{B}, \mathbf{B})$ *and* $\mathbf{B}$ *is defined as*

$$\operatorname{diag}\left(\frac{(1-\rho a_1)^2 + \rho^2\lambda_1^2}{a_1 + \rho(\lambda_1^2 - a_1^2)}, \cdots, \frac{(1-\rho a_d)^2 + \rho^2\lambda_d^2}{a_d + \rho(\lambda_d^2 - a_d^2)}\right). \tag{38}$$

Proof. Since the SDE is linear, the closed-form formula of the solution Z_t is known. We use the martingale property of Brownian motion to calculate $\mathbb{E}\left[Z_t\right]$ while that of the second moment uses the Itô Isometry. □

We verify Eq. (38) in Figure 8 in Appendix.

On the sign of ρ and its magnitude:

If one can chose ρ_i for each coordinate:

1. Eq. (34) implies that SEG converges only if $\rho_i\left(a_i^2 - \lambda_i^2\right) - a_i < 0$, and that $\rho_i\left(a_i^2 - \lambda_i^2\right) < 0$ is necessary to be faster than SGDA, meaning that **negative** ρ_i might be convenient if $a_i > \lambda_i$;
2. If ρ_i has the correct sign, a **larger** absolute value implies **faster convergence**;
3. Eq. (38) implies that the asymptotic variance along the i-th coordinate $\mathbf{B}_{i,i}(\rho_i)$ **explodes** if $|\rho_i|$ is too **large** or if $\rho_i \to \frac{-a_i}{\lambda_i^2 - a_i^2}$;
4. $B_{i,i}(\rho_i)$ is a convex function of ρ_i whose minimum is realized at $\rho_i^{\mathrm{V}} = \frac{1}{a_i + \lambda_i}$; However, if ρ_i^{V} is **small**, it **slows down** the convergence.

If one has to choose a single value of ρ:

1. One has to select it as it will (de)accelerate different coordinates based on its sign;
2. The trace of $\mathbf{B}$ is a convex function of ρ, meaning that there is an optimal ρ^* that minimizes it.

5.2 Exact Dynamics - SHGD

Theorem 5.2 (Exact Dynamics of SHGD)**.** *Under the assumptions of Corollary 3.7, we take the covariance of the noise on the gradients to be equal to $\sigma^2 \mathbf{I}_d$ and have*

$$Z_t = \tilde{\mathbf{E}}(t)\left(z + \sqrt{\eta}\sigma \int_0^t \tilde{\mathbf{E}}(-s)\mathbf{M} dW_s\right), \tag{39}$$

$$\tilde{\mathbf{E}}(t) = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix}, \mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{\Lambda} \\ \mathbf{\Lambda} & -\mathbf{A} \end{bmatrix}, \text{ where}$$

$$\mathbf{E}(t) := \operatorname{diag}\left(e^{-\left(\lambda_1^2 + a_1^2\right)t}, \cdots, e^{-\left(\lambda_d^2 + a_d^2\right)t}\right). \tag{40}$$

In particular, we have that

1. $\mathbb{E}\left[Z_t\right] = \tilde{\mathbf{E}}(t) z \overset{t\to\infty}{=} 0$;
2. *The covariance matrix of Z_t is equal to*

$$\eta\frac{\sigma^2}{2}\begin{bmatrix} \mathbf{I}_d - \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d - \mathbf{E}(2t) \end{bmatrix} \bar{\Sigma} \overset{t\to\infty}{=} \frac{\eta\sigma^2}{2}\bar{\Sigma}, \tag{41}$$

where $\bar{\Sigma} := \operatorname{diag}(\mathbf{I}_d, \mathbf{I}_d)$.

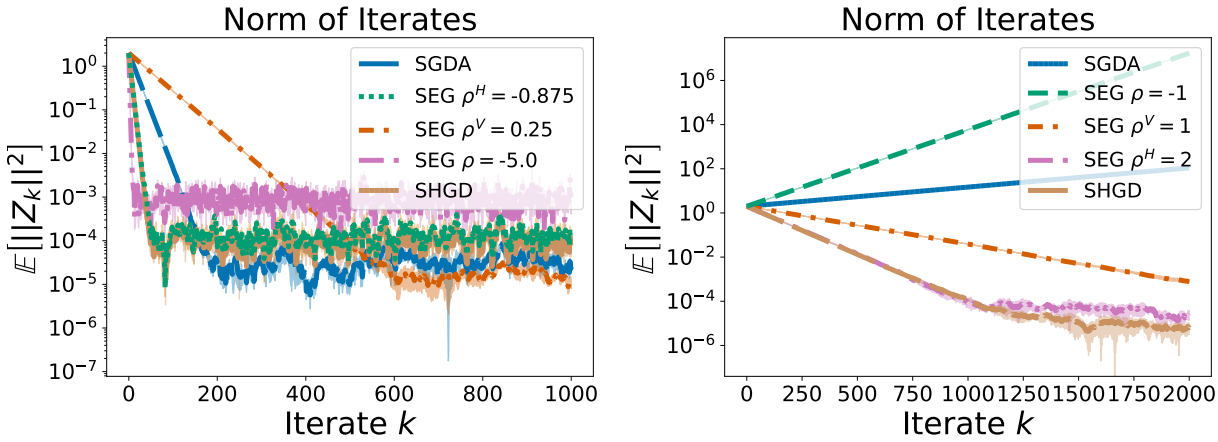

Figure 4: Comparison between SEG and SHGD on Quadratic Games: (Left), ρ^V and ρ^H meet the designated goals, sometimes **negative** ρ is desirable as positive ones **slow down** the convergence. Large $|\rho|$ induces faster convergence which in turn results in larger suboptimality. (Right), negative ρ escapes the *bad saddle* faster than SGDA, positive ones induce convergence, and ρ^H matches the decay of SHGD. In both experiments, $\eta = 0.01$.

SEG vs SHGD: Insights

1. The curvature influences the convergence speed of SHGD, but differently than for SEG, it does not affect the asymptotic covariance matrix;
2. If $(\lambda_i^2 - a_i^2)\rho_i^H > a_i^2 + \lambda_i^2 - a_i$, SEG exponentially decays **faster** than SHGD. However, this means that SEG has a **larger** asymptotic variance;
3. If $\rho_i^V = \frac{1}{\lambda_i + a_i}$, SEG attains its lowest asymptotic variance $\frac{\eta\sigma^2\lambda_i}{2(a_i+\lambda_i)^2}$, which is smaller than $\frac{\eta\sigma^2}{2}$ reached by SHGD only if $a_i^2 + \lambda_i^2 - \lambda_i > 0$;
4. If $\lambda_i^2 + a_i^2 \sim 0$, SHGD is essentially stuck and SEG is intrinsically faster;
5. If $a_i < 0$, $Z = 0$ is a *bad saddle*. While SEG can escape it, SHGD is pulled towards it.

Conclusion Our results allowed us to carry out a 1-to-1 comparison of the two methods, shedding light on the role of ρ in influencing the behavior of SEG w.r.t. SHGD. The interaction between the curvature and the noise implies that selecting ρ_i is a trade-off between the speed of convergence and the asymptotic variance. There is no clear winner between SEG and SHGD, as they are preferable for different landscapes. Figure 4 shows experiments that support the latter claim.

6 CONCLUSIONS

We have presented and analyzed the first formal SDE models for SGDA, SEG, and SHGD. We have shown the *implicit* regularization in SEG in contrast with the *explicit* use of curvature-based information in SHGD, which leads to different noise structures and asymptotic suboptimality.

Furthermore, we have used these SDEs to fully characterize the evolution of the Hamiltonian under the dynamics of these algorithms in useful scenarios. We derived convergence bounds and established conditions under which stepsize schedulers guarantee convergence.

Finally, our comparative analysis of SEG and SHGD for Quadratic Games sheds light on the role of ρ, revealing a trade-off between convergence speed and suboptimality. We also presented the first theoretical and experimental evidence that, depending on the curvature of the loss, the optimal ρ might be negative.

Outlook. Our framework offers a unified and structured analytical approach rooted in Itô calculus to study minimax optimizers. Our approach not only facilitates the derivation of novel insights but also enables straightforward comparisons between discrete algorithms. We believe our findings provide a foundation for future research, which may include the analysis of momentum, adaptive methods, derivation of scaling laws, and the design of new optimizers.

Limitations. Modeling discrete-time algorithms using SDEs hinges on Assumption 3.1. As documented (Li et al., 2021), large values of the stepsize η or the absence of specific conditions on ∇f and the noise covariance matrix can result in an approximation failure. While these shortcomings can be mitigated by increasing the order of the weak approximation, our perspective aligns with the idea that SDEs should primarily serve as simplification tools — to solidify our intuition — and might not gain substantial benefits from additional complexity.

7 ACKNOWLEDGEMENTS

We thank the reviewers for their feedback which greatly helped us improve this manuscript.

Additionally, we thank Dr. Junchi Yang and Prof. Dr. Niao He for the insightful discussions.

Enea Monzio Compagnoni and Aurelien Lucchi acknowledge the financial support of the Swiss National Foundation, SNF grant No 207392. Antonio Orvieto acknowledges the financial support of the Hector Foundation. Frank Norbert Proske acknowledges the financial support of the Norwegian Research Council (project No 274410) and MSCA4Ukraine (project No 101101923).

References

An, J., Lu, J., and Ying, L. (2020). Stochastic modified equations for the asynchronous stochastic gradient descent. *Information and Inference: A Journal of the IMA*, 9(4):851–873.

Balduzzi, D., Racaniere, S., Martens, J., Foerster, J., Tuyls, K., and Graepel, T. (2018). The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 354–363. PMLR.

Chaudhari, P. and Soatto, S. (2018). Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE.

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

Chavdarova, T., Hsieh, Y.-P., and Jordan, M. I. (2022). Continuous-time analysis for variational inequalities: An overview and desiderata. *arXiv preprint arXiv:2207.07105*.

Chavdarova, T., Jordan, M. I., and Zampetakis, M. (2023). Last-iterate convergence of saddle point optimizers via high-resolution differential equations. In *Minimax Theory and its Applications 08 (2023), No. 2*, pages 333–380. Heldermann Verlag.

Chen, C., Ding, N., and Carin, L. (2015). On the convergence of stochastic gradient mcmc algorithms with high-order integrators. *Advances in neural information processing systems*, 28.

Chen, G. H. and Rockafellar, R. T. (1997). Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444.

Compagnoni, E. M., Biggio, L., Orvieto, A., Proske, F. N., Kersting, H., and Lucchi, A. (2023). An sde for modeling sam: Theory and insights. In *International Conference on Machine Learning*, pages 25209–25253. PMLR.

Du, S. S., Gidel, G., Jordan, M. I., and Li, C. J. (2022). Optimal extragradient-based bilinearly-coupled saddle-point optimization. *arXiv preprint arXiv:2206.08573*.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. (2019). A variational inequality perspective on generative adversarial networks. *ICLR*.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT Press.

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

He, L., Meng, Q., Chen, W., Ma, Z.-M., and Liu, T.-Y. (2018). Differential equations for modeling asynchronous algorithms. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 2220–2226. AAAI Press.

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

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

Hsieh, Y.-P., Mertikopoulos, P., and Cevher, V. (2021). The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. In *International Conference on Machine Learning*, pages 4337–4348. PMLR.

Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., and Storkey, A. (2018). Three factors influencing minima in sgd. *ICANN 2018*.

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

Korpelevich, G. M. (1976). The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

Li, C. J., Yu, Y., Loizou, N., Gidel, G., Ma, Y., Le Roux, N., and Jordan, M. (2022). On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging. In *International Conference on Artificial Intelligence and Statistics*, pages 9793–9826. PMLR.

Li, Q., Tai, C., and Weinan, E. (2017). Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110. PMLR.

Li, Q., Tai, C., and Weinan, E. (2019). Stochastic modified equations and dynamics of stochastic gradient algorithms i: Mathematical foundations. *The Journal of Machine Learning Research*, 20(1):1474–1520.

Li, Z., Malladi, S., and Arora, S. (2021). On the validity of modeling SGD with stochastic differential equations (SDEs). In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Ljung, L., Pflug, G., and Walk, H. (2012). *Stochastic approximation and optimization of random systems*, volume 17. Birkhäuser.

Loizou, N., Berard, H., Gidel, G., Mitliagkas, I., and Lacoste-Julien, S. (2021). 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.

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

Lu, H. (2022). An o (sr)-resolution ode framework for understanding discrete-time algorithms and applications to the linear convergence of minimax problems. *Mathematical Programming*, 194(1-2):1061–1112.

Malladi, S., Lyu, K., Panigrahi, A., and Arora, S. (2022). On the SDEs and scaling rules for adaptive gradient algorithms. In *Advances in Neural Information Processing Systems*.

Mandt, S., Hoffman, M. D., Blei, D. M., et al. (2015). Continuous-time limit of stochastic gradient descent revisited. *NIPS-2015*.

Mao, X. (2007). *Stochastic differential equations and applications*. Elsevier.

Mil'shtein, G. (1986). Weak approximation of solutions of systems of stochastic differential equations. *Theory of Probability & Its Applications*, 30(4):750–766.

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

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

Noor, M. A. (2003). New extragradient-type methods for general variational inequalities. *Journal of Mathematical Analysis and Applications*, 277(2):379–394.

Øksendal, B. (1990). When is a stochastic integral a time change of a diffusion? *Journal of theoretical probability*, 3(2):207–226.

Orvieto, A. and Lucchi, A. (2019). Continuous-time models for stochastic optimization algorithms. *Advances in Neural Information Processing Systems*, 32.

Ryu, E. K., Yuan, K., and Yin, W. (2019). Ode analysis of stochastic gradient methods with optimism and anchoring for minimax problems. *arXiv preprint arXiv:1905.10899*.

Su, W., Boyd, S., and Candes, E. (2014). A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*.

Xie, Z., Sato, I., and Sugiyama, M. (2021). A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. In *International Conference on Learning Representations*.

Xu, M. et al. (2022). *Experimental Evaluation of Iterative Methods for Games*. PhD thesis, Johns Hopkins University.

Zhao, J., Lucchi, A., Proske, F. N., Orvieto, A., and Kersting, H. (2022). Batch size selection by stochastic optimal control. In *Has it Trained Yet? NeurIPS 2022 Workshop*.

Zhu, Z., Wu, J., Yu, B., Wu, L., and Ma, J. (2019). The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. *ICML 2019*.

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, we provide the assumptions of the mathematical setting we work with, all assumptions of theorems, a clear definition of the algorithms, and of the models analyzed.]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes, we provide convergence conditions and convergence rates.]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [The source code will be provided.]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes. May statements are informal in the main paper, but their formal counterpart are clearly stated in the appendix.]

(b) Complete proofs of all theoretical results. [Yes, all proofs are in the appendix]

(c) Clear explanations of any assumptions. [Yes, we cite prior work that used similar assumptions]

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]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, details are presented in the appendix.]

(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, details are presented in the appendix.]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes, details are presented in the appendix.]

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. [Not Applicable]

(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. [Not Applicable]

(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]

APPENDIX

A ADDITIONAL RELATED WORKS

SGDA is one of the most popular algorithms for solving min-max optimization problems that arise in machine learning. Since it does not converge even on simple landscapes (Chen and Rockafellar, 1997; Noor, 2003; Gidel et al., 2019; Loizou et al., 2021), researchers have derived several advanced extensions such as the Extragradient method (Korpelevich, 1976) and variants with arbitrary sampling and variance (Gorbunov et al., 2022), as well as alternative optimizers such as (Stochastic) Hamiltonian Gradient Descent (Balduzzi et al., 2018; Loizou et al., 2020).

Stochastic ExtraGradient (SEG) is a prominent extension of SGDA that has been studied extensively in recent years. Indeed, many versions have been proposed and studied: (Nemirovski et al., 2009; Juditsky et al., 2011) studied Independent-Samples SEG, while Mishchenko et al. (2020) and Li et al. (2022) showed that the average iterate of Same-Sample SEG converges to a neighbor of the optimum. While (Chavdarova et al., 2019) showed that same-stepsize SEG diverges in the unconstrained monotone case, Mishchenko et al. (2020); Hsieh et al. (2020) focused on two-scale SEG, showcasing how this design choice is crucial by deriving schedulers that guarantee convergence. Hsieh et al. (2019) studies the convergence of variations of SEG engineered to mitigate the cost of the extra gradient. Finally, Gorbunov et al. (2022) provides a rich analysis that encompasses several variants of SEG with different choices of stepsizes and sampling techniques, and ends up designing new promising methods. The latter endeavor is key for future research: As highlighted by (Hsieh et al., 2019), existing min-max algorithms may be subject to inescapable convergence failures in important cases.

Among other works, we refer the reader interested in previous analyses of bilinear and quadratic games to (Hsieh et al., 2021; Li et al., 2022; Xu et al., 2022; Chavdarova et al., 2023): These give a detailed presentation of the behavior of GDA, EG and HGD, and their stochastic versions on such tasks.

We highlight that some convergence conditions and some of the convergence bounds derived in the literature for SEG (see among others (Hsieh et al., 2020; Mishchenko et al., 2020; Hsieh et al., 2019; Gorbunov et al., 2022; Lu, 2022; Li et al., 2022)) and SHGD (see (Loizou et al., 2020, 2021)) are somehow related to those we present in this paper.

B STOCHASTIC CALCULUS

In this section, we summarize some important results in the analysis of Stochastic Differential Equations Mao (2007); Øksendal (1990). The notation and the results in this section will be used extensively in all proofs in this paper. We assume the reader to have some familiarity with Brownian motion and with the definition of stochastic integral (Ch. 1.4 and 1.5 in Mao (2007)).

B.1 Itô's Lemma

We start with some notation: Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be a filtered probability space. We say that an event $E \in \mathcal{F}$ holds almost surely (a.s.) in this space if $\mathbb{P}(E) = 1$. We call $\mathcal{L}^p([a,b], \mathbb{R}^d)$, with $p > 0$, the family of $\mathbb{R}^d$-valued $\mathcal{F}_t$-adapted processes $\{f_t\}_{a\leq t\leq b}$ such that

$$\int_a^b \|f_t\|^p dt \leq \infty.$$

Moreover, we denote by $\mathcal{M}^p([a,b], \mathbb{R}^d)$, with $p > 0$, the family of $\mathbb{R}^d$-valued processes $\{f_t\}_{a\leq t\leq b}$ in $\mathcal{L}([a,b], \mathbb{R}^d)$ such that $\mathbb{E}\left[\int_a^b \|f_t\|^p dt\right] \leq \infty$. We will write $h \in \mathcal{L}^p\left(\mathbb{R}_+, \mathbb{R}^d\right)$, with $p > 0$, if $h \in \mathcal{L}^p\left([0,T], \mathbb{R}^d\right)$ for every $T > 0$. Similar definitions hold for matrix-valued functions using the Frobenius norm $\|A\| := \sqrt{\sum_{ij} |A_{ij}|^2}$.

Let $W = \{W_t\}_{t\geq 0}$ be a one-dimensional Brownian motion defined on our probability space and let $X = \{X_t\}_{t\geq 0}$ be an $\mathcal{F}_t$-adapted process taking values on $\mathbb{R}^d$.

Definition B.1. Let the *drift* be $b \in \mathcal{L}^1\left(\mathbb{R}_+, \mathbb{R}^d\right)$ and the diffusion term be $\sigma \in \mathcal{L}^2\left(\mathbb{R}_+, \mathbb{R}^{d\times m}\right)$. X_t is an Itô process if it takes the form

$$X_t = x_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s.$$

We shall say that X_t has the stochastic differential

$$dX_t = b_t dt + \sigma_t dW_t. \tag{42}$$

Theorem B.2 (Itô's Lemma)**.** *Let X_t be an Itô process with stochastic differential $dX_t = b_t dt + \sigma_t dW_t$. Let $f(x,t)$ be twice continuously differentiable in x and continuously differentiable in t, taking values in $\mathbb{R}$. Then $f(X_t,t)$ is again an Itô process with stochastic differential*

$$df(X_t,t) = \partial_t f(X_t,t))dt + \langle \nabla f(X_t,t), b_t\rangle dt + \frac{1}{2} Tr\left(\sigma_t \sigma_t^\top \nabla^2 f(X_t,t)\right) dt + \langle \nabla f(X_t,t), \sigma_t\rangle dW_t. \tag{43}$$

B.2 Stochastic Differential Equations

Stochastic Differential Equations (SDEs) are equations of the form

$$dX_t = b(X_t,t)dt + \sigma(X_t,t)dW_t.$$

First of all, we need to define what it means for a stochastic process $X = \{X_t\}_{t\geq 0}$ with values in $\mathbb{R}^d$ to solve an SDE.

Definition B.3. Let X_t be as above with deterministic initial condition $X_0 = x_0$. Assume $b : \mathbb{R}^d \times [0,T] \to \mathbb{R}^d$ and $\sigma : \mathbb{R}^d \times [0,T] \to \mathbb{R}^{d\times m}$ are Borel measurable; X_t is called a solution to the corresponding SDE if

1. X_t is continuous and $\mathcal{F}_t$-adapted;
2. $b \in \mathcal{L}^1\left([0,T], \mathbb{R}^d\right)$;
3. $\sigma \in \mathcal{L}^2\left([0,T], \mathbb{R}^{d\times m}\right)$;
4. For every $t \in [0,T]$
$$X_t = x_0 + \int_0^t b(X_s,s)ds + \int_0^t \sigma(X_s,s)dW(s) \quad a.s.$$

Moreover, the solution X_t is said to be unique if any other solution $X_t^\star$ is such that

$$\mathbb{P}\left\{X_t = X_t^\star, \text{ for all } 0 \leq t \leq T\right\} = 1.$$

Notice that since the solution to an SDE is an Itô process, we can use Itô's Lemma. The following theorem gives a sufficient condition on b and σ for the existence of a solution to the corresponding SDE.

Theorem B.4. *Assume that there exist two positive constants $\bar{K}$ and K such that*

1. *(Global Lipschitz condition) for all $x, y \in \mathbb{R}^d$ and $t \in [0,T]$*
$$\max\{\|b(x,t) - b(y,t)\|^2, \ \|\sigma(x,t) - \sigma(y,t)\|^2\} \leq \bar{K}\|x-y\|^2;$$
2. *(Linear growth condition) for all $x \in \mathbb{R}^d$ and $t \in [0,T]$*
$$\max\{\|b(x,t)\|^2, \ \|\sigma(x,t)\|^2\} \leq K(1+\|x\|^2).$$

Then, there exists a unique solution X_t to the corresponding SDE, and $X_t \in \mathcal{M}^2([0,T], \mathbb{R}^d)$.

Numerical approximation. Often, SDEs are solved numerically. The simplest algorithm to provide a sample path $(\hat{x}_k)_{k\geq 0}$ for X_t, so that $X_{k\Delta t} \approxeq \hat{x}_k$ for some small Δt and for all $k\Delta t \leq M$ is called Euler-Maruyama (Algorithm 1). For more details on this integration method and its approximation properties, the reader can check Mao (2007).

Algorithm 1 Euler-Maruyama Integration Method for SDEs

input The drift b, the volatility σ, and the initial condition x_0.
  Fix a stepsize Δt;
  Initialize $\hat{x}_0 = x_0$;
  $k = 0$;
  while $k \leq \left\lfloor \frac{T}{\Delta t} \right\rfloor$ **do**
    Sample some d-dimensional Gaussian noise $Z_k \sim \mathcal{N}(0, I_d)$;
    Compute $\hat{x}_{k+1} = \hat{x}_k + \Delta t \; b(\hat{x}_k, k\Delta t) + \sqrt{\Delta t} \; \sigma(\hat{x}_k, k\Delta t) Z_k$;
    $k = k + 1$;
  end while
output The approximated sample path $(\hat{x}_k)_{0 \leq k \leq \lfloor \frac{T}{\Delta t} \rfloor}$.

C THEORETICAL FRAMEWORK - WEAK APPROXIMATION

In this section, we introduce the theoretical framework used in the paper, together with its assumptions and notations.

First of all, many proofs will use Taylor expansions in powers of η. For ease of notation, we introduce the shorthand that whenever we write $\mathcal{O}(\eta^\alpha)$, we mean that there exists a function $K(z) \in G$ such that the error terms are bounded by $K(z)\eta^\alpha$. For example, we write

$$b(z + \eta) = b_0(z) + \eta b_1(z) + \mathcal{O}\left(\eta^2\right)$$

to mean: there exists $K \in G$ such that

$$|b(z + \eta) - b_0(z) - \eta b_1(z)| \leq K(z)\eta^2.$$

Additionally, we introduce the following shorthand:

- A multi-index is $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ such that $\alpha_j \in \{0, 1, 2, \ldots\}$;
- $|\alpha| := \alpha_1 + \alpha_2 + \cdots + \alpha_n$;
- $\alpha! := \alpha_1! \alpha_2! \cdots \alpha_n!$;
- For $z = (z_1, z_2, \ldots, z_n) \in \mathbb{R}^n$, we define $z^\alpha := z_1^{\alpha_1} z_2^{\alpha_2} \cdots z_n^{\alpha_n}$;
- For a multi-index β, $\partial_\beta^{|\beta|} f(z) := \frac{\partial^{|\beta|}}{\partial_{z_1}^{\beta_1} \partial_{z_2}^{\beta_2} \cdots \partial_{z_n}^{\beta_n}} f(z)$;
- We also denote the partial derivative with respect to z_i by ∂_{e_i}.

Definition C.1 (G Set)**.** Let G denote the set of continuous functions $\mathbb{R}^{2d} \to \mathbb{R}$ of at most polynomial growth, i.e. $g \in G$ if there exists positive integers $\nu_1, \nu_2 > 0$ such that $|g(z)| \leq \nu_1 \left(1 + |z|^{2\nu_2}\right)$, for all $z \in \mathbb{R}^{2d}$.

The next results are inspired by Theorem 1 of Li et al. (2017) and are derived under some regularity assumption on the function f.

Assumption C.2. Assume that the following conditions on f, f_i, and their gradients are satisfied:

- $\nabla f, \nabla f_i$ satisfy a Lipschitz condition: There exists $L > 0$ such that

$$|\nabla f(u) - \nabla f(v)| + \sum_{i=1}^{N} |\nabla f_i(u) - \nabla f_i(v)| \leq L|u - v|;$$

- f, f_i and its partial derivatives up to order 7 belong to G;
- $\nabla f, \nabla f_i$ satisfy a growth condition: There exists $M > 0$ such that

$$|\nabla f(z)| + \sum_{i=1}^{N} |\nabla f_i(z)| \leq M(1 + |z|).$$

Lemma C.3 (Lemma 1 Li et al. (2017))**.** *Let $0 < \eta < 1$. Consider a stochastic process $Z_t, t \geq 0$ satisfying the SDE*

$$dZ_t = b\left(Z_t\right) dt + \sqrt{\eta}\sigma\left(Z_t\right) dW_t$$

with $Z_0 = z \in \mathbb{R}^{2d}$ and b, σ together with their derivatives belong to G. Define the one-step difference $\Delta = Z_\eta - z$, and indicate the i-th component of Δ with Δ_i. Then we have

1. $\mathbb{E}\Delta_i = b_i\eta + \frac{1}{2}\left[\sum_{j=1}^{d} b_j \partial_{e_j} b_i\right]\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i = 1, \ldots, 2d;$
2. $\mathbb{E}\Delta_i\Delta_j = \left[b_i b_j + \sigma\sigma^T_{(ij)}\right]\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \ldots, 2d;$
3. $\mathbb{E}\prod_{j=1}^{s} \Delta_{(i_j)} = \mathcal{O}\left(\eta^3\right)$ *for all* $s \geq 3, i_j = 1, \ldots, 2d.$

All functions above are evaluated at z.

Theorem C.4 (Theorem 2 and Lemma 5, Mil'shtein (1986))**.** *Let Assumption C.2 hold and let us define $\bar{\Delta} = z_1 - z$ to be the increment in the discrete-time algorithm, and indicate the i-th component of $\bar{\Delta}$ with $\bar{\Delta}_i$. If in addition there exists $K_1, K_2, K_3, K_4 \in G$ so that*

1. $\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(z)\eta^2, \quad \forall i = 1, \ldots, 2d;$
2. $\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(z)\eta^2, \quad \forall i, j = 1, \ldots, 2d;$
3. $\left|\mathbb{E}\prod_{j=1}^{s} \Delta_{i_j} - \mathbb{E}\prod_{j=1}^{s} \bar{\Delta}_{i_j}\right| \leq K_3(z)\eta^2, \quad \forall s \geq 3, \quad \forall i_j \in \{1, \ldots, 2d\};$
4. $\mathbb{E}\prod_{j=1}^{3} \left|\bar{\Delta}_{i_j}\right| \leq K_4(z)\eta^2, \quad \forall i_j \in \{1, \ldots, 2d\}.$

Then, there exists a constant C so that for all $k = 0, 1, \ldots, N$ we have

$$\left|\mathbb{E}g\left(Z_{k\eta}\right) - \mathbb{E}g\left(z_k\right)\right| \leq C\eta.$$

C.1 Formal Derivation - SGDA

In this subsection, we provide the first formal derivation of an SDE model for SGDA. Let us consider the stochastic process $Z_t \in \mathbb{R}^{2d}$ defined as the solution of

$$dZ_t = -\eta_t \circ F\left(Z_t\right) dt + \sqrt{\eta}(\eta_t 1^\top) \circ \sqrt{\Sigma\left(Z_t\right)} dW_t, \tag{44}$$

where

$$\Sigma(z) := \mathbb{E}\left[\left(F\left(z\right) - F_{\gamma}\left(z\right)\right)\left(F\left(z\right) - F_{\gamma}\left(z\right)\right)^{\top}\right], \tag{45}$$

and the $\circ$ symbol represents the Hadamard product. The following theorem guarantees that such a process is a 1-order SDE of the discrete-time algorithm of SGDA

$$z_{k+1} = z_k - \eta\eta_k \circ F_{\gamma_k}\left(z_k\right) \tag{46}$$

with $z_0 := z = (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, which is an extension of 2.

Theorem C.5 (Stochastic modified equations)**.** *Let $0 < \eta < 1, T > 0$ and set $N = \lfloor T/\eta \rfloor$. Let $z_k \in \mathbb{R}^{2d}, 0 \leq k \leq N$ denote a sequence of SGDA iterations defined by Eq.* (46)*. Consider the stochastic process Z_t defined in Eq.* (44) *and fix some test function $g \in G$ and suppose that g and its partial derivatives up to order 6 belong to G.*
Then, under Assumption C.2, there exists a constant $C > 0$ independent of η such that for all $k = 0, 1, \ldots, N$, we have

$$\left|\mathbb{E}g\left(Z_{k\eta}\right) - \mathbb{E}g\left(z_k\right)\right| \leq C\eta.$$

That is, the SDE (44) *is an order* 1 *weak approximation of the SGDA iterations* (46)*.*

Lemma C.6. *Under the assumptions of Theorem C.5, let $0 < \eta < 1$ and consider $z_k, k \geq 0$ satisfying the SGDA iterations*

$$z_{k+1} = z_k - \eta\eta_k \circ F_{\gamma_k}\left(z_k\right)$$

with $z_0 := z = (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$. From the definition the one-step difference $\bar{\Delta} = z_1 - z$, then we have

1. $\mathbb{E}\bar{\Delta}_i = -\eta_0^i F_i \eta \quad \forall i = 1, \ldots, 2d;$
2. $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \eta_0^i\eta_0^j F_i F_j \eta^2 + \eta_0^i\eta_0^j\Sigma_{(ij)}\eta^2 \quad \forall i, j = 1, \ldots, 2d;$
3. $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, 2d\}.$

All the functions above are evaluated at z.

Proof of Lemma C.6. First of all, we have that by definition

$$\mathbb{E}\left[z_1 - z\right] = -\eta\eta_0 \circ F(z), \tag{47}$$

which implies

$$\mathbb{E}\bar{\Delta}_i = -\eta_0^i F_i\left(z\right)\eta \quad \forall i = 1, \ldots, 2d. \tag{48}$$

Second, we have that by definition

$$\begin{aligned}\mathbb{E}\left[\left(z_1 - z\right)\left(z_1 - z\right)^{\top}\right] &= \eta^2\eta_0 \circ F(z)F(z)^{\top} \circ \eta_0^{\top} + \eta^2\mathbb{E}\left[\eta_0 \circ \left(F(z) - F_{\gamma}(z)\right)\left(F(z) - F_{\gamma}(z)\right)^{\top} \circ \eta_0^{\top}\right] \\ &= \eta^2\eta_0 \circ F(z)F(z)^{\top} \circ \eta_0^{\top} + \eta^2(\eta_0 1^{\top}) \circ \Sigma(z) \circ (\eta_0 1^{\top})^{\top},\end{aligned} \tag{49}$$

which implies that

$$\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = \eta_0^i\eta_0^j F_i\left(z\right)F_j\left(z\right)\eta^2 + \eta_0^i\eta_0^j\Sigma_{(ij)}\left(z\right)\eta^2 \quad \forall i, j = 1, \ldots, 2d. \tag{50}$$

Finally, by definition

$$\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j}=\mathcal{O}\left(\eta^3\right)\quad\forall s\geq 3,\quad i_j\in\{1,\ldots,2d\},\tag{51}$$

which concludes our proof. □

Proof of Theorem C.5. To prove this result, all we need to do is check the conditions in Theorem C.4. As we apply Lemma C.3, we make the following choices:

- $b(z)=-\eta_t\circ F(z)$;
- $\sigma(z)=(\eta_t 1^\top)\circ\sqrt{\Sigma(z)}$.

First of all, we notice that $\forall i=1,\ldots,2d$, it holds that

- $\mathbb{E}\bar{\Delta}_i \overset{\text{1. Lemma C.6}}{=} -\eta_0^i F_i\left(z\right)\eta$;
- $\mathbb{E}\Delta_i \overset{\text{1. Lemma C.3}}{=} -\eta_0^i F_i\left(z\right)\eta+\mathcal{O}\left(\eta^2\right)$.

Therefore, we have that for some $K_1(z)\in G$,

$$\left|\mathbb{E}\Delta_i-\mathbb{E}\bar{\Delta}_i\right|\leq K_1(z)\eta^2,\quad\forall i=1,\ldots,2d.\tag{52}$$

Additionally, we notice that $\forall i,j=1,\ldots,d$, it holds that

- $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j \overset{\text{2. Lemma C.6}}{=} \eta_0^i\eta_0^j F_i\left(z\right)F_j\left(z\right)\eta^2+\eta_0^i\eta_0^j\Sigma_{(ij)}\left(z\right)\eta^2$;
- $\mathbb{E}\Delta_i\Delta_j \overset{\text{2. Lemma C.3}}{=} \eta_0^i\eta_0^j F_i\left(z\right)F_j\left(z\right)\eta^2+\eta_0^i\eta_0^j\Sigma_{(ij)}\left(z\right)\eta^2+\mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_2(z)\in G$,

$$\left|\mathbb{E}\Delta_i\Delta_j-\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right|\leq K_2(z)\eta^2,\quad\forall i,j=1,\ldots,2d.\tag{53}$$

Additionally, we notice that $\forall s\geq 3,\forall i_j\in\{1,\ldots,2d\}$, it holds that

- $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} \overset{\text{3. Lemma C.6}}{=} \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} \overset{\text{3. Lemma C.3}}{=} \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_3(z)\in G$,

$$\left|\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j}-\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j}\right|\leq K_3(z)\eta^2.\tag{54}$$

Additionally, for some $K_4(z)\in G$, $\forall i_j\in\{1,\ldots,d\}$,

$$\mathbb{E}\prod_{j=1}^{3}\left|\bar{\Delta}_{(i_j)}\right| \overset{\text{3. Lemma C.6}}{\leq} K_4(z)\eta^2.\tag{55}$$

To conclude, Eq. (52), Eq. (53), Eq. (54), and Eq. (55) allow us to conclude the proof. □

Corollary C.7. *Let us take the same assumptions of Theorem C.5. Additionally, let us assume that stochastic gradients can be written as* $\nabla_x f_\gamma(z) = \nabla_x f(z) + U^x$ *and* $\nabla_y f_\gamma(z) = \nabla_y f(z) + U^y$ *such that* U^x *and* U^y *are independent noises that do not depend on* z*, whose expectation is* 0*, and whose covariance matrix is* Σ*. For* $\eta_t = \mathbf{1}$*, the SDE* (44) *becomes*

$$\begin{aligned} dX_t &= -\nabla_x f\left(Z_t\right) dt + \sqrt{\eta\Sigma} dW_t^x, \\ dY_t &= +\nabla_y f\left(Z_t\right) dt + \sqrt{\eta\Sigma} dW_t^y. \end{aligned} \tag{56}$$

Proof of Corollary C.7. It follows directly by the independence of the noise, the definition of the scheduler, and the definition of the covariance matrix. □

C.2 Formal Derivation - SEG

In this subsection, we provide the first formal derivation of an SDE model for SEG. Before presenting the proof, we introduce some notation. Let $\boldsymbol{\gamma} := (\gamma^1, \gamma^2)$, $\bar{F}_{\boldsymbol{\gamma}}(z) := \nabla F_{\gamma^1}(z) F_{\gamma^2}(z)$, and $\bar{F}(z) := \mathbb{E}[\bar{F}_{\boldsymbol{\gamma}}(z)]$ be its expectation. We denote the noise in $\bar{F}$ as $\bar{\xi}_{\boldsymbol{\gamma}}(z) := \bar{F}_{\boldsymbol{\gamma}}(z) - \bar{F}(z)$.

Let us consider the stochastic process $Z_t \in \mathbb{R}^{2d}$ defined as the solution of

$$dZ_t = -F^{\mathrm{SEG}}\left(Z_t\right) dt + \sqrt{\eta \Sigma^{\mathrm{SEG}}\left(Z_t\right)} dW_t, \tag{57}$$

with

$$F^{\mathrm{SEG}}(z) := F(z) - \rho \bar{F}(z), \tag{58}$$

$$\Sigma^{\mathrm{SEG}}(z) := \Sigma(z) + \rho\left[\bar{\Sigma}(z) + \bar{\Sigma}(z)^\top\right], \tag{59}$$

where $\bar{\Sigma}(z)$ is defined as

$$\mathbb{E}\left[\xi_{\gamma^1}(z) \bar{\xi}_{\boldsymbol{\gamma}}(z)^\top\right] = \mathbb{E}\left[\left(F\left(z\right) - F_{\gamma^1}\left(z\right)\right)\left(\mathbb{E}\left[\nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right] - \nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right)^\top\right]. \tag{60}$$

Theorem C.8 (Stochastic modified equations)**.** *Let* $0 < \eta < 1, T > 0$ *and set* $N = \lfloor T/\eta \rfloor$*. Let* $z_k \in \mathbb{R}^{2d}, 0 \leq k \leq N$ *denote a sequence of SEG iterations defined by Eq.* (3)*. Additionally, let us take*

$$\rho = \mathcal{O}\left(\sqrt{\eta}\right). \tag{61}$$

Consider the stochastic process Z_t *defined in Eq.* (57) *and fix some test function* $g \in G$ *and suppose that* g *and its partial derivatives up to order 6 belong to* G*.*
Then, under Assumption C.2, there exists a constant $C > 0$ *independent of* η *such that for all* $k = 0, 1, \ldots, N$*, we have*

$$\left|\mathbb{E} g\left(Z_{k\eta}\right) - \mathbb{E} g\left(z_k\right)\right| \leq C\eta.$$

That is, the SDE (57) *is an order* 1 *weak approximation of the SEG iterations* (3)*.*

Lemma C.9. *Under the assumptions of Theorem C.8, let $0 < \eta < 1$ and consider $z_k, k \geq 0$ satisfying the SEG iterations* (3)

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \eta \begin{bmatrix} +\nabla_x f_{\gamma_k^1}\left(x_k - \rho \nabla_x f_{\gamma_k^2}(x_k, y_k), y_k + \rho \nabla_y f_{\gamma_k^2}(x_k, y_k)\right) \\ -\nabla_y f_{\gamma_k^1}\left(x_k - \rho \nabla_x f_{\gamma_k^2}(x_k, y_k), y_k + \rho \nabla_y f_{\gamma_k^2}(x_k, y_k)\right) \end{bmatrix} \tag{62}$$

with $z_0 := z = (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$. From the definition the one-step difference $\bar{\Delta} = z_1 - z$, then we have

1. $\mathbb{E}\bar{\Delta}_i = -F_i^{SEG}\eta + \mathcal{O}\left(\eta^2\right) \quad \forall i = 1, \ldots, 2d;$
2. $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = F_i^{SEG}F_j^{SEG}\eta^2 + \Sigma_{(ij)}^{SEG}\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \ldots, 2d;$
3. $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, 2d\}.$

All the functions above are evaluated at z.

Proof of Lemma C.9. First of all, we have that by definition and using a Taylor expansion,

$$\begin{aligned} x_{k+1} &= x_k - \eta \nabla_x f_{\gamma_k^1}\left(x_k - \rho \nabla_x f_{\gamma_k^2}(x_k, y_k), y_k + \rho \nabla_y f_{\gamma_k^2}(x_k, y_k)\right) \\ &= x_k - \eta \nabla_x f_{\gamma_k^1}(x_k, y_k) + \eta\rho \nabla_{xx}^2 f_{\gamma_k^1}(x_k, y_k) \nabla_x f_{\gamma_k^2}(x_k, y_k) - \eta\rho \nabla_{xy}^2 f_{\gamma_k^1}(x_k, y_k) \nabla_y f_{\gamma_k^2}(x_k, y_k) + \mathcal{O}\left(\eta\rho^2\right), \end{aligned} \tag{63}$$

and

$$\begin{aligned} y_{k+1} &= y_k + \eta \nabla_y f_{\gamma_k^1}\left(x_k - \rho \nabla_x f_{\gamma_k^2}(x_k, y_k), y_k + \rho \nabla_y f_{\gamma_k^2}(x_k, y_k)\right) \\ &= y_k + \eta \nabla_y f_{\gamma_k^1}(x_k, y_k) - \eta\rho \nabla_{xy}^2 f_{\gamma_k^2}(x_k, y_k) \nabla_x f_{\gamma_k^2}(x_k, y_k) + \eta\rho \nabla_{yy}^2 f_{\gamma_k}(x_k, y_k) \nabla_y f_{\gamma_k}(x_k, y_k) + \mathcal{O}\left(\eta\rho^2\right). \end{aligned} \tag{64}$$

Therefore

$$z_1 = z - \eta F_{\gamma^1}(z) + \eta\rho \nabla F_{\gamma^1}(z) F_{\gamma^2}(z) + \mathcal{O}\left(\eta^2\right), \tag{65}$$

which implies that

$$\begin{aligned} \mathbb{E}\left[z_1 - z\right] &= -\eta F(z) + \eta\rho \mathbb{E}\left[\nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right] + \mathcal{O}\left(\eta^2\right) \\ &= z - \eta F^{\mathrm{SEG}}(z) + \mathcal{O}\left(\eta^2\right), \end{aligned} \tag{66}$$

where $F^{\mathrm{SEG}}(z) := F(z) - \rho \mathbb{E}\left[\nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right]$, which in turn implies that

$$\mathbb{E}\bar{\Delta}_i = -F_i^{\mathrm{SEG}}(z)\,\eta + \mathcal{O}\left(\eta^2\right) \quad \forall i = 1, \ldots, 2d. \tag{67}$$

Second, we have that

$$\begin{aligned} \mathbb{E}\left[(z_1 - z)(z_1 - z)^\top\right] &= \eta^2 \left[\left(F^{\mathrm{SEG}}(z)\right)\left(F^{\mathrm{SEG}}(z)\right)^\top\right] \\ &\quad + \eta^2 \mathbb{E}\left[\left(F(z) - F_{\gamma^1}(z)\right)\left(F(z) - F_{\gamma^1}(z)\right)^\top\right] \\ &\quad + \eta^2\rho \left(\mathbb{E}\left[\left(F(z) - F_{\gamma^1}(z)\right)\left(\mathbb{E}\left[\nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right] - \nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right)^\top\right]\right) \\ &\quad + \eta^2\rho \left(\mathbb{E}\left[\left(F(z) - F_{\gamma^1}(z)\right)\left(\mathbb{E}\left[\nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right] - \nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right)^\top\right]\right)^\top + \mathcal{O}\left(\eta^3\right) \\ &= \eta^2 \left[\left(F^{\mathrm{SEG}}(z)\right)\left(F^{\mathrm{SEG}}(z)\right)^\top\right] + \eta^2 \Sigma^{\mathrm{SEG}}(z) + \mathcal{O}\left(\eta^3\right), \end{aligned} \tag{68}$$

which implies that

$$\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = F_i^{\text{SEG}}(z)\,F_j^{\text{SEG}}(z)\,\eta^2 + \Sigma_{(ij)}^{\text{SEG}}(z)\,\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i,j = 1,\dots,2d. \tag{69}$$

Finally, by definition

$$\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1,\dots,2d\}, \tag{70}$$

which concludes our proof. □

Proof of Theorem C.8. To prove this result, all we need to do is check the conditions in Theorem C.4. As we apply Lemma C.3, we make the following choices:

- $b(z) = -F^{\text{SEG}}(z)$;
- $\sigma(z) = \Sigma^{\text{SEG}}(z)^{\frac{1}{2}}$.

First of all, we notice that $\forall i = 1,\dots,2d$, it holds that

- $\mathbb{E}\bar{\Delta}_i \overset{\text{1. Lemma C.9}}{=} -F_i^{\text{SEG}}(z)\,\eta + \mathcal{O}\left(\eta^2\right)$;
- $\mathbb{E}\Delta_i \overset{\text{1. Lemma C.3}}{=} -F_i^{\text{SEG}}(z)\,\eta + \mathcal{O}\left(\eta^2\right)$.

Therefore, we have that for some $K_1(z) \in G$,

$$\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(z)\eta^2, \quad \forall i = 1,\dots,2d. \tag{71}$$

Additionally, we notice that $\forall i,j = 1,\dots,d$, it holds that

- $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j \overset{\text{2. Lemma C.9}}{=} F_i^{\text{SEG}}(z)\,F_j^{\text{SEG}}(z)\,\eta^2 + \Sigma_{(ij)}^{\text{SEG}}(z)\,\eta^2 + \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\Delta_i\Delta_j \overset{\text{2. Lemma C.3}}{=} F_i^{\text{SEG}}(z)\,F_j^{\text{SEG}}(z)\,\eta^2 + \Sigma_{(ij)}^{\text{SEG}}(z)\,\eta^2 + \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_2(z) \in G$,

$$\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(z)\eta^2, \quad \forall i,j = 1,\dots,2d. \tag{72}$$

Additionally, we notice that $\forall s \geq 3, \forall i_j \in \{1,\dots,2d\}$, it holds that

- $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} \overset{\text{3. Lemma C.9}}{=} \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} \overset{\text{3. Lemma C.3}}{=} \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_3(z) \in G$,

$$\left|\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} - \mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j}\right| \leq K_3(z)\eta^2. \tag{73}$$

Additionally, for some $K_4(z) \in G$, $\forall i_j \in \{1,\dots,d\}$,

$$\mathbb{E}\prod_{j=1}^{3}\left|\bar{\Delta}_{(i_j)}\right| \overset{\text{3. Lemma C.9}}{\leq} K_4(z)\eta^2. \tag{74}$$

Finally, Eq. (71), Eq. (72), Eq. (73), and Eq. (74) allow us to conclude the proof.

□

Corollary C.10. *Let us take the same assumptions of Theorem C.8. Additionally, let us assume that $\gamma^1 = \gamma^2 = \gamma$, the stochastic gradients can be written as $\nabla_x f_\gamma(z) = \nabla_x f(z) + U^x$ and $\nabla_y f_\gamma(z) = \nabla_y f(z) + U^y$ such that U^x and U^y are independent noises that do not depend on z, whose expectation is 0, and whose covariance matrix is Σ. Therefore, the SDE (57) is*

$$dZ_t = -F(Z_t) + \rho \nabla F(Z_t) F(Z_t) dt + \sqrt{\eta} \left(\mathbf{I}_{2d} - \rho \nabla F\left(Z_t\right)\right) \sqrt{\Sigma} dW_t. \tag{75}$$

Proof of Corollary C.10. First of all, we notice that

$$\begin{aligned} F(z) - \rho \mathbb{E}\left[\nabla F_\gamma(z) F_\gamma(z)\right] &= F(z) - \rho \mathbb{E}\left[\nabla F(z) F_\gamma(z)\right] \\ &= F(z) - \rho \nabla F(z) F(z). \end{aligned} \tag{76}$$

Second, based on our assumption of the noise structure, we can rewrite Eq. (60) of the matrix Σ^{SEG} as

$$\begin{aligned} \Sigma^{\text{SEG}} &= \eta^2 \mathbb{E}\left[\left(F(z) - F_\gamma(z)\right)\left(F(z) - F_\gamma(z)\right)^\top\right] \\ &+ \eta^2 \rho \left(\mathbb{E}\left[\left(F\left(z\right) - F_\gamma\left(z\right)\right)\left(\mathbb{E}\left[\nabla F_\gamma(z) F_\gamma(z)\right] - \nabla F_\gamma(z) F_\gamma(z)\right)^\top\right]\right) \\ &+ \eta^2 \rho \left(\mathbb{E}\left[\left(F\left(z\right) - F_\gamma\left(z\right)\right)\left(\mathbb{E}\left[\nabla F_\gamma(z) F_\gamma(z)\right] - \nabla F_\gamma(z) F_\gamma(z)\right)^\top\right]\right)^\top + \mathcal{O}\left(\eta^3\right) \\ &= \eta^2 \mathbb{E}\left[\left(F(z) - F_\gamma(z)\right)\left(F(z) - F_\gamma(z)\right)^\top\right] \\ &+ \eta^2 \rho \left(\mathbb{E}\left[\left(F\left(z\right) - F_\gamma\left(z\right)\right)\left(\mathbb{E}\left[\nabla F(z) F_\gamma(z)\right] - \nabla F(z) F_\gamma(z)\right)^\top\right]\right) \\ &+ \eta^2 \rho \left(\mathbb{E}\left[\left(F\left(z\right) - F_\gamma\left(z\right)\right)\left(\mathbb{E}\left[\nabla F(z) F_\gamma(z)\right] - \nabla F(z) F_\gamma(z)\right)^\top\right]\right)^\top + \mathcal{O}\left(\eta^3\right) \\ &= \eta^2 \left(\Sigma + \rho \Sigma \nabla F\left(z\right)^\top + \rho \nabla F\left(z\right) \Sigma\right) + \mathcal{O}\left(\eta^3\right). \end{aligned} \tag{77}$$

By observing that $(\mathbf{I}_{2d} - \rho \nabla F\left(z\right)) \sqrt{\Sigma} \sqrt{\Sigma} (\mathbf{I}_{2d} - \rho \nabla F\left(z\right)^\top) = \Sigma + \rho \Sigma \nabla F\left(z\right)^\top + \rho \nabla F\left(z\right) \Sigma + \mathcal{O}\left(\eta\right)$, we conclude the proof. □

Corollary C.11. *Let us take the same assumptions of Theorem C.8. Additionally, let us assume that γ^1 and γ^2, are independent and the stochastic gradients can be written as $\nabla_x f_{\gamma^i}(z) = \nabla_x f(z) + U_x^i$ and $\nabla_y f_{\gamma^i}(z) = \nabla_y f(z) + U_y^i$ such that U_x^i and U_y^i are independent noises that do not depend on z, whose expectation is 0, and whose covariance matrix is Σ. Therefore, the SDE (57) is*

$$dZ_t = -F(Z_t) + \rho \nabla F(Z_t) F(Z_t) dt + \sqrt{\eta \Sigma} dW_t. \tag{78}$$

Proof of Corollary C.11. First of all, we notice that

$$F(z) - \rho \mathbb{E}\left[\nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right] = F(z) - \rho \nabla F(z) F(z). \tag{79}$$

Second, based on our assumption of the noise structure, we can rewrite Eq. (60) of the matrix Σ^{SEG} as

$$\begin{aligned} \Sigma^{\text{SEG}} &= \eta^2 \mathbb{E}\left[\left(F(z) - F_{\gamma^1}(z)\right)\left(F(z) - F_{\gamma^1}(z)\right)^\top\right] \\ &+ \eta^2 \rho \left(\mathbb{E}\left[\left(F\left(z\right) - F_{\gamma^1}\left(z\right)\right)\left(\nabla F(z) F(z) - \nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right)^\top\right]\right) \\ &+ \eta^2 \rho \left(\mathbb{E}\left[\left(F\left(z\right) - F_{\gamma^1}\left(z\right)\right)\left(\nabla F(z) F(z) - \nabla F_{\gamma^1}(z) F_{\gamma^2}(z)\right)^\top\right]\right)^\top + \mathcal{O}\left(\eta^3\right) \\ &= \eta^2 \mathbb{E}\left[\left(F(z) - F_{\gamma^1}(z)\right)\left(F(z) - F_{\gamma^1}(z)\right)^\top\right] \\ &+ \eta^2 \rho \left(\mathbb{E}\left[\left(F\left(z\right) - F_{\gamma^1}\left(z\right)\right)\left(\nabla F(z) F(z) - \nabla F(z) F_{\gamma^2}(z)\right)^\top\right]\right) \\ &+ \eta^2 \rho \left(\mathbb{E}\left[\left(F\left(z\right) - F_{\gamma^1}\left(z\right)\right)\left(\nabla F(z) F(z) - \nabla F(z) F_{\gamma^2}(z)\right)^\top\right]\right)^\top + \mathcal{O}\left(\eta^3\right) \\ &= \eta^2 \Sigma + \mathcal{O}\left(\eta^3\right), \end{aligned} \tag{80}$$

which concludes the proof.

□

C.2.1 Continuous-time SEG is equivalent to SGDA if $\rho = \mathcal{O}(\eta)$

In this subsection, we provide a formal proof that if $\rho = \mathcal{O}(\eta)$, the first-order weak approximation of SEG is the same as that of SGDA. This is consistent with the ODE literature on ODEs for these models (Chavdarova et al., 2023; Lu, 2022).

We will consider the stochastic process $Z_t \in \mathbb{R}^{2d}$ defined as the solution of

$$dZ_t = -F\left(Z_t\right) dt + \sqrt{\eta\Sigma} dW_t. \tag{81}$$

Theorem C.12 (Stochastic modified equations)**.** *Let $0 < \eta < 1, T > 0$ and set $N = \lfloor T/\eta \rfloor$. Let $z_k \in \mathbb{R}^{2d}, 0 \leq k \leq N$ denote a sequence of SEG iterations defined by Eq.* (3)*. Additionally, let us take*

$$\rho = \mathcal{O}\left(\eta\right). \tag{82}$$

Consider the stochastic process Z_t defined in Eq. (81) *and fix some test function $g \in G$ and suppose that g and its partial derivatives up to order 6 belong to G.*
Then, under Assumption C.2, there exists a constant $C > 0$ independent of η such that for all $k = 0, 1, \ldots, N$, we have

$$\left|\mathbb{E} g\left(Z_{k\eta}\right) - \mathbb{E} g\left(z_k\right)\right| \leq C\eta.$$

That is, the SDE (81) *is an order* 1 *weak approximation of the SEG iterations* (3)*.*

Lemma C.13. *Under the assumptions of Theorem C.12, let $0 < \eta < 1$ and consider $z_k, k \geq 0$ satisfying the SEG iterations* (3)

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \eta \begin{bmatrix} +\nabla_x f_{\gamma_k^1}\left(x_k - \rho\nabla_x f_{\gamma_k^2}\left(x_k, y_k\right), y_k + \rho\nabla_y f_{\gamma_k^2}\left(x_k, y_k\right)\right) \\ -\nabla_y f_{\gamma_k^1}\left(x_k - \rho\nabla_x f_{\gamma_k^2}\left(x_k, y_k\right), y_k + \rho\nabla_y f_{\gamma_k^2}\left(x_k, y_k\right)\right) \end{bmatrix} \tag{83}$$

with $z_0 := z = (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$. From the definition the one-step difference $\bar{\Delta} = z_1 - z$, then we have

1. $\mathbb{E}\bar{\Delta}_i = -F_i\left(z\right)\eta + \mathcal{O}\left(\eta^2\right) \quad \forall i = 1, \ldots, 2d;$
2. $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = F_i\left(z\right)F_j\left(z\right)\eta^2 + \Sigma_{(ij)}\left(z\right)\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \ldots, 2d;$
3. $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, 2d\}.$

All the functions above are evaluated at z.

Proof of Lemma C.13. First of all, we have that by definition and using a Taylor expansion,

$$\begin{aligned} x_{k+1} &= x_k - \eta\nabla_x f_{\gamma_k^1}\left(x_k - \rho\nabla_x f_{\gamma_k^2}\left(x_k, y_k\right), y_k + \rho\nabla_y f_{\gamma_k^2}\left(x_k, y_k\right)\right) \\ &= x_k - \eta\nabla_x f_{\gamma_k^1}\left(x_k, y_k\right) + \eta\rho\nabla^2_{xx} f_{\gamma_k^1}\left(x_k, y_k\right)\nabla_x f_{\gamma_k^2}\left(x_k, y_k\right) - \eta\rho\nabla^2_{xy} f_{\gamma_k^1}\left(x_k, y_k\right)\nabla_y f_{\gamma_k^2}\left(x_k, y_k\right) + \mathcal{O}\left(\eta\rho^2\right), \end{aligned} \tag{84}$$

and

$$\begin{aligned} y_{k+1} &= y_k + \eta\nabla_y f_{\gamma_k^1}\left(x_k - \rho\nabla_x f_{\gamma_k^2}\left(x_k, y_k\right), y_k + \rho\nabla_y f_{\gamma_k^2}\left(x_k, y_k\right)\right) \\ &= y_k + \eta\nabla_y f_{\gamma_k^1}\left(x_k, y_k\right) - \eta\rho\nabla^2_{xy} f_{\gamma_k^2}\left(x_k, y_k\right)\nabla_x f_{\gamma_k^2}\left(x_k, y_k\right) + \eta\rho\nabla^2_{yy} f_{\gamma_k}\left(x_k, y_k\right)\nabla_y f_{\gamma_k}\left(x_k, y_k\right) + \mathcal{O}\left(\eta\rho^2\right). \end{aligned} \tag{85}$$

Therefore

$$z_1 = z - \eta F_{\gamma^1}(z) + \mathcal{O}\left(\eta^2\right), \tag{86}$$

which implies that

$$\mathbb{E}\left[z_{k+1} - z_k\right] = -\eta F(z_k) + \mathcal{O}\left(\eta^2\right), \tag{87}$$

which in turn implies that

$$\mathbb{E}\bar{\Delta}_i = -F_i\left(z\right)\eta + \mathcal{O}\left(\eta^2\right) \quad \forall i = 1, \ldots, 2d. \tag{88}$$

Second, we have that

$$\mathbb{E}\left[\left(z_1 - z\right)\left(z_1 - z\right)^\top\right] = \eta^2\left[\left(F(z)\right)\left(F(z)\right)^\top\right] + \eta^2\Sigma(z) + \mathcal{O}\left(\eta^3\right), \tag{89}$$

which implies that

$$\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j = F_i\left(z\right)F_j\left(z\right)\eta^2 + \Sigma_{(ij)}\left(z\right)\eta^2 + \mathcal{O}\left(\eta^3\right) \quad \forall i, j = 1, \ldots, 2d. \tag{90}$$

Finally, by definition,

$$\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, 2d\}, \tag{91}$$

which concludes our proof. □

Proof of Theorem C.12. To prove this result, all we need to do is check the conditions in Theorem C.4. As we apply Lemma C.3, we make the following choices:

- $b(z) = -F(z)$;
- $\sigma(z) = \Sigma(z)^{\frac{1}{2}}$.

First of all, we notice that $\forall i = 1, \ldots, 2d$, it holds that

- $\mathbb{E}\bar{\Delta}_i \overset{\text{1. Lemma C.13}}{=} -F_i\left(z\right)\eta + \mathcal{O}\left(\eta^2\right)$;
- $\mathbb{E}\Delta_i \overset{\text{1. Lemma C.3}}{=} -F_i\left(z\right)\eta + \mathcal{O}\left(\eta^2\right)$.

Therefore, we have that for some $K_1(z) \in G$,

$$\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(z)\eta^2, \quad \forall i = 1, \ldots, 2d. \tag{92}$$

Additionally, we notice that $\forall i, j = 1, \ldots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j \overset{\text{2. Lemma C.13}}{=} F_i\left(z\right)F_j\left(z\right)\eta^2 + \Sigma_{(ij)}\left(z\right)\eta^2 + \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\Delta_i\Delta_j \overset{\text{2. Lemma C.3}}{=} F_i\left(z\right)F_j\left(z\right)\eta^2 + \Sigma_{(ij)}\left(z\right)\eta^2 + \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_2(z) \in G$,

$$\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(z)\eta^2, \quad \forall i, j = 1, \ldots, 2d. \tag{93}$$

Additionally, we notice that $\forall s \geq 3, \forall i_j \in \{1, \ldots, 2d\}$, it holds that

- $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} \overset{\text{3. Lemma C.13}}{=} \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} \overset{\text{3. Lemma C.3}}{=} \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_3(z) \in G$

$$\left| \mathbb{E} \prod_{j=1}^{s} \Delta_{i_j} - \mathbb{E} \prod_{j=1}^{s} \bar{\Delta}_{i_j} \right| \leq K_3(z)\eta^2. \tag{94}$$

Additionally, for some $K_4(z) \in G$, $\forall i_j \in \{1, \ldots, d\}$

$$\mathbb{E} \prod_{j=1}^{3} \left| \bar{\Delta}_{(i_j)} \right| \overset{\text{3. Lemma C.13}}{\leq} K_4(z)\eta^2. \tag{95}$$

Finally, Eq. (92), Eq. (93), Eq. (94), and Eq. (95) allow us to conclude the proof.

□

C.3 Formal Derivation - SHGD

In this subsection, we present the first formal derivation of an SDE model for SHGD. We will consider the stochastic process $Z_t \in \mathbb{R}^d$ defined as the solution of

$$dZ_t = -F^{\text{SHGD}}(Z_t)\, dt + \sqrt{\eta \Sigma^{\text{SHGD}}(Z_t)} dW_t. \tag{96}$$

with

$$F^{\text{SHGD}}(z) := \nabla \mathbb{E}\left[\mathcal{H}_{\boldsymbol{\gamma}}(z)\right], \tag{97}$$

$$\Sigma^{\text{SHGD}}(z) := \mathbb{E}\left[\hat{\xi}_{\boldsymbol{\gamma}}(z)\hat{\xi}_{\boldsymbol{\gamma}}(z)^\top\right], \tag{98}$$

$$\hat{\xi}_{\boldsymbol{\gamma}}(z) = F^{\text{SHGD}}(z) - \nabla \mathcal{H}_{\boldsymbol{\gamma}}(z). \tag{99}$$

We remind the following equalities that will come in handy in the subsequent proofs:

$$\Sigma^{\text{SHGD}}(z) = \mathbb{E}\left[\left(\nabla \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] - \nabla \mathcal{H}_{\gamma^1,\gamma^2}(z)\right)\left(\nabla \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] - \nabla \mathcal{H}_{\gamma^1,\gamma^2}(z)\right)^\top\right], \tag{100}$$

$$\mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] = \mathbb{E}\left[\frac{F_{\gamma^1}^\top(z)\, F_{\gamma^2}(z)}{2}\right], \quad \text{and} \quad \mathbb{E}\left[\nabla \mathcal{H}_{\gamma^1,\gamma^2}(z)\right] = \mathbb{E}\left[\frac{F_{\gamma^1}^\top(z)\, \nabla F_{\gamma^2}(z) + F_{\gamma^2}^\top(z)\, \nabla F_{\gamma^1}(z)}{2}\right]. \tag{101}$$

Theorem C.14 (Stochastic modified equations)**.** *Let $0 < \eta < 1, T > 0$ and set $N = \lfloor T/\eta \rfloor$. Let $z_k \in \mathbb{R}^{2d}, 0 \leq k \leq N$ denote a sequence of SHGD iterations defined by Eq.* (4)*. Consider the stochastic process Z_t defined in Eq.* (96) *and fix some test function $g \in G$ and suppose that g and its partial derivatives up to order 6 belong to G.*
Then, under Assumption C.2, there exists a constant $C > 0$ independent of η such that for all $k = 0, 1, \ldots, N$, we have

$$\left|\mathbb{E}g(Z_{k\eta}) - \mathbb{E}g(z_k)\right| \leq C\eta.$$

That is, the SDE (96) *is an order* 1 *weak approximation of the SHGD iterations* (4)*.*

Lemma C.15. *Under the assumptions of Theorem C.14, let $0 < \eta < 1$ and consider $z_k, k \geq 0$ satisfying the SHGD iterations* (4)

$$z_{k+1} = z_k - \eta \nabla \mathcal{H}_{\gamma_k^1 \gamma_k^2}(z_k)$$

with $z_0 := z = (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$. From the definition the one-step difference $\bar{\Delta} = z_1 - z$, then we have

1. $\mathbb{E}\bar{\Delta}_i = -\partial_{e_i} \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}\right] \eta \quad \forall i = 1, \ldots, 2d;$
2. $\mathbb{E}\bar{\Delta}_i \bar{\Delta}_j = \partial_{e_i} \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}\right] \partial_{e_j} \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}\right] \eta^2 + \Sigma_{(ij)}^{SHGD} \eta^2 \quad \forall i, j = 1, \ldots, 2d;$
3. $\mathbb{E}\prod_{j=1}^{s} \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, 2d\}.$

All the functions above are evaluated at z.

Proof of Lemma C.15. First of all, we have that by definition

$$\mathbb{E}\left[z_1 - z\right] = -\eta \nabla \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right], \tag{102}$$

which implies

$$\mathbb{E}\bar{\Delta}_i = -\partial_{e_i} \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] \eta \quad \forall i = 1, \ldots, 2d. \tag{103}$$

Second, we have that by definition

$$\begin{aligned}
\mathbb{E}\left[(z_1 - z)(z_1 - z)^\top\right] &= \eta^2 \left[\nabla \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] \nabla \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right]^\top\right] \\
&\quad + \eta^2 \mathbb{E}\left[\left(\nabla \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] - \nabla \mathcal{H}_{\gamma^1,\gamma^2}(z)\right)\left(\nabla \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] - \nabla \mathcal{H}_{\gamma^1,\gamma^2}(z)\right)^\top\right] \\
&= \eta^2 \left[\nabla \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] \nabla \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right]^\top\right] + \eta^2 \Sigma^{\mathrm{SHGD}}(z),
\end{aligned} \tag{104}$$

which implies that

$$\mathbb{E}\bar{\Delta}_i \bar{\Delta}_j = \partial_{e_i} \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] \partial_{e_j} \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] \eta^2 + \Sigma_{(ij)}^{\mathrm{SHGD}}(z) \eta^2 \quad \forall i, j = 1, \ldots, 2d. \tag{105}$$

Finally, by definition

$$\mathbb{E}\prod_{j=1}^{s} \bar{\Delta}_{i_j} = \mathcal{O}\left(\eta^3\right) \quad \forall s \geq 3, \quad i_j \in \{1, \ldots, 2d\}, \tag{106}$$

which concludes our proof. □

Proof of Theorem C.14. To prove this result, all we need to do is check the conditions in Theorem C.4. As we apply Lemma C.3, we make the following choices:

- $b(z) = -\nabla \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right]$;
- $\sigma(z) = \Sigma^{\mathrm{SHGD}}(z)^{\frac{1}{2}}$.

First of all, we notice that $\forall i = 1, \ldots, 2d$, it holds that

- $\mathbb{E}\bar{\Delta}_i \overset{\text{1. Lemma C.15}}{=} -\partial_{e_i} \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] \eta;$
- $\mathbb{E}\Delta_i \overset{\text{1. Lemma C.3}}{=} -\partial_{e_i} \mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right] \eta + \mathcal{O}\left(\eta^2\right).$

Therefore, we have that for some $K_1(z) \in G$,

$$\left|\mathbb{E}\Delta_i - \mathbb{E}\bar{\Delta}_i\right| \leq K_1(z) \eta^2, \quad \forall i = 1, \ldots, 2d. \tag{107}$$

Additionally, we notice that $\forall i, j = 1, \ldots, d$, it holds that

- $\mathbb{E}\bar{\Delta}_i\bar{\Delta}_j \overset{\text{2. Lemma C.15}}{=} \partial_{e_i}\mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right]\partial_{e_j}\mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right]\eta^2 + \Sigma^{\text{SHGD}}_{(ij)}(z)\,\eta^2$;
- $\mathbb{E}\Delta_i\Delta_j \overset{\text{2. Lemma C.3}}{=} \partial_{e_i}\mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right]\partial_{e_j}\mathbb{E}\left[\mathcal{H}_{\gamma^1,\gamma^2}(z)\right]\eta^2 + \Sigma^{\text{SHGD}}_{(ij)}(z)\,\eta^2 + \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_2(z) \in G$,

$$\left|\mathbb{E}\Delta_i\Delta_j - \mathbb{E}\bar{\Delta}_i\bar{\Delta}_j\right| \leq K_2(z)\eta^2, \quad \forall i,j = 1,\dots,2d. \tag{108}$$

Additionally, we notice that $\forall s \geq 3, \forall i_j \in \{1,\dots,2d\}$, it holds that

- $\mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j} \overset{\text{3. Lemma C.15}}{=} \mathcal{O}\left(\eta^3\right)$;
- $\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} \overset{\text{3. Lemma C.3}}{=} \mathcal{O}\left(\eta^3\right)$.

Therefore, we have that for some $K_3(z) \in G$,

$$\left|\mathbb{E}\prod_{j=1}^{s}\Delta_{i_j} - \mathbb{E}\prod_{j=1}^{s}\bar{\Delta}_{i_j}\right| \leq K_3(z)\eta^2. \tag{109}$$

Additionally, for some $K_4(z) \in G$, $\forall i_j \in \{1,\dots,d\}$,

$$\mathbb{E}\prod_{j=1}^{3}\left|\bar{\Delta}_{(i_j)}\right| \overset{\text{3. Lemma C.15}}{\leq} K_4(z)\eta^2. \tag{110}$$

Finally, Eq. (107), Eq. (108), Eq. (109), and Eq. (110) allow us to conclude the proof.

□

Corollary C.16. *Under the assumptions of Theorem C.14. Additionally, let us assume that $\gamma^1 = \gamma^2 = \gamma$, the stochastic gradients are $\nabla_x f_\gamma(z) = \nabla_x f(z) + U^x$ and $\nabla_y f_\gamma(z) = \nabla_y f(z) + U^y$ such that U^x and U^y are independent noises that do not depend on z, whose expectation is 0, and whose covariance matrix is Σ. Therefore, the SDE is:*

$$dZ_t = -\nabla\mathcal{H}(Z_t)\,dt + \sqrt{\eta}\nabla^2 f(Z_t)\sqrt{\Sigma}dW_t. \tag{111}$$

Proof of Corollary C.16. First of all, we notice that

$$\begin{aligned}\mathbb{E}\left[\mathcal{H}_\gamma(Z_t)\right] &= \mathbb{E}\left[\frac{\|\nabla_x f_\gamma(Z_t)\|_2^2 + \|\nabla_y f_\gamma(Z_t)\|_2^2}{2}\right] = \mathbb{E}\left[\frac{\|\nabla_x f(Z_t)\|_2^2 + \|\nabla_y f(Z_t)\|_2^2}{2}\right] \\ &+ \frac{\mathbb{E}\left[(U^x)(U^x)^\top\right] + \mathbb{E}\left[(U^y)(U^y)^\top\right]}{2} = \mathcal{H}(Z_t) + \frac{\mathbb{E}\left[(U^x)(U^x)^\top\right] + \mathbb{E}\left[(U^y)(U^y)^\top\right]}{2}.\end{aligned} \tag{112}$$

Since $\frac{\mathbb{E}\left[(U^x)(U^x)^\top\right]+\mathbb{E}\left[(U^y)(U^y)^\top\right]}{2}$ is independent on z, we ignore it as its gradient is 0.

Second, based on our assumption of the noise structure, we can rewrite Eq. (100) of the matrix $\Sigma^{\text{SHGD}}(z)$ as

$$\mathbb{E}\left[\left(F^\top(z)\nabla F(z) - \frac{F^\top_{\gamma^1}(z) + F^\top_{\gamma^2}(z)}{2}\nabla F(z)\right)\left(F^\top(z)\nabla F(z) - \frac{F^\top_{\gamma^1}(z) + F^\top_{\gamma^2}(z)}{2}\nabla F(z)\right)^\top\right]. \tag{113}$$

Since $\gamma^1 = \gamma^2 = \gamma$, and noticing that $F^\top(z)\nabla F(z) = \nabla^2 f(z)\nabla f(z)$, we have

$$\Sigma^{\text{SHGD}}(z) = \nabla^2 f(z)\,\Sigma\nabla^2 f(z), \tag{114}$$

which implies that

$$dZ_t = -\nabla\mathcal{H}(Z_t)\,dt + \sqrt{\eta}\nabla^2 f(Z_t)\sqrt{\Sigma}dW_t. \tag{115}$$

□

Corollary C.17. *Under the assumptions of Theorem C.14. Additionally, let us assume that γ^1 and γ^2, are independent and the stochastic gradients can be written as $\nabla_x f_{\gamma^i}(z) = \nabla_x f(z) + U_x^i$ and $\nabla_y f_{\gamma^i}(z) = \nabla_y f(z) + U_y^i$ such that U_x^i and U_y^i are independent noises that do not depend on z. Therefore, the SDE is:*

$$dZ_t = -\nabla \mathcal{H}\left(Z_t\right) dt + \sqrt{\frac{\eta}{2}} \nabla^2 f\left(Z_t\right) \sqrt{\Sigma} dW_t. \tag{116}$$

Proof of Corollary C.17. First of all, we notice that

$$\mathbb{E}\left[\nabla \mathcal{H}_{\gamma^1,\gamma^2}(z)\right] := \mathbb{E}\left[\frac{F_{\gamma^1}^\top(z) \nabla F_{\gamma^2}(z) + F_{\gamma^2}^\top(z) \nabla F_{\gamma^1}(z)}{2}\right] = F^\top(z) \nabla F(z) = \nabla \mathcal{H}(z). \tag{117}$$

Second, based on our assumption of the noise structure, we can rewrite Eq. (100) of the matrix $\Sigma^{\text{SHGD}}(z)$ as

$$\mathbb{E}\left[\left(F^\top(z) \nabla F(z) - \frac{F_{\gamma^1}^\top(z) + F_{\gamma^2}^\top(z)}{2} \nabla F(z)\right)\left(F^\top(z) \nabla F(z) - \frac{F_{\gamma^1}^\top(z) + F_{\gamma^2}^\top(z)}{2} \nabla F(z)\right)^\top\right]. \tag{118}$$

Since γ^1 and γ^2 are independent, and noticing that $F^\top(z) \nabla F(z) = \nabla^2 f(z) \nabla f(z)$, we have

$$\Sigma^{\text{SHGD}}(z) = \frac{1}{2} \nabla^2 f(z) \Sigma \nabla^2 f(z), \tag{119}$$

which implies that

$$dZ_t = -\nabla \mathcal{H}\left(Z_t\right) dt + \sqrt{\frac{\eta}{2}} \nabla^2 f\left(Z_t\right) \sqrt{\Sigma} dW_t. \tag{120}$$

□

D BILINEAR GAMES - INSIGHTS

In this section, we study Bilinear Games of the form $f(x, y) = x^\top \mathbf{\Lambda} y$, where $\mathbf{\Lambda}$ is a square, diagonal, and positive semidefinite matrix.

D.1 SEG

Theorem D.1 (Exact Dynamics of SEG)**.** *Under the assumptions of Corollary C.10, for $f(x, y) = x^\top \mathbf{\Lambda} y$ and noise covariance matrices equal to $\sigma \mathbf{I}_d$, we have that*

$$Z_t = \tilde{\mathbf{E}}(t) \tilde{\mathbf{R}}(t) \left(z + \sqrt{\eta} \sigma \int_0^t \tilde{\mathbf{E}}(-s) \tilde{\mathbf{R}}(-s) \mathbf{M} dW_s\right), \tag{121}$$

with $\tilde{\mathbf{E}}(t) = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix}, \tilde{\mathbf{R}}(t) = \begin{bmatrix} \mathbf{C}(t) & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{C}(t) \end{bmatrix}$, *and* $\mathbf{M} = \begin{bmatrix} \mathbf{I}_d & -\rho \mathbf{\Lambda} \\ \rho \mathbf{\Lambda} & \mathbf{I}_d \end{bmatrix}$, *where*

$$\mathbf{E}(t) := \operatorname{diag}\left(e^{-\rho \lambda_1^2 t}, \cdots, e^{-\rho \lambda_d^2 t}\right), \tag{122}$$

$$\mathbf{C}(t) := \operatorname{diag}\left(\cos\left(\lambda_1 t\right), \cdots, \cos\left(\lambda_d t\right)\right), \tag{123}$$

and

$$\mathbf{S}(t) := \operatorname{diag}\left(\sin\left(\lambda_1 t\right), \cdots, \sin\left(\lambda_d t\right)\right). \tag{124}$$

In particular, we have that

1. $\mathbb{E}\left[Z_t\right] = \tilde{\mathbf{E}}(t) \tilde{\mathbf{R}}(t) z \overset{t \to \infty}{=} 0$;

2. The covariance matrix of is equal to

$$\eta\sigma^2 \begin{bmatrix} \mathbf{I}_d - \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d - \mathbf{E}(2t) \end{bmatrix} \bar{\Sigma} \overset{t\to\infty}{=} \eta\sigma^2\bar{\Sigma}, \tag{125}$$

where

$$\bar{\Sigma} := \begin{bmatrix} \mathbf{B} & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{B} \end{bmatrix}, \tag{126}$$

and $\mathbf{B} := \operatorname{diag}\left(\frac{1+\rho^2\lambda_1^2}{2\rho\lambda_1^2}, \cdots, \frac{1+\rho^2\lambda_d^2}{2\rho\lambda_d^2}\right)$*;*

3. If $\rho = 0$*, SGDA would indeed diverge.*

Proof. The SDEs of SEG are:

$$dX_t = -\mathbf{\Lambda} Y_t dt - \rho\mathbf{\Lambda}^2 X_t dt + \sqrt{\eta}\mathbf{I}_d\sigma dW_t^x - \sqrt{\eta}\sigma\rho\mathbf{\Lambda} dW_t^y \tag{127}$$

and

$$dY_t = +\mathbf{\Lambda} X_t dt - \rho\mathbf{\Lambda}^2 Y t dt + \sqrt{\eta}\sigma\mathbf{I}_d W_t^y + \sqrt{\eta}\sigma\rho\mathbf{\Lambda} dW_t^x, \tag{128}$$

which can be rewritten as

$$dZ_t = \mathbf{A} Z_t dt + \sqrt{\eta}\sigma\mathbf{B} dW_t, \tag{129}$$

where

$$\mathbf{A} = \begin{bmatrix} -\rho\mathbf{\Lambda}^2 & -\mathbf{\Lambda} \\ \mathbf{\Lambda} & -\rho\mathbf{\Lambda}^2 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{I}_d & -\rho\mathbf{\Lambda} \\ \rho\mathbf{\Lambda} & \mathbf{I}_d \end{bmatrix}. \tag{130}$$

Therefore, the solution is

$$Z_t = e^{\mathbf{A}t}\left(z + \sqrt{\eta}\sigma\int_0^t e^{-\mathbf{A}s}\mathbf{B} dW_s\right). \tag{131}$$

We observe that $\mathbf{A} = \mathbf{A}_1 + \mathbf{A}_2$ s.t.

$$\mathbf{A}_1 = \begin{bmatrix} -\rho\mathbf{\Lambda}^2 & \mathbf{0}_d \\ \mathbf{0}_d & -\rho\mathbf{\Lambda}^2 \end{bmatrix} \quad \text{and} \quad \mathbf{A}_2 = \begin{bmatrix} \mathbf{0}_d & -\mathbf{\Lambda} \\ \mathbf{\Lambda} & \mathbf{0}_d \end{bmatrix}, \tag{132}$$

and that since these two matrix commute, $e^{\mathbf{A}t} = e^{\mathbf{A}_1 t}e^{\mathbf{A}_2 t}$. Clearly, we have that

$$\tilde{\mathbf{E}}(t) := e^{\mathbf{A}_1 t} = \operatorname{diag}\left(e^{-\rho\lambda_1^2 t}, \cdots, e^{-\rho\lambda_d^2 t}, e^{-\rho\lambda_1^2 t}, \cdots, e^{-\rho\lambda_d^2 t}\right) = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix}, \tag{133}$$

where $\mathbf{E}(t) := \operatorname{diag}\left(e^{-\rho\lambda_1^2 t}, \cdots, e^{-\rho\lambda_d^2 t}\right)$.

Regarding $\mathbf{A}_2$, we observe that

$$(\mathbf{A}_2 t)^{2k} = \operatorname{diag}\left((\lambda_1 t)^{2k}(-1)^k, \cdots, (\lambda_d t)^{2k}(-1)^k, (\lambda_1 t)^{2k}(-1)^k, \cdots, (\lambda_d t)^{2k}(-1)^k\right) \tag{134}$$

and that

$$(\mathbf{A}_2 t)^{2k+1} = \begin{bmatrix} \mathbf{0_d} & \mathbf{P} \\ \mathbf{Q} & \mathbf{0_d} \end{bmatrix}, \tag{135}$$

where

$$\mathbf{P} := \operatorname{diag}\left((\lambda_1 t)^{2k+1}(-1)^{k+1}, \cdots, (\lambda_d t)^{2k+1}(-1)^{k+1}\right) \tag{136}$$

and

$$\mathbf{Q} := \operatorname{diag}\left((\lambda_1 t)^{2k+1}(-1)^k, \cdots, (\lambda_d t)^{2k+1}(-1)^k\right). \tag{137}$$

Therefore,

$$\begin{aligned}\tilde{\mathbf{R}}(t) := e^{\mathbf{A}_2 t} &= \sum_{k=0}^{\infty} \frac{(\mathbf{A}_2 t)^{2k}}{(2k)!} + \sum_{k=0}^{\infty} \frac{(\mathbf{A}_2 t)^{2k+1}}{(2k+1)!} \\ &= \begin{bmatrix} \mathbf{C}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{C}(t) \end{bmatrix} + \begin{bmatrix} \mathbf{0}_d & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{0}_d \end{bmatrix} = \begin{bmatrix} \mathbf{C}(t) & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{C}(t) \end{bmatrix}, \end{aligned} \tag{138}$$

where

$$\mathbf{C}(t) := \operatorname{diag}\left(\cos\left(\lambda_1 t\right), \cdots, \cos\left(\lambda_d t\right)\right) \tag{139}$$

and

$$\mathbf{S}(t) := \operatorname{diag}\left(\sin\left(\lambda_1 t\right), \cdots, \sin\left(\lambda_d t\right)\right). \tag{140}$$

Automatically, we get that

$$e^{-\mathbf{A}_1 s} = \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix} \tag{141}$$

and

$$e^{-\mathbf{A}_2 s} = \begin{bmatrix} \mathbf{C}(s) & \mathbf{S}(s) \\ -\mathbf{S}(s) & \mathbf{C}(s) \end{bmatrix}, \tag{142}$$

which imply that

$$\begin{aligned} Z_t = &\begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix} \begin{bmatrix} \mathbf{C}(t) & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{C}(t) \end{bmatrix} \Bigg(z \\ &+ \sqrt{\eta}\sigma \int_0^t \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix} \begin{bmatrix} \mathbf{C}(s) & \mathbf{S}(s) \\ -\mathbf{S}(s) & \mathbf{C}(s) \end{bmatrix} \begin{bmatrix} \mathbf{I}_d & -\rho\mathbf{\Lambda} \\ \rho\mathbf{\Lambda} & \mathbf{I}_d \end{bmatrix} \begin{bmatrix} dW_s^x \\ dW_s^y \end{bmatrix} \Bigg). \end{aligned} \tag{143}$$

To conclude, we have that

$$Z_t = \tilde{\mathbf{E}}(t)\tilde{\mathbf{R}}(t) \left(z + \sqrt{\eta}\sigma \int_0^t \tilde{\mathbf{E}}(-s)\tilde{\mathbf{R}}(-s)\mathbf{M} dW_s \right), \tag{144}$$

where $\mathbf{M} = \begin{bmatrix} \mathbf{I}_d & -\rho\mathbf{\Lambda} \\ \rho\mathbf{\Lambda} & \mathbf{I}_d \end{bmatrix}$.

We observe that since the expected value of the noise terms is 0, we have that

$$\mathbb{E}[Z_t] = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix} \begin{bmatrix} \mathbf{C}(t) & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{C}(t) \end{bmatrix} z. \tag{145}$$

Therefore, the expectation of Z_t converges to 0 exponentially fast, while spiraling around the origin. We observe that larger values of ρ encourage a faster convergence of $\mathbb{E}[Z_t]$ to 0.

Let us now have a look at the covariance matrix of this process:

$Var(Z_t) = \eta\sigma^2 \begin{bmatrix} \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(2t) \end{bmatrix} \tilde{\mathbf{R}}(t) Var(V_t) \tilde{\mathbf{R}}(t)^\top$, where

$$
\begin{aligned}
V_t &:= \int_0^t \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix} \begin{bmatrix} \mathbf{C}(s) & \mathbf{S}(s) \\ -\mathbf{S}(s) & \mathbf{C}(s) \end{bmatrix} \begin{bmatrix} \mathbf{I}_d & -\rho\mathbf{\Lambda} \\ \rho\mathbf{\Lambda} & \mathbf{I}_d \end{bmatrix} \begin{bmatrix} dW_s^x \\ dW_s^y \end{bmatrix} \\
&= \int_0^t \begin{bmatrix} \mathbf{E}(-s)(\mathbf{C}(s)+\rho\mathbf{\Lambda}\mathbf{S}(s)) & \mathbf{E}(-s)(\mathbf{S}(s)-\rho\mathbf{\Lambda}\mathbf{C}(s)) \\ \mathbf{E}(-s)(\rho\mathbf{\Lambda}\mathbf{C}(s)-\mathbf{S}(s)) & \mathbf{E}(-s)(\mathbf{C}(s)+\rho\mathbf{\Lambda}\mathbf{S}(s)) \end{bmatrix} \begin{bmatrix} dW_s^x \\ dW_s^y \end{bmatrix} \\
&= \begin{bmatrix} \int_0^t e^{\rho\lambda_1^2 s}(\cos(\lambda_1 s)+\rho\lambda\sin(\lambda_1 s))dW_s^{x_1} + \int_0^t e^{\rho\lambda_1^2 s}(\sin(\lambda_1 s)-\rho\lambda_1\cos(\lambda_1 s))dW_s^{y_1} \\ \vdots \\ \int_0^t e^{\rho\lambda_d^2 s}(\cos(\lambda_d s)+\rho\lambda\sin(\lambda_d s))dW_s^{x_d} + \int_0^t e^{\rho\lambda_d^2 s}(\sin(\lambda_d s)-\rho\lambda_d\cos(\lambda_d s))dW_s^{y_d} \\ \int_0^t e^{\rho\lambda_1^2 s}(-\sin(\lambda_1 s)+\rho\lambda\cos(\lambda_1 s))dW_s^{x_1} + \int_0^t e^{\rho\lambda_1^2 s}(\cos(\lambda_1 s)+\rho\lambda_1\sin(\lambda_1 s))dW_s^{y_1} \\ \vdots \\ \int_0^t e^{\rho\lambda_d^2 s}(-\sin(\lambda_d s)+\rho\lambda\cos(\lambda_d s))dW_s^{x_d} + \int_0^t e^{\rho\lambda_d^2 s}(\cos(\lambda_d s)+\rho\lambda_d\sin(\lambda_d s))dW_s^{y_d} \end{bmatrix} \\
&=: \begin{bmatrix} a_1^{x_1}(t)+a_2^{y_1}(t) \\ \vdots \\ a_1^{x_d}(t)+a_2^{y_d}(t) \\ a_3^{x_1}(t)+a_4^{y_1}(t) \\ \vdots \\ a_3^{x_d}(t)+a_4^{y_d}(t) \end{bmatrix}.
\end{aligned} \tag{146}
$$

Therefore,

$$
Var(V_t) = \begin{bmatrix} \mathbf{V}^{1,2}(t) & \mathbf{C}^{1,2,3,4}(t) \\ \mathbf{C}^{1,2,3,4}(t) & \mathbf{V}^{3,4}(t) \end{bmatrix}, \tag{147}
$$

such that

$$
\mathbf{V}_{i,i}^{1,2}(t) = Var(a_1^{x_i}(t)) + Var(a_2^{y_i}(t)), \quad \forall i \in \{1,\cdots,d\}, \tag{148}
$$

$$
\mathbf{V}_{i,i}^{3,4}(t) = Var(a_3^{x_i}(t)) + Var(a_4^{y_i}(t)), \quad \forall i \in \{1,\cdots,d\}, \tag{149}
$$

and

$$
\mathbf{C}_{i,i}^{1,2,3,4}(t) = Cov(a_1^{x_i}(t), a_3^{x_i}(t)) + Cov(a_1^{y_i}(t), a_3^{y_i}(t)), \quad \forall i \in \{1,\cdots,d\}. \tag{150}
$$

Using the well-known Itô Isometry:

$$
\mathbb{E}\left[\left(\int_0^t H_s dW_s\right)^2\right] = \mathbb{E}\left[\int_0^t H_s^2 ds\right],
$$

we get that

$$
\begin{aligned}
&Var(a_1^{x_i}(t)) + Var(a_2^{y_i}(t)) = \int_0^t e^{2\rho\lambda_i^2 s}(\cos(\lambda_i s)+\rho\lambda_i\sin(\lambda_i s))^2 ds + \int_0^t e^{2\rho\lambda_i^2 s}(\sin(\lambda_i s)-\rho\lambda_i\cos(\lambda_i s))^2 ds \\
&= \int_0^t e^{2\rho\lambda_i^2 s}\left[(\cos(\lambda_i s)+\rho\lambda_i\sin(\lambda_i s))^2 + (\sin(\lambda_i s)-\rho\lambda_i\cos(\lambda_i s))^2\right] ds \\
&= \int_0^t e^{2\rho\lambda_i^2 s}\left[1+\rho^2\lambda_i^2\right] ds \\
&= \frac{1+\rho^2\lambda_i^2}{2\rho\lambda_i^2}\left(e^{2\rho\lambda_i^2 t}-1\right).
\end{aligned} \tag{151}
$$

We observe that if $\rho = 0$, this quantity is equal to t.

Then, we do a similar calculation:

$$
\begin{aligned}
&Var(a_3^{x_i}(t)) + Var(a_4^{y_i}(t)) = \int_0^t e^{2\rho\lambda_i^2 s}(-\sin(\lambda_i s) + \rho\lambda_i \cos(\lambda_i s))^2 ds + \int_0^t e^{2\rho\lambda_i^2 s}(\cos(\lambda_i s) + \rho\lambda_i \sin(\lambda_i s))^2 ds \\
&= \int_0^t e^{2\rho\lambda_i^2 s}\left[(-\sin(\lambda_i s) + \rho\lambda_i \cos(\lambda_i s))^2 + (\cos(\lambda_i s) + \rho\lambda_i \sin(\lambda_i s))^2\right] ds \\
&= \int_0^t e^{2\rho\lambda_i^2 s}\left[1 + \rho^2\lambda_i^2\right] ds \\
&= \frac{1+\rho^2\lambda_i^2}{2\rho\lambda_i^2}\left(e^{2\rho\lambda_i^2 t} - 1\right).
\end{aligned} \tag{152}
$$

We observe that if $\rho = 0$, also this quantity is equal to t.

Remembering now that

$$
\mathbb{E}\left[\left(\int_0^t X_s dW_s\right)\left(\int_0^t Y_s dW_s\right)\right] = \mathbb{E}\left[\int_0^t X_s Y_s ds\right],
$$

we have that

$$
\begin{aligned}
Cov(a_1^{x_i}(t), a_3^{x_i}(t)) + Cov(a_2^{y_i}(t), a_4^{y_i}(t)) &= \int_0^t e^{2\rho\lambda_i^2 s}(\cos(\lambda_i s) + \rho\lambda_i \sin(\lambda_i s))(-\sin(\lambda_i s) + \rho\lambda_i \cos(\lambda_i s)) ds \\
&+ \int_0^t e^{2\rho\lambda_i^2 s}(\sin(\lambda_i s) - \rho\lambda_i \cos(\lambda_i s))(\cos(\lambda_i s) + \rho\lambda_i \sin(\lambda_i s)) ds = 0.
\end{aligned} \tag{153}
$$

To conclude, the covariance matrix of Z_t is

$$
Var(Z_t) = \eta\sigma^2 \begin{bmatrix} \mathbf{I}_d - \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d - \mathbf{E}(2t) \end{bmatrix} \bar{\Sigma} \overset{t\to\infty}{=} \eta\sigma^2\bar{\Sigma}, \tag{154}
$$

where

$$
\bar{\Sigma} := \begin{bmatrix} \mathbf{B} & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{B} \end{bmatrix} \tag{155}
$$

and $\mathbf{B} := \operatorname{diag}\left(\frac{1+\rho^2\lambda_1^2}{2\rho\lambda_1^2}, \cdots, \frac{1+\rho^2\lambda_d^2}{2\rho\lambda_d^2}\right)$.

Of course, if $\rho = 0$ the covariance matrix is actually $\eta\sigma^2 t\mathbf{I}_d$, meaning that the variance of SGDA diverges. □

Lemma D.2. *Let us define the variance $\mathbf{B}_{i,i}(\rho) = \frac{1+\rho^2\lambda_i^2}{2\rho\lambda_i^2}$ and consider it as a function of ρ. The following hold:*

1. $\lim_{\rho\to 0} \mathbf{B}_{i,i}(\rho) = \infty$;
2. $\lim_{\rho\to\infty} \mathbf{B}_{i,i}(\rho) = \infty$;
3. $\mathbf{B}_{i,i}(\rho)$ *is convex in* ρ;
4. $\rho = \frac{1}{\lambda_i}$ *realizes the minimum and* $\mathbf{B}_{i,i}\left(\frac{1}{\lambda_i}\right) = \frac{1}{\lambda_i}$;
5. *The trace of* $\bar{\Sigma}$ *is minimized by* $\rho = \sqrt{\frac{\sum \frac{1}{\lambda_i^2}}{d}}$.

Proof. The first four points are obvious while we spell out the last one. We observe that the trace of $\mathbf{B}$ is convex as the sum of convex functions and its derivative w.r.t. ρ is

$$
\frac{d}{d\rho}\left(\sum_{i=1}^d \frac{1+\rho^2\lambda_i^2}{2\rho\lambda_i^2}\right) = \sum_{i=1}^d \frac{1}{2} - \frac{1}{2\rho^2\lambda_i^2}, \tag{156}
$$

which implies that the optimal ρ is indeed $\rho = \sqrt{\frac{\sum \frac{1}{\lambda_i^2}}{d}}$. □

Insights - The trade-off in selecting ρ The curvature of the landscape influences the speed of convergence. Indeed, larger values of λ_i, which correspond to stronger interaction, speed up the exponential decay in the expected value of the iterates. Additionally, ρ impacts the convergence speed in expectation as larger values boost such a decay. However, the peculiar way in which the noise and the landscape interact implies that larger values of ρ might actually result in larger asymptotic variance. One observes that both $\rho \to 0$ and $\rho \to \infty$ result in infinite asymptotic variance. On the bright side, $\rho_i = \frac{1}{\lambda_i}$ is the optimal choice to reduce the variance along the i-th dimension. Unfortunately, this could possibly be very small and thus slow down the convergence. Finally, if one can only select a single ρ across all parameters, then one might want to minimize the trace of the covariance matrix using $\rho = \sqrt{\frac{\sum \frac{1}{\lambda_i^2}}{d}}$.

D.2 SHGD

Theorem D.3 (Exact Dynamics of SHGD)**.** *Under the assumptions of Corollary C.16, for $f(x,y) = x^\top \mathbf{\Lambda} y$ and noise covariance matrices equal to $\sigma \mathbf{I}_d$, we have that*

$$Z_t = \tilde{\mathbf{E}}(t)\left(z + \sqrt{\eta}\sigma \int_0^t \tilde{\mathbf{E}}(-s)\mathbf{M} dW_s\right), \tag{157}$$

$\tilde{\mathbf{E}}(t) = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix}$, $\mathbf{M} = \begin{bmatrix} \mathbf{0}_d & \mathbf{\Lambda} \\ \mathbf{\Lambda} & \mathbf{0}_d \end{bmatrix}$, *where*

$$\mathbf{E}(t) := \operatorname{diag}\left(e^{-\lambda_1^2 t}, \cdots, e^{-\lambda_d^2 t}\right). \tag{158}$$

In particular, we have that

1. $\mathbb{E}[Z_t] = \tilde{\mathbf{E}}(t) z \overset{t\to\infty}{=} 0$;
2. *The covariance matrix of Z_t is equal to*

$$\eta\frac{\sigma^2}{2}\begin{bmatrix} \mathbf{I}_d - \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d - \mathbf{E}(2t) \end{bmatrix} \bar{\Sigma} \overset{t\to\infty}{=} \eta\sigma^2\bar{\Sigma}, \tag{159}$$

where

$$\bar{\Sigma} := \begin{bmatrix} \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix}. \tag{160}$$

Proof. The SDEs of SHGD are:

$$dX_t = -\mathbf{\Lambda}^2 X_t dt + \sqrt{\eta}\sigma\mathbf{\Lambda} dW_t^y \tag{161}$$

and

$$dY_t = -\mathbf{\Lambda}^2 Y t dt + \sqrt{\eta}\sigma\mathbf{\Lambda} dW_t^x, \tag{162}$$

which can be rewritten as

$$dZ_t = \mathbf{A} Z_t dt + \sqrt{\eta}\sigma\mathbf{B} dW_t, \tag{163}$$

where

$$\mathbf{A} = \begin{bmatrix} -\mathbf{\Lambda}^2 & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{\Lambda}^2 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{0}_d & \mathbf{\Lambda} \\ \mathbf{\Lambda} & \mathbf{0}_d \end{bmatrix}. \tag{164}$$

Therefore, the solution is

$$Z_t = e^{\mathbf{A}t}\left(z + \sqrt{\eta}\sigma \int_0^t e^{-\mathbf{A}s}\mathbf{B} dW_s\right). \tag{165}$$

Clearly, we have that

$$\tilde{\mathbf{E}}(t) := e^{\mathbf{A}t} = \operatorname{diag}\left(e^{-\lambda_1^2 t}, \cdots, e^{-\lambda_d^2 t}, e^{-\lambda_1^2 t}, \cdots, e^{-\lambda_d^2 t}\right) = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix}, \tag{166}$$

where $\mathbf{E}(t) := \text{diag}\left(e^{-\lambda_1^2 t}, \cdots, e^{-\lambda_d^2 t}\right)$. Automatically, we get that

$$e^{-\mathbf{A}_1 s} = \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix}, \tag{167}$$

which implies that

$$Z_t = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix} \left(z + \sqrt{\eta}\sigma \int_0^t \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix} \begin{bmatrix} \mathbf{0}_d & \mathbf{\Lambda} \\ \mathbf{\Lambda} & \mathbf{0}_d \end{bmatrix} \begin{bmatrix} dW_s^x \\ dW_s^y \end{bmatrix} \right). \tag{168}$$

To conclude, we have that

$$Z_t = \tilde{\mathbf{E}}(t) \left(z + \sqrt{\eta}\sigma \int_0^t \tilde{\mathbf{E}}(-s) \mathbf{M} dW_s \right), \tag{169}$$

where $\mathbf{M} = \begin{bmatrix} \mathbf{0}_d & \mathbf{\Lambda} \\ \mathbf{\Lambda} & \mathbf{0}_d \end{bmatrix}$. We observe that

$$\mathbb{E}[Z_t] = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix} z \tag{170}$$

because the parts dependent on dW are martingales. Therefore, $\mathbb{E}[Z_t]$ converges to 0 exponentially fast.

Let us now have a look at the covariance matrix of this process:

$Var(Z_t) = \eta\sigma^2 \begin{bmatrix} \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(2t) \end{bmatrix} Var(V_t)$, where

$$\begin{aligned} V_t &:= \int_0^t \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix} \begin{bmatrix} \mathbf{0}_d & \mathbf{\Lambda} \\ \mathbf{\Lambda} & \mathbf{0}_d \end{bmatrix} \begin{bmatrix} dW_s^x \\ dW_s^y \end{bmatrix} \\ &= \begin{bmatrix} \int_0^t \lambda_1 e^{\lambda_1^2 s} dW_s^{y_1} \\ \vdots \\ \int_0^t \lambda_d e^{\lambda_d^2 s} dW_s^{y_d} \\ \int_0^t \lambda_1 e^{\lambda_1^2 s} dW_s^{x_1} \\ \vdots \\ \int_0^t \lambda_d e^{\lambda_d^2 s} dW_s^{x_d} \end{bmatrix} =: \begin{bmatrix} a^{y_1}(t) \\ \vdots \\ a^{y_d}(t) \\ a^{x_1}(t) \\ \vdots \\ a^{x_d}(t) \end{bmatrix}. \end{aligned} \tag{171}$$

Therefore,

$$Var(V_t) = \begin{bmatrix} \mathbf{V}^{1,y}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{V}^{1,x}(t) \end{bmatrix}, \tag{172}$$

such that

$$\mathbf{V}_{i,i}^{1,y}(t) = Var(a^{y_i}(t)), \quad \forall i \in \{1, \cdots, d\}, \tag{173}$$

and

$$\mathbf{V}_{i,i}^{1,x}(t) = Var(a^{x_i}(t)), \quad \forall i \in \{1, \cdots, d\}. \tag{174}$$

Using the well-known Itô Isometry

$$\mathbb{E}\left[\left(\int_0^t H_s dW_s \right)^2 \right] = \mathbb{E}\left[\int_0^t H_s^2 ds \right],$$

we get that

$$Var(a^{y_i}(t)) = \int_0^t e^{2\lambda_i^2 s} \lambda_i^2 ds = \frac{1}{2}\left(e^{2\lambda_i^2 t} - 1 \right). \tag{175}$$

Similarly, we get that

$$Var(a^{x_i}(t)) = \int_0^t e^{2\lambda_i^2 s} \lambda_i^2 ds = \frac{1}{2}\left(e^{2\lambda_i^2 t} - 1 \right). \tag{176}$$

Therefore, we conclude that the covariance matrix of Z_t is

$$Var(Z_t) = \eta\sigma^2 \begin{bmatrix} \mathbf{I}_d - \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d - \mathbf{E}(2t) \end{bmatrix} \bar{\Sigma} \overset{t\to\infty}{=} \frac{\eta\sigma^2}{2}\bar{\Sigma}, \tag{177}$$

where

$$\bar{\Sigma} := \begin{bmatrix} \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix}. \tag{178}$$

□

Insights Just like for SEG, the curvature influences the speed of convergence of the algorithm. On the other, the asymptotic variance is independent of the curvature.

SEG vs SHGD We notice that if $\rho = 1$, the exponential decay of SEG and SHGD is the same. However, in such a case, the asymptotic variance of SEG along the i-th dimension is $\eta\sigma^2\left(\frac{1}{2} + \frac{1}{2\lambda_i}\right)$ which is larger than that of SHGD which attains $\frac{\eta\sigma^2}{2}$. Differently, one can select $\rho_i^{\mathrm{V}} = \frac{1}{\lambda_i}$, which realizes the minimum variance of SEG along the i-th dimension. Thus, the resulting variance is $\frac{\eta\sigma^2}{2\lambda_i}$ which is smaller than $\frac{\eta\sigma^2}{2}$ if and only if $\lambda_i > 1$. In such a case, SEG can be more optimal than SHGD, but since $\rho_i = \frac{1}{\lambda_i} < 1$, it will converge more slowly than SHGD.

Therefore, selecting the size of ρ or ρ_i leads to a trade-off between the speed of convergence and the asymptotic variance. To conclude, there is no clear winner between the two methods as their performance depends on the curvature of the landscape.

E QUADRATIC GAMES - INSIGHTS

In this section, we study Quadratic Games of the form $f(x,y) = \frac{x^\top \mathbf{A} x}{2} + x^\top \mathbf{\Lambda} y - \frac{y^\top \mathbf{A} y}{2}$, where $\mathbf{\Lambda}$ and $\mathbf{A}$ are square, diagonal and positive semidefinite matrices. We notice that if $\mathbf{A} = \mathbf{0}$, these are classic Bilinear Games.

E.1 SEG

Theorem E.1 (Exact Dynamics of SEG)**.** *Under the assumptions of Corollary C.10, for $f(x,y) = \frac{x^\top \mathbf{A} x}{2} + x^\top \mathbf{\Lambda} y - \frac{y^\top \mathbf{A} y}{2}$ and noise covariance matrices equal to $\sigma\mathbf{I}_d$, we have that*

$$Z_t = \tilde{\mathbf{E}}(t)\tilde{\mathbf{R}}(t)\left(z + \sqrt{\eta}\sigma\int_0^t \tilde{\mathbf{E}}(-s)\tilde{\mathbf{R}}(-s)\mathbf{M}dW_s\right), \tag{179}$$

$\tilde{\mathbf{E}}(t) = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix}, \tilde{\mathbf{R}}(t) = \begin{bmatrix} \mathbf{C}(t) & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{C}(t) \end{bmatrix}$, *and* $M = \begin{bmatrix} \mathbf{I}_d - \rho\mathbf{A} & -\rho\mathbf{\Lambda} \\ \rho\mathbf{\Lambda} & \mathbf{I}_d - \rho\mathbf{A} \end{bmatrix}$, *where*

$$\mathbf{E}(t) := \mathrm{diag}\left(e^{\rho\left(a_1^2-\lambda_1^2\right)t - a_1 t}, \cdots, e^{\rho\left(a_d^2-\lambda_d^2\right)t - a_d t}\right), \tag{180}$$

$$\mathbf{C}(t) := \mathrm{diag}\left(\cos\left(\hat{\lambda}_1 t\right), \cdots, \cos\left(\hat{\lambda}_d t\right)\right), \tag{181}$$

$$\mathbf{S}(t) := \mathrm{diag}\left(\sin\left(\hat{\lambda}_d t\right), \cdots, \sin\left(\hat{\lambda}_d t\right)\right), \tag{182}$$

and $\hat{\lambda}_i := \lambda_i(1 - 2\rho a_i)$. In particular, if $\rho\left(a_i^2 - \lambda_1^2\right) - a_i < 0$:

1. $\mathbb{E}\left[Z_t\right] = \tilde{\mathbf{E}}(t)\tilde{\mathbf{R}}(t)z \overset{t\to\infty}{=} 0$;
2. *The covariance matrix of Z_t is equal to*

$$\eta\sigma^2 \begin{bmatrix} \mathbf{I}_d - \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d - \mathbf{E}(2t) \end{bmatrix} \bar{\Sigma} \overset{t\to\infty}{=} \eta\sigma^2\bar{\Sigma}, \tag{183}$$

where

$$\bar{\Sigma} := \begin{bmatrix} \mathbf{B} & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{B} \end{bmatrix}, \tag{184}$$

and $\mathbf{B} := \operatorname{diag}\left(\frac{(1-\rho a_1)^2+\rho^2\lambda_1^2}{2(a_1+\rho(\lambda_1^2-a_1^2))}, \cdots, \frac{(1-\rho a_d)^2+\rho^2\lambda_d^2}{2(a_d+\rho(\lambda_d^2-a_d^2))}\right)$*;*

3. *If* $\rho = 0$*, SGDA would indeed always converge.*

Proof. The SDE is

$$dZ_t = \mathbf{D}Z_t dt + \sqrt{\eta}\sigma\mathbf{B}dW_t, \tag{185}$$

where

$$\mathbf{D} = \begin{bmatrix} \rho(\mathbf{A}^2-\mathbf{\Lambda}^2)-\mathbf{A} & -\Lambda(\mathbf{I}_d-2\rho\mathbf{A}) \\ \mathbf{\Lambda}(\mathbf{I}_d-2\rho\mathbf{A}) & \rho(\mathbf{A}^2-\mathbf{\Lambda}^2)-\mathbf{A} \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{I}_d-\rho\mathbf{A} & -\rho\mathbf{\Lambda} \\ \rho\mathbf{\Lambda} & \mathbf{I}_d-\rho\mathbf{A} \end{bmatrix}. \tag{186}$$

Therefore, the solution is

$$Z_t = e^{\mathbf{D}t}\left(z + \sqrt{\eta}\sigma\int_0^t e^{-\mathbf{D}s}\mathbf{B}dW_s\right). \tag{187}$$

We observe that $\mathbf{D} = \mathbf{D_1} + \mathbf{D_2}$ s.t.

$$\mathbf{D}_1 = \begin{bmatrix} \rho(\mathbf{A}^2-\mathbf{\Lambda}^2)-\mathbf{A} & \mathbf{0}_d \\ \mathbf{0}_d & \rho(\mathbf{A}^2-\mathbf{\Lambda}^2)-\mathbf{A} \end{bmatrix} \quad \text{and} \quad \mathbf{D_2} = \begin{bmatrix} \mathbf{0}_d & -\Lambda(\mathbf{I}_d-2\rho\mathbf{A}) \\ \mathbf{\Lambda}(\mathbf{I}_d-2\rho\mathbf{A}) & \mathbf{0}_d \end{bmatrix} \tag{188}$$

and that since these two matrix commute, $e^{\mathbf{D}t} = e^{\mathbf{D}_1 t}e^{\mathbf{D}_2 t}$. Clearly, we have that

$$\tilde{\mathbf{E}}(t) := e^{\mathbf{D}_1 t} = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix}, \tag{189}$$

where $\mathbf{E}(t) := \operatorname{diag}\left(e^{\rho(a_1^2-\lambda_1^2)t-a_1 t}, \cdots, e^{\rho(a_d^2-\lambda_1^2)t-a_d t}\right)$.

Regarding $\mathbf{D}_2$, we observe that $(\mathbf{D}_2 t)^{2k}$ is equal to

$$\operatorname{diag}\left((\lambda_1(1-2\rho a_1)t)^{2k}(-1)^k, \cdots, (\lambda_d(1-2\rho a_d)t)^{2k}(-1)^k, (\lambda_1(1-2\rho a_1)t)^{2k}(-1)^k, \cdots, (\lambda_d(1-2\rho a_d)t)^{2k}(-1)^k\right) \tag{190}$$

and

$$(\mathbf{D}_2 t)^{2k+1} = \begin{bmatrix} \mathbf{0_d} & \mathbf{P} \\ \mathbf{Q} & \mathbf{0_d} \end{bmatrix}, \tag{191}$$

with

$$\mathbf{P} := \operatorname{diag}\left((\lambda_1(1-2\rho a_1)t)^{2k+1}(-1)^{k+1}, \cdots, (\lambda_d(1-2\rho a_d)t)^{2k+1}(-1)^{k+1}\right) \tag{192}$$

and

$$\mathbf{Q} := \operatorname{diag}\left((\lambda_1(1-2\rho a_1)t)^{2k+1}(-1)^{k}, \cdots, (\lambda_d(1-2\rho a_d)t)^{2k+1}(-1)^{k}\right). \tag{193}$$

Therefore,

$$\begin{aligned} \tilde{\mathbf{R}}(t) := e^{\mathbf{D}_2 t} &= \sum_{k=0}^{\infty}\frac{(\mathbf{D}_2 t)^{2k}}{(2k)!} + \sum_{k=0}^{\infty}\frac{(\mathbf{D}_2 t)^{2k+1}}{(2k+1)!} \\ &= \begin{bmatrix} \mathbf{C}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{C}(t) \end{bmatrix} + \begin{bmatrix} \mathbf{0}_d & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{0}_d \end{bmatrix} = \begin{bmatrix} \mathbf{C}(t) & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{C}(t) \end{bmatrix}, \end{aligned} \tag{194}$$

where

$$\mathbf{C}(t) := \operatorname{diag}\left(\cos\left(\lambda_1(1-2\rho a_1)t\right), \cdots, \cos\left(\lambda_d(1-2\rho a_d)t\right)\right), \tag{195}$$

and

$$\mathbf{S}(t) := \operatorname{diag}\left(\sin\left(\lambda_1(1-2\rho a_1)t\right), \cdots, \sin\left(\lambda_d(1-2\rho a_d)t\right)\right). \tag{196}$$

Automatically, we get that

$$e^{-\mathbf{D}_1 s} = \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix} \tag{197}$$

and

$$e^{-\mathbf{D}_2 s} = \begin{bmatrix} \mathbf{C}(s) & \mathbf{S}(s) \\ -\mathbf{S}(s) & \mathbf{C}(s) \end{bmatrix}, \tag{198}$$

which implies that

$$\begin{aligned} Z_t = &\begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix} \begin{bmatrix} \mathbf{C}(t) & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{C}(t) \end{bmatrix} \Bigg(z \\ &+ \sqrt{\eta}\sigma \int_0^t \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix} \begin{bmatrix} \mathbf{C}(s) & \mathbf{S}(s) \\ -\mathbf{S}(s) & \mathbf{C}(s) \end{bmatrix} \begin{bmatrix} \mathbf{I}_d - \rho\mathbf{A} & -\rho\mathbf{\Lambda} \\ \rho\mathbf{\Lambda} & \mathbf{I}_d - \rho\mathbf{A} \end{bmatrix} \begin{bmatrix} dW_s^x \\ dW_s^y \end{bmatrix} \Bigg). \end{aligned} \tag{199}$$

We observe that since the expected value of the noise terms is 0, we have

$$\mathbb{E}[Z_t] = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix} \begin{bmatrix} \mathbf{C}(t) & -\mathbf{S}(t) \\ \mathbf{S}(t) & \mathbf{C}(t) \end{bmatrix} z. \tag{200}$$

Therefore if $\rho(a_i^2 - \lambda_i^2) - a_i < 0$, $\mathbb{E}[Z_t]$ converges to 0 exponentially fast, while spiraling around the origin.

Let us now have a look at the variance of this process:

$Var(Z_t) = \eta\sigma^2 \begin{bmatrix} \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(2t) \end{bmatrix} \tilde{\mathbf{R}}(t) Var(V_t) \tilde{\mathbf{R}}(t)^\top$, where

$$\begin{aligned} V_t &:= \int_0^t \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix} \begin{bmatrix} \mathbf{C}(s) & \mathbf{S}(s) \\ -\mathbf{S}(s) & \mathbf{C}(s) \end{bmatrix} \begin{bmatrix} \mathbf{I}_d - \rho\mathbf{A} & -\rho\mathbf{\Lambda} \\ \rho\mathbf{\Lambda} & \mathbf{I}_d - \rho\mathbf{A} \end{bmatrix} \begin{bmatrix} dW_s^x \\ dW_s^y \end{bmatrix} \\ &= \int_0^t \begin{bmatrix} \mathbf{E}(-s)(\mathbf{C}(s)(\mathbf{I}_d - \rho\mathbf{A}) + \rho\mathbf{\Lambda}\mathbf{S}(s)) & \mathbf{E}(-s)(\mathbf{S}(s)(\mathbf{I}_d - \rho\mathbf{A}) - \rho\mathbf{\Lambda}\mathbf{C}(s)) \\ \mathbf{E}(-s)(\rho\mathbf{\Lambda}\mathbf{C}(s) - \mathbf{S}(s)(\mathbf{I}_d - \rho\mathbf{A})) & \mathbf{E}(-s)(\mathbf{C}(s)(\mathbf{I}_d - \rho\mathbf{A}) + \rho\mathbf{\Lambda}\mathbf{S}(s)) \end{bmatrix} \begin{bmatrix} dW_s^x \\ dW_s^y \end{bmatrix} \\ &= \begin{bmatrix} \int_0^t e^{a_1 s - \rho(a_1^2 - \lambda_1^2)s}(\cos(\hat{\lambda}_1 s)(1 - \rho a_1) + \rho\lambda_1 \sin(\hat{\lambda}_1 s))dW_s^{x_1} + \int_0^t e^{a_1 s - \rho(a_1^2 - \lambda_1^2)s}(\sin(\hat{\lambda}_1 s)(1 - \rho a_1) - \rho\lambda_1 \cos(\hat{\lambda}_1 s))dW_s^{y_1} \\ \vdots \\ \int_0^t e^{a_d s - \rho(a_d^2 - \lambda_d^2)s}(\cos(\hat{\lambda}_d s)(1 - \rho a_d) + \rho\lambda_d \sin(\hat{\lambda}_d s))dW_s^{x_d} + \int_0^t e^{a_d s - \rho(a_d^2 - \lambda_d^2)s}(\sin(\hat{\lambda}_d s)(1 - \rho a_d) - \rho\lambda_d \cos(\hat{\lambda}_d s))dW_s^{y_d} \\ \int_0^t e^{a_1 s - \rho(a_1^2 - \lambda_1^2)s}(\rho\lambda_1 \cos(\hat{\lambda}_1 s) - \sin(\hat{\lambda}_1 s)(1 - \rho a_1))dW_s^{x_1} + \int_0^t e^{a_1 s - \rho(a_1^2 - \lambda_1^2)s}(\cos(\hat{\lambda}_1 s)(1 - \rho a_1) + \rho\lambda_1 \sin(\hat{\lambda}_1 s))dW_s^{y_1} \\ \vdots \\ \int_0^t e^{a_d s - \rho(a_d^2 - \lambda_1^2)s}(\rho\lambda_d \cos(\hat{\lambda}_d s) - \sin(\hat{\lambda}_d s)(1 - \rho a_d))dW_s^{x_d} + \int_0^t e^{a_d s - \rho(a_d^2 - \lambda_1^2)s}(\cos(\hat{\lambda}_d s)(1 - \rho a_d) + \rho\lambda_d \sin(\hat{\lambda}_d s))dW_s^{y_d} \end{bmatrix} \\ &=: \begin{bmatrix} a_1^{x_1}(t) + a_2^{y_1}(t) \\ \vdots \\ a_1^{x_d}(t) + a_2^{y_d}(t) \\ a_3^{x_1}(t) + a_4^{y_1}(t) \\ \vdots \\ a_3^{x_d}(t) + a_4^{y_d}(t) \end{bmatrix} \text{ and } \hat{\lambda}_i := \lambda_i(1 - 2\rho a_i). \end{aligned} \tag{201}$$

Therefore,

$$Var(V_t) = \begin{bmatrix} \mathbf{V}^{1,2}(t) & \mathbf{C}^{1,2,3,4}(t) \\ \mathbf{C}^{1,2,3,4}(t) & \mathbf{V}^{3,4}(t) \end{bmatrix}, \tag{202}$$

such that

$$\mathbf{V}_{i,i}^{1,2}(t) = Var(a_1^{x_i}(t)) + Var(a_2^{y_i}(t)), \quad \forall i \in \{1, \cdots, d\}, \tag{203}$$

$$\mathbf{V}_{i,i}^{3,4}(t) = Var(a_3^{x_i}(t)) + Var(a_4^{y_i}(t)), \quad \forall i \in \{1, \cdots, d\}, \tag{204}$$

and

$$\mathbf{C}_{i,i}^{1,2,3,4}(t) = Cov(a_1^{x_i}(t), a_3^{x_i}(t)) + Cov(a_1^{y_i}(t), a_3^{y_i}(t)), \quad \forall i \in \{1, \cdots, d\}. \tag{205}$$

Using the well-known Itô Isometry,

$$\mathbb{E}\left[\left(\int_0^t H_s dW_s\right)^2\right] = \mathbb{E}\left[\int_0^t H_s^2 ds\right],$$

we get that

$$\begin{aligned} Var(a_1^{x_i}(t)) + Var(a_2^{y_i}(t)) &= \int_0^t e^{2(a_i - \rho(a_i^2-\lambda_i^2))s}(\cos(\hat{\lambda}_i s)(1-\rho a_i) + \rho\lambda_i \sin(\hat{\lambda}_i s))^2 ds \\ &+ \int_0^t e^{2(a_i - \rho(a_i^2-\lambda_i^2))s}(\sin(\hat{\lambda}_i s)(1-\rho a_i) - \rho\lambda_i \cos(\hat{\lambda}_i s))^2 ds \\ &= \int_0^t e^{2(a_i - \rho(a_i^2-\lambda_i^2))s}\left[(1-\rho a_i)^2 + \rho^2\lambda_i^2\right] ds \\ &= \frac{(1-\rho a_i)^2 + \rho^2\lambda_i^2}{2(a_i + \rho(\lambda_i^2 - a_i^2))}\left(e^{2(a_i+\rho(\lambda_i^2-a_i^2))t} - 1\right). \end{aligned} \tag{206}$$

Then, we do a similar calculation and find that

$$Var(a_3^{x_i}(t)) + Var(a_4^{y_i}(t)) = \frac{(1-\rho a_i)^2 + \rho^2\lambda_i^2}{2(a_i + \rho(\lambda_i^2 - a_i^2))}\left(e^{2(a_i+\rho(\lambda_i^2-a_i^2))t} - 1\right). \tag{207}$$

Remembering now that

$$\mathbb{E}\left[\left(\int_0^t X_s dW_s\right)\left(\int_0^t Y_s dW_s\right)\right] = \mathbb{E}\left[\int_0^t X_s Y_s ds\right],$$

we have that

$$\begin{aligned} Cov(a_1^{x_i}(t), a_3^{x_i}(t)) + Cov(a_2^{y_i}(t), a_4^{y_i}(t)) &= \int_0^t e^{2\rho\lambda_i^2 s}(\cos(\lambda_i s) + \rho\lambda_i \sin(\lambda_i s))(-\sin(\lambda_i s) + \rho\lambda_i \cos(\lambda_i s)) ds \\ &+ \int_0^t e^{2\rho\lambda_i^2 s}(\sin(\lambda_i s) - \rho\lambda_i \cos(\lambda_i s))(\cos(\lambda_i s) + \rho\lambda_i \sin(\lambda_i s)) ds = 0. \end{aligned} \tag{208}$$

Therefore, we conclude that the covariance matrix of Z_t is

$$Var(Z_t) = \eta\sigma^2 \begin{bmatrix} \mathbf{I}_d - \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d - \mathbf{E}(2t) \end{bmatrix} \bar{\Sigma} \overset{t\to\infty}{=} \eta\sigma^2\bar{\Sigma}, \tag{209}$$

with

$$\bar{\Sigma} := \begin{bmatrix} \mathbf{B} & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{B} \end{bmatrix}, \tag{210}$$

where

$$\mathbf{B} := \operatorname{diag}\left(\frac{(1-\rho a_1)^2 + \rho^2\lambda_1^2}{2(a_1 + \rho(\lambda_1^2 - a_1^2))}, \dots, \frac{(1-\rho a_d)^2 + \rho^2\lambda_d^2}{2(a_d + \rho(\lambda_d^2 - a_d^2))}\right). \tag{211}$$

Empirical validation of Eq. (211) is provided in Figure 8.

□

Lemma E.2. *Let us define the variance* $\mathbf{B}_{i,i}(\rho) = \frac{(1-\rho a_i)^2+\rho^2\lambda_i^2}{2(a_i+\rho(\lambda_i^2-a_i^2))}$ *and consider it as a function of* ρ*. The following hold:*

1. $\rho_i\left(a_i^2 - \lambda_i^2\right) < 0$ *is necessary to converge faster than SGDA;*
2. *If* $\lambda_i > a_i$ *and* $\lim_{\rho\to\infty} \mathbf{B}_{i,i}(\rho) = \infty$*;*
3. *If* $\lambda_i < a_i$ *and* $\lim_{\rho\to-\infty} \mathbf{B}_{i,i}(\rho) = \infty$*;*
4. $\lim_{\rho\to\frac{-a_i}{\lambda_i^2-a_i^2}} \mathbf{B}_{i,i}(\rho) = \infty$*;*

5. $\rho = \frac{1}{\lambda_i + a_i}$ *realizes the minimum of* $\mathbf{B}_{i,i}$ *and* $\mathbf{B}_{i,i}\left(\frac{1}{\lambda_i+a_i}\right) = \frac{\eta\sigma^2}{2}\frac{\lambda_i}{(a_i+\lambda_i)^2}$;

6. *The trace of* $\mathbf{B}$ *is is strictly convex in* ρ*, meaning that there is a unique minimizer.*

Proof. All points above can be proven easily and are left an exercise for the reader. □

Insights - The trade-off in selecting ρ The curvature of the landscape influences the speed of convergence. Indeed, larger values of a_i, which correspond to stronger convexity/concavity, speed up the exponential decay in the expected value of the iterates. Differently, the relative size of λ_i and a_i influences the convergence depending on the sign of ρ. First of all, if $\rho_i\left(a_i^2 - \lambda_i^2\right) > 0$, SEG is slower than SGDA at converging and $\rho_i\left(a_i^2 - \lambda_i^2\right) < 0$ is necessary to converge faster than SGDA. This means that **negative** ρ_i might be convenient if $a_i > \lambda_i$. Therefore, if ρ_i has the correct sign, a **larger** absolute value implies **faster convergence**. However, we also have that the asymptotic variance along the i-th coordinate $\mathbf{B}_{i,i}(\rho_i)$ **explodes** if $|\rho_i|$ is too **large** or if $\rho_i \to \frac{-a_i}{\lambda_i^2 - a_i^2}$. On the bright side, $B_{i,i}(\rho_i)$ is a convex function of ρ_i whose minimum is realized at $\rho_i^{\mathrm{V}} = \frac{1}{a_i+\lambda_i}$. However, if ρ_i^{V} is **small**, it **slows down** the convergence. Finally, if one has to choose a single value of ρ, one has to carefully select it as it will (de)accelerate different coordinates based on its sign. Fortunately, the trace of $\mathbf{B}$ is a convex function of ρ, meaning that there is an optimal ρ^* that minimizes it.

E.2 SHGD

Theorem E.3 (Exact Dynamics of SHGD)**.** *Under the assumptions of Corollary C.16, for* $f(x,y) = \frac{x^\top \mathbf{A} x}{2} + x^\top \mathbf{\Lambda} y - \frac{y^\top \mathbf{A} y}{2}$ *and noise covariance matrices equal to* $\sigma\mathbf{I}_d$*, we have that*

$$Z_t = \tilde{\mathbf{E}}(t)\left(z + \sqrt{\eta}\sigma\int_0^t \tilde{\mathbf{E}}(-s)\mathbf{M}dW_s\right), \tag{212}$$

with $\tilde{\mathbf{E}}(t) = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix}$, $M = \begin{bmatrix} \mathbf{A} & \mathbf{\Lambda} \\ \mathbf{\Lambda} & -\mathbf{A} \end{bmatrix}$, *where*

$$\mathbf{E}(t) := \mathrm{diag}\left(e^{-\left(\lambda_1^2+a_1^2\right)t}, \cdots, e^{-\left(\lambda_d^2+a_d^2\right)t}\right). \tag{213}$$

In particular, we have that

1. $\mathbb{E}\left[Z_t\right] = \tilde{\mathbf{E}}(t)z \overset{t\to\infty}{=} 0$;

2. *The covariance matrix of* Z_t *is equal to*

$$\eta\frac{\sigma^2}{2}\begin{bmatrix} \mathbf{I}_d - \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d - \mathbf{E}(2t) \end{bmatrix}\bar{\Sigma} \overset{t\to\infty}{=} \eta\sigma^2\bar{\Sigma}, \tag{214}$$

where

$$\bar{\Sigma} := \begin{bmatrix} \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix}. \tag{215}$$

Proof. The SDE is

$$dZ_t = \mathbf{D}Z_t dt + \sqrt{\eta}\sigma\mathbf{B}dW_t, \tag{216}$$

where

$$\mathbf{D} = \begin{bmatrix} -(\mathbf{\Lambda}^2 + \mathbf{A}^2) & \mathbf{0}_d \\ \mathbf{0}_d & -(\mathbf{\Lambda}^2 + \mathbf{A}^2) \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{A} & \mathbf{\Lambda} \\ \mathbf{\Lambda} & -\mathbf{A} \end{bmatrix}. \tag{217}$$

Therefore, the solution is

$$Z_t = e^{\mathbf{D}t}\left(z + \sqrt{\eta}\sigma\int_0^t e^{-\mathbf{D}s}\mathbf{B}dW_s\right). \tag{218}$$

Clearly, we have that

$$\tilde{\mathbf{E}}(t) := e^{\mathbf{D}t} = \text{diag}\left(e^{-(\lambda_1^2+a_1^2)t}, \cdots, e^{-(\lambda_d^2+a_d^2)t}, e^{-(\lambda_1^2+a_1^2)t}, \cdots, e^{-(\lambda_d^2+a_d^2)t}\right) = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix}, \tag{219}$$

where $\mathbf{E}(t) := \text{diag}\left(e^{-(\lambda_1^2+a_1^2)t}, \cdots, e^{-(\lambda_d^2+a_d^2)t}\right)$. Automatically, we get that

$$e^{-\mathbf{D}s} = \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix}, \tag{220}$$

which implies that

$$Z_t = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix} \left(z + \sqrt{\eta}\sigma \int_0^t \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{\Lambda} \\ \mathbf{\Lambda} & -\mathbf{A} \end{bmatrix} \begin{bmatrix} dW_s^x \\ dW_s^y \end{bmatrix} \right). \tag{221}$$

To conclude, we have that

$$Z_t = \tilde{\mathbf{E}}(t) \left(z + \sqrt{\eta}\sigma \int_0^t \tilde{\mathbf{E}}(-s)\mathbf{M} dW_s \right), \tag{222}$$

where $\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{\Lambda} \\ \mathbf{\Lambda} & -\mathbf{A} \end{bmatrix}$.

We observe that since the expected value of the noise is 0,

$$\mathbb{E}[Z_t] = \begin{bmatrix} \mathbf{E}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(t) \end{bmatrix} z. \tag{223}$$

Therefore, it converges to 0 exponentially fast.

Let us now have a look at the covariance matrix of this process:

$Var(Z_t) = \eta\sigma^2 \begin{bmatrix} \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(2t) \end{bmatrix} Var(V_t)$, where,

$$\begin{aligned} V_t &:= \int_0^t \begin{bmatrix} \mathbf{E}(-s) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{E}(-s) \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{\Lambda} \\ \mathbf{\Lambda} & -\mathbf{A} \end{bmatrix} \begin{bmatrix} dW_s^x \\ dW_s^y \end{bmatrix} \\ &= \begin{bmatrix} \int_0^t \lambda_1 e^{(\lambda_1^2+a_1^2)s} dW_s^{y_1} + \int_0^t a_1 e^{(\lambda_1^2+a_1^2)s} dW_s^{x_1} \\ \vdots \\ \int_0^t \lambda_d e^{(\lambda_d^2+a_d^2)s} dW_s^{y_d} + \int_0^t a_d e^{(\lambda_d^2+a_d^2)s} dW_s^{x_d} \\ \int_0^t \lambda_1 e^{(\lambda_1^2+a_1^2)s} dW_s^{x_1} + \int_0^t a_1 e^{(\lambda_1^2+a_1^2)s} dW_s^{y_1} \\ \vdots \\ \int_0^t \lambda_d e^{(\lambda_d^2+a_d^2)s} dW_s^{x_d} + \int_0^t a_d e^{(\lambda_d^2+a_d^2)s} dW_s^{y_d} \end{bmatrix} =: \begin{bmatrix} a_1^{x_1}(t) + a_2^{y_1}(t) \\ \vdots \\ a_1^{x_d}(t) + a_2^{y_d}(t) \\ a_3^{x_1}(t) + a_4^{y_1}(t) \\ \vdots \\ a_3^{x_d}(t) + a_4^{y_d}(t) \end{bmatrix}. \end{aligned} \tag{224}$$

Therefore,

$$Var(V_t) = \begin{bmatrix} \mathbf{V}^{1,2}(t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{V}^{3,4}(t) \end{bmatrix}, \tag{225}$$

such that

$$\mathbf{V}_{i,i}^{1,2}(t) = Var(a_1^{x_i}(t)) + Var(a_2^{y_i}(t)), \quad \forall i \in \{1, \cdots, d\}, \tag{226}$$

and

$$\mathbf{V}_{i,i}^{3,4}(t) = Var(a_3^{x_i}(t)) + Var(a_4^{y_i}(t)), \quad \forall i \in \{1, \cdots, d\}. \tag{227}$$

Using the well-known Itô Isometry

$$\mathbb{E}\left[\left(\int_0^t H_s dW_s\right)^2\right] = \mathbb{E}\left[\int_0^t H_s^2 ds\right],$$

we get that

$$Var(a_1^{x_i}(t)) + Var(a_2^{y_i}(t)) = \int_0^t e^{2(\lambda_i^2+a_i^2)s}\lambda_i^2 ds + \int_0^t e^{2(\lambda_i^2+a_i^2)s}a_i^2 ds = \frac{1}{2}\left(e^{2(\lambda_i^2+a_i^2)t} - 1\right). \tag{228}$$

Similarly, we get that

$$Var(a_3^{x_i}(t)) + Var(a_4^{y_i}(t)) = \frac{1}{2}\left(e^{2(\lambda_i^2+a_i^2)t} - 1\right). \tag{229}$$

Therefore, we conclude that the covariance matrix of Z_t is

$$Var(Z_t) = \eta\sigma^2 \begin{bmatrix} \mathbf{I}_d - \mathbf{E}(2t) & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d - \mathbf{E}(2t) \end{bmatrix} \bar{\Sigma} \overset{t\to\infty}{=} \frac{\eta\sigma^2}{2}\bar{\Sigma}, \tag{230}$$

where

$$\bar{\Sigma} := \begin{bmatrix} \mathbf{I}_d & \mathbf{0}_d \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix}. \tag{231}$$

□

SEG vs SHGD Interestingly, for both algorithms, the curvature of the landscape influences the speed of convergence. However, the asymptotic covariance matrix of SHGD is not influenced by it. Much differently, the speed of convergence of the expected value of Z_t is strongly influenced by the values of a_i, λ_i, and for SEG also by the sign and magnitude of ρ. For example, if $(\lambda_i^2 - a_i^2)\rho_i^H > a_i^2 + \lambda_i^2 - a_i$, SEG exponentially decays **faster** than SHGD. However, this results in SEG having a **larger** asymptotic variance. Another interesting choice is to reduce the asymptotic variance of SEG by selecting $\rho_i^V = \frac{1}{\lambda_i + a_i}$. In this case, SEG attains its lowest asymptotic variance $\frac{\eta\sigma^2\lambda_i}{2(a_i+\lambda_i)^2}$, which is smaller than $\frac{\eta\sigma^2}{2}$ reached by SHGD only if $a_i^2 + \lambda_i^2 - \lambda_i > 0$. Finally, if $a_i < 0$, $Z = 0$ is a *bad saddle* that one wishes to escape. While SEG can escape it, SHGD is pulled towards it.

Therefore, selecting the size of ρ or ρ_i leads to a trade-off between the speed of convergence and the asymptotic variance. To conclude, there is no clear winner between the two methods as the entire dynamics depend on the curvature of the landscape.

F CONVERGENCE GUARANTEES

In this section, we provide a complete and precise characterization of the dynamics of the Hamiltonian under the dynamics of SEG and SHGD. We then use the latter to derive convergence bounds and establish conditions under which stepsize schedulers guarantee asymptotic convergence. For simplicity, we write $\mathcal{O}(\text{Noise})$ for the terms that depend on dW_t as they vanish once we take an expectation.

In this section, we will often make use of the following shorthand:

1. $H_t = \mathbb{E}_{\boldsymbol{\gamma}}[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)] = \mathbb{E}_{\gamma^1,\gamma^2}\left[\mathcal{H}_{\gamma^1,\gamma^2}(Z_t)\right] = \mathbb{E}_{\boldsymbol{\gamma}}\left[\frac{F_{\gamma^1}^\top(Z_t)F_{\gamma^2}(Z_t)}{2}\right]$;
2. $\Sigma_t^{\text{SHGD}} := \Sigma^{\text{SHGD}}(Z_t)$ and $\Sigma_t^{\text{SEG}} := \Sigma^{\text{SEG}}(Z_t)$;
3. $F_t = F(Z_t)$, $\nabla F_t = \nabla F(Z_t)$, $\nabla f_t = \nabla f(Z_t)$, and $\nabla^2 f_t = \nabla^2 f(Z_t)$.

F.1 SHGD

Lemma F.1 (Dynamics of Hamiltonian)**.** *Let Z_t be the solution of the SDE of SHGD. Then,*

$$\mathbb{E}\left[\dot{H}_t\right] = -\mathbb{E}\left[\|\nabla H_t\|^2\right] + \frac{\eta}{2}Tr\left(\mathbb{E}\left[\Sigma_t^{SHGD}\nabla^2 H_t\right]\right).$$

Proof. The SDE for SHGD is:

$$dZ_t = -\nabla\mathbb{E}_{\boldsymbol{\gamma}}\left[\mathcal{H}_{\boldsymbol{\gamma}}\left(Z_t\right)\right]dt + \sqrt{\eta}\sqrt{\Sigma_t^{\text{SHGD}}}dW_t. \tag{232}$$

Therefore, by Itô's Lemma we have

$$d\mathbb{E}_{\boldsymbol{\gamma}}\left[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)\right] = -\left\|\nabla\mathbb{E}_{\boldsymbol{\gamma}}\left[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)\right]\right\|_2^2 dt + \frac{\eta}{2}\mathrm{Tr}\left(\sqrt{\Sigma_t^{\mathrm{SHGD}}}\nabla^2\mathbb{E}_{\boldsymbol{\gamma}}\left[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)\right]\sqrt{\Sigma_t^{\mathrm{SHGD}}}\right)dt + \mathcal{O}(\text{Noise}). \tag{233}$$

Therefore,

$$d\mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}\left[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)\right]\right] = -\mathbb{E}\left[\left\|\nabla\mathbb{E}_{\boldsymbol{\gamma}}\left[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)\right]\right\|_2^2\right]dt + \frac{\eta}{2}\mathrm{Tr}\left(\mathbb{E}\left[\sqrt{\Sigma_t^{\mathrm{SHGD}}}\nabla^2\mathbb{E}_{\boldsymbol{\gamma}}\left[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)\right]\sqrt{\Sigma_t^{\mathrm{SHGD}}}\right]\right)dt, \tag{234}$$

which implies that

$$\mathbb{E}\left[\dot{H}_t\right] = -\mathbb{E}\left[\left\|\nabla H_t\right\|^2\right] + \frac{\eta}{2}\mathrm{Tr}\left(\mathbb{E}\left[\Sigma_t^{\mathrm{SHGD}}\nabla^2 H_t\right]\right). \tag{235}$$

□

Theorem F.2 (SHGD General Convergence)**.** *Consider the solution Z_t of the SHGD SDE with $\gamma^1 \neq \gamma^2$. Let $v_t := \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}\left[\left\|\nabla\mathcal{H}\left(Z_t\right) - \nabla\mathcal{H}_{\boldsymbol{\gamma}}\left(Z_t\right)\right\|_2^2\right]\right]$ measure the error in $\nabla\mathcal{H}$, in expectation over the whole randomness up to time t. Suppose that:*

1. *The smallest eigenvalue (in absolute value) μ of $\nabla^2 f(z)$ is non-zero;*
2. *$\left\|\nabla^2 H(z)\right\|_{op} < \mathcal{L}_{\mathcal{T}}$, for all $z \in \mathbb{R}^{2d}$.*

Then,

$$\mathbb{E}\left[H_t\right] \leq e^{-2\mu^2 t}\left[H_0 + \frac{\eta\mathcal{L}_{\mathcal{T}}}{2}\int_0^t v_s e^{2\mu^2 s}ds\right]. \tag{236}$$

Proof. Under these assumptions,

$$\begin{aligned}
dH_t &= -\left\|\nabla H_t\right\|^2 dt + \frac{\eta}{2}\mathrm{Tr}\left(\nabla^2 H_t\Sigma_t^{\mathrm{SHGD}}\right)dt + \mathcal{O}(\text{Noise}) \\
&= -\left\|F_t^\top\nabla F_t\right\|^2 dt + \frac{\eta}{2}\mathrm{Tr}\left(\nabla^2 H_t\Sigma_t^{\mathrm{SHGD}}\right)dt + \mathcal{O}(\text{Noise}) \\
&= -\left\|\nabla^2 f_t\nabla f_t\right\|^2 dt + \frac{\eta}{2}\mathrm{Tr}\left(\nabla^2 H_t\Sigma_t^{\mathrm{SHGD}}\right)dt + \mathcal{O}(\text{Noise}) \\
&\leq -\mu^2\left\|\nabla f_t\right\|^2 dt + \frac{\eta}{2}\left\|\nabla^2 H_t\right\|_{\mathrm{op}}\mathrm{Tr}\left(\Sigma_t^{\mathrm{SHGD}}\right)dt + \mathcal{O}(\text{Noise}) \\
&\leq -\mu^2\left\|\nabla f_t\right\|^2 dt + \frac{\eta}{2}\mathcal{L}_{\mathcal{T}}\mathrm{Tr}\left(\Sigma_t^{\mathrm{SHGD}}\right)dt + \mathcal{O}(\text{Noise}),
\end{aligned} \tag{237}$$

where we used that $\mathrm{Tr}(\mathbf{AB}) \leq \mathrm{Tr}(\mathbf{A})\|\mathbf{B}\|_{\mathrm{op}}$ if $\mathbf{A}$ is a real positive semi-definite matrix and $\mathbf{B}$ is of the same size. Then, we observe that

$$\mathbb{E}\left[\mathrm{Tr}\left(\Sigma_t^{\mathrm{SHGD}}\right)\right] = \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}\left[\left\|\nabla\mathcal{H}\left(Z_t\right) - \nabla\mathcal{H}_{\boldsymbol{\gamma}}\left(Z_t\right)\right\|_2^2\right]\right] = v_t, \tag{238}$$

which implies that

$$\begin{aligned}
d\mathbb{E}\left[H_t\right] &\leq -\mu^2\mathbb{E}\left[\left\|\nabla f_t\right\|^2\right]dt + \frac{\eta}{2}\mathcal{L}_{\mathcal{T}}v_t dt \\
&= -2\mu^2\mathbb{E}\left[H_t\right]dt + \frac{\eta}{2}\mathcal{L}_{\mathcal{T}}v_t dt,
\end{aligned} \tag{239}$$

which in turn implies that

$$\mathbb{E}\left[H_t\right] \leq e^{-2\mu^2 t}\left[H_0 + \frac{\eta\mathcal{L}_{\mathcal{T}}}{2}\int_0^t v_s e^{2\mu^2 s}ds\right]. \tag{240}$$

□

Corollary F.3. *Under the assumptions of Theorem F.2, if for $\mathcal{L}_{\mathcal{V}} > 0$*

$$v_t \leq \mathcal{L}_{\mathcal{V}}\mathbb{E}\left[H_t\right], \tag{241}$$

the solution is

$$\mathbb{E}\left[H_t\right] \leq H_0 e^{\left(-2\mu^2+\eta\mathcal{L}_{\mathcal{V}}\mathcal{L}_{\mathcal{T}}\right)t}. \tag{242}$$

If instead

$$v_t \leq \mathcal{L}_{\mathcal{V}}, \tag{243}$$

we have

$$\mathbb{E}\left[H_t\right] \leq H_0 e^{-2\mu^2 t} + \left(1 - e^{-2\mu^2 t}\right) \frac{\eta\mathcal{L}_{\mathcal{V}}\mathcal{L}_{\mathcal{T}}}{2\mu^2}. \tag{244}$$

In more generality, if

$$v_t \leq \mathcal{L}_{\mathcal{V}}\mathbb{E}\left[H_t\right]^{\alpha}, \quad \alpha \in [0,1) \cup (1,\infty), \tag{245}$$

the solution is even more interesting as:

1. *If* $\alpha > 1$, $\mathbb{E}\left[H_t\right] \to 0$ *as* $e^{-2\mu^2 t}$*;*
2. *If* $\alpha < 1$, $\mathbb{E}\left[H_t\right] \to \left(\frac{\eta\mathcal{L}_{\mathcal{T}}\mathcal{L}_{\mathcal{V}}}{2\mu^2}\right)^{\frac{1}{1-\alpha}}$.

Proof. Let us first consider the case where, for some $\mathcal{L}_{\mathcal{V}} > 0$,

$$v_t \leq \mathcal{L}_{\mathcal{V}}\mathbb{E}\left[H_t\right]. \tag{246}$$

This implies that

$$d\mathbb{E}\left[H_t\right] \leq -2\mu^2\mathbb{E}\left[H_t\right]dt + \frac{\eta}{2}\mathcal{L}_{\mathcal{T}}\mathcal{L}_{\mathcal{V}}\mathbb{E}\left[H_t\right]dt. \tag{247}$$

By renaming $r_t := \mathbb{E}\left[H_t\right]$, we have

$$dr_t \leq -2\mu^2 r_t dt + \eta\mathcal{L}_{\mathcal{T}}\mathcal{L}_{\mathcal{V}} r_t dt, \tag{248}$$

which results in

$$\mathbb{E}\left[H_t\right] \leq H_0 e^{\left(-2\mu^2+\eta\mathcal{L}_{\mathcal{V}}\mathcal{L}_{\mathcal{T}}\right)t}. \tag{249}$$

If instead

$$v_t \leq \mathcal{L}_{\mathcal{V}}, \tag{250}$$

we automatically have

$$\mathbb{E}\left[H(Z_t)\right] \leq H(Z_0)e^{-2\mu^2 t} + \left(1 - e^{-2\mu^2 t}\right) \frac{\eta\mathcal{L}_{\mathcal{V}}\mathcal{L}_{\mathcal{T}}}{2\mu^2}. \tag{251}$$

More in general, if we assume that for some $\mathcal{L}_{\mathcal{V}} > 0$,

$$v_t \leq \mathcal{L}_{\mathcal{V}}\mathbb{E}\left[H_t\right]^{\alpha}, \tag{252}$$

we have

$$dr_t \leq -2\mu^2 r_t dt + \eta\mathcal{L}_{\mathcal{T}}\mathcal{L}_{\mathcal{V}} r_t^{\alpha} dt, \tag{253}$$

which implies that for $\alpha \neq 1$:

$$r_t \leq \sqrt[1-\alpha]{\frac{r_0^{-\alpha}e^{(\alpha-1)2\mu^2 t}\left(r_0 2\mu^2 - \eta\mathcal{L}_{\mathcal{T}}\mathcal{L}_{\mathcal{V}} r_0^{\alpha}\right) + \eta\mathcal{L}_{\mathcal{T}}\mathcal{L}_{\mathcal{V}}}{2\mu^2}} =: b_t \tag{254}$$

In these cases, we have that the bound b_t converges or diverges depending on the magnitude of α:

1. If $\alpha > 1$, $b_t \to 0$ as $e^{-2\mu^2 t}$;
2. If $\alpha < 1$, $b_t \to \left(\frac{\eta\mathcal{L}_{\mathcal{T}}\mathcal{L}_{\mathcal{V}}}{2\mu^2}\right)^{\frac{1}{1-\alpha}}$.

□

Corollary F.4. *β-Error Bound on F and L-Lipschitzianity on $\nabla\mathcal{H}_{\gamma^1,\gamma^2}$ and $\nabla\mathcal{H}$, implies that $v_t \leq 8L^2\beta^2\mathbb{E}\left[H_t\right]$.*

Proof.

$$\begin{aligned} v_t =& \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}[\|\nabla \mathcal{H}_{\boldsymbol{\gamma}}(Z_t) - \nabla \mathcal{H}(Z_t)\|_2^2]\right] = \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}[\|\nabla \mathcal{H}_{\boldsymbol{\gamma}}(Z_t) - \mathcal{H}_{\boldsymbol{\gamma}}(Z^*) + \nabla \mathcal{H}(Z^*) - \nabla \mathcal{H}(Z_t)\|_2^2]\right] \\ \leq& 2\mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}[\|\nabla \mathcal{H}_{\boldsymbol{\gamma}}(Z_t) - \mathcal{H}_{\boldsymbol{\gamma}}(Z^*)\|_2^2]\right] + 2\mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}[\|\nabla \mathcal{H}(Z^*) - \nabla \mathcal{H}(Z_t)\|_2^2]\right] \leq 4L^2\mathbb{E}\left[\|Z_t - Z^*\|_2^2\right] \\ \leq& 4\beta^2 L^2 \mathbb{E}\left[\|F(Z_t)\|_2^2\right] = 8L^2\beta^2\mathbb{E}\left[H_t\right]. \end{aligned} \tag{255}$$

□

The following corollary exemplifies the case where v_t is bounded and we achieve convergence only up to a certain ball.

Corollary F.5. *Under the assumptions of Theorem D.3, for $f(x,y) := x^\top \mathbf{\Lambda} y$, we have:*

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{=} \frac{\eta}{2}\sum_{i=1}^{d}\sigma_i^2 > 0. \tag{256}$$

Proof. It is easy to see that

$$\frac{\|Z_t\|^2}{2} = \sum_{i=1}^{d}\frac{\|Z_t^i\|^2}{2}, \tag{257}$$

where $Z^i := (X^i, Y^i)$, and that

$$d\left(\frac{\|Z_t^i\|^2}{2}\right) = -2\lambda_i^2\frac{\|Z_t^i\|^2}{2}dt + \eta\sigma_i^2\lambda_i^2 dt + \mathcal{O}(\text{Noise}). \tag{258}$$

This implies that

$$d\left(\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2}\right) = -2\lambda_i^2\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2}dt + \eta\sigma_i^2\lambda_i^2 dt, \tag{259}$$

which implies that

$$\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} = \frac{\|Z_0^i\|^2}{2}e^{-2\lambda_i^2 t} + (1 - e^{-2\lambda_i^2 t})\frac{\eta\sigma_i^2}{2} \overset{t\to\infty}{\to} \frac{\eta\sigma_i^2}{2}. \tag{260}$$

□

Interestingly, one can recover convergence by allowing stepsize schedulers. In the following result, we derive a necessary and sufficient condition to craft such schedulers. Then, we provide two concrete examples.

Corollary F.6 (SHGD Insights)**.** *Under the assumptions of Theorem D.3, for $f(x,y) := x^\top \mathbf{\Lambda} y$, for any positive scheduler η_t we have*

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} = \sum_{i=1}^{d} e^{-2\lambda_i^2\int_0^t \eta_s ds}\left(\frac{\|Z_0^i\|^2}{2} + \eta\sigma_i^2\lambda_i^2\int_0^t e^{2\lambda_i^2\int_0^s \eta_r dr}\eta_s^2 ds\right). \tag{261}$$

Therefore,

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{\to} 0 \iff \int_0^\infty \eta_s ds = \infty \text{ and } \lim_{t\to\infty}\eta_t = 0. \tag{262}$$

In particular,

1. $\eta_t = 1$ *implies that*
$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{=} \frac{\eta}{2}\sum_{i=1}^{d}\sigma_i^2 > 0; \tag{263}$$

2. $\eta_t = \frac{1}{(t+1)^\gamma}$ *for* $\gamma \in \{0.5, 1\}$, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \to 0$;

3. $\eta_t = \frac{1}{(t+1)^2}$, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \nrightarrow 0$.

Proof. For $f(x,y) := x^\top \mathbf{\Lambda} y$, with the noise on gradient assumption, the SDE when we include a scheduler η_t is

$$dZ_t = \mathbf{A} Z_t \eta_t dt + \sqrt{\eta}\eta_t \sigma \mathbf{B} dW_t, \tag{264}$$

where

$$\mathbf{A} = \begin{bmatrix} -\mathbf{\Lambda}^2 & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{\Lambda}^2 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{0}_d & \mathbf{\Lambda} \\ \mathbf{\Lambda} & \mathbf{0}_d \end{bmatrix}. \tag{265}$$

Therefore,

$$d\left(\frac{\|Z_t^i\|^2}{2}\right) = -2\lambda_i^2 \eta_t \frac{\|Z_t^i\|^2}{2} dt + \eta \sigma_i^2 \lambda_i^2 \eta_t^2 dt + \mathcal{O}(\text{Noise}), \tag{266}$$

which implies that

$$d\left(\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2}\right) = -2\lambda_i^2 \frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} \eta_t dt + \eta \sigma_i^2 \lambda_i^2 \eta_t^2 dt, \tag{267}$$

which ultimately implies that

$$\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} = e^{-2\lambda_i^2 \int_0^t \eta_s ds} \left(\frac{\|Z_0^i\|^2}{2} + \eta \sigma_i^2 \lambda_i^2 \int_0^t e^{2\lambda_i^2 \int_0^s \eta_r dr} \eta_s^2 ds \right). \tag{268}$$

Let us write out the necessary conditions for this quantity to go to 0:

1. For the first part, $e^{-2\lambda_i^2 \int_0^t \eta_s ds} \frac{\|Z_0^i\|^2}{2}$ goes to 0, if and only if $\int_0^\infty \eta_s ds = \infty$;
2. For the second part, we need $\eta \sigma_i^2 \lambda_i^2 e^{-2\lambda_i^2 \int_0^t \eta_s ds} \int_0^t e^{2\lambda_i^2 \int_0^s \eta_r dr} \eta_s^2 ds$ to go to 0 as well.

Let us rewrite the second condition in a more convenient way:

$$\eta \sigma_i^2 \lambda_i^2 e^{-2\lambda_i^2 \int_0^t \eta_s ds} \int_0^t e^{2\lambda_i^2 \int_0^s \eta_r dr} \eta_s^2 ds = \eta \sigma_i^2 \lambda_i^2 \frac{\int_0^t e^{2\lambda_i^2 \int_0^s \eta_r dr} \eta_s^2 ds}{e^{2\lambda_i^2 \int_0^t \eta_s ds}}. \tag{269}$$

Since $\int_0^\infty \eta_s ds = \infty$, both the numerator and denominator diverge. Therefore, we can use L'Hôpital's rule:

$$\lim_{t\to\infty} \eta \sigma_i^2 \lambda_i^2 \frac{\int_0^t e^{2\lambda_i^2 \int_0^s \eta_r dr} \eta_s^2 ds}{e^{2\lambda_i^2 \int_0^t \eta_s ds}} = \lim_{t\to\infty} \eta \sigma_i^2 \lambda_i^2 \frac{e^{2\lambda_i^2 \int_0^t \eta_s ds} \eta_t^2}{e^{2\lambda_i^2 \int_0^t \eta_s ds} 2\lambda_i^2 \eta_t} = \lim_{t\to\infty} \frac{\eta \sigma_i^2}{2} \eta_t, \tag{270}$$

which converges to 0, if and only if $\lim_{t\to\infty} \eta_t = 0$.

Note that, if the first condition is violated, the first component does not go to 0. If the second condition is not satisfied, the second component does not go to 0.

In particular,

1. $\eta_t = \frac{1}{t+1}$ and $2\lambda_i^2 \neq 1 \implies \frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} = \frac{(t+1)^{-2\lambda_i^2 - 1}\left(\frac{\|Z_0^i\|^2}{2}(2\lambda_i^2 - 1)(t+1) + \eta\sigma_i^2\lambda_i^2\left((t+1)^{2\lambda_i^2} - t - 1\right)\right)}{2\lambda_i^2 - 1} \overset{t\to\infty}{\to} 0$;
2. $\eta_t = \frac{1}{t+1}$ and $2\lambda_i^2 = 1 \implies \frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} = \frac{\frac{\|Z_0^i\|^2}{2} + \eta\sigma_i^2\lambda_i^2 \log(t+1)}{t+1} \overset{t\to\infty}{\to} 0$;
3. $\eta_t = \frac{1}{\sqrt{t+1}} \implies \frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} = e^{-4\lambda_i^2\sqrt{t+1}}\left(\frac{\|Z_0^i\|^2}{2} e^{4\lambda_i^2} + 2\eta\sigma_i^2\lambda_i^2\left(Ei(4\lambda_i^2\sqrt{t+1}) - Ei(4\lambda_i^2)\right)\right) \overset{t\to\infty}{\to} 0$;
4. $\eta_t = \frac{1}{(t+1)^2} \implies \frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} = \frac{\|Z_0^i\|^2}{2} e^{-\frac{2\lambda_i^2 t}{t+1}} - \eta\sigma_i^2\lambda_i^2 \left(\frac{\left(4\lambda_i^4 + 4\lambda_i^2 + 2\right)e^{-\frac{2\lambda_i^2 t}{t+1}}}{8\lambda_i^6} + \frac{\left(4\lambda_i^4 + 4\lambda_i^2(t+1) + 2(t+1)^2\right)}{8\lambda_i^6 (t+1)^2} \right) \overset{t\to\infty}{\to} 0$.

□

Now we study a case where the noise structure itself is enough to guarantee the convergence. In this case, v_t scales with $\mathbb{E}[H_t]$.

Corollary F.7 (Noise on Data). *For $f(x,y) = x^\top \mathbb{E}_\xi[\mathbf{\Lambda}_\xi]\, y$ such that $\mathbf{\Lambda}_\xi$ is diagonal, we have*

1. $\mathbb{E}[Z_t] = \tilde{\mathbf{E}}(t) z \overset{t\to\infty}{=} 0$;
2. $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} = \sum_{i=1}^d \frac{\|Z_0^i\|^2}{2} e^{-\left(2\lambda_i^2 - \eta\sigma_i^2\left(2\lambda_1^2+\sigma_i^2\right)\right)t}$.

In particular, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \to 0$ *if* $\eta < \frac{2\lambda_i^2}{\sigma_i^2\left(2\lambda_1^2+\sigma_i^2\right)}, \quad \forall i.$

Proof. The derivation of the SDE is straightforward and the formula is

$$dZ_t = \mathbf{A} Z_t dt + \sqrt{\eta}\mathbf{B} dW_t, \tag{271}$$

where

$$\mathbf{A} = \begin{bmatrix} -\mathbf{\Lambda}^2 & \mathbf{0}_d \\ \mathbf{0}_d & -\mathbf{\Lambda}^2 \end{bmatrix} \quad \text{and} \quad \mathbf{B}\mathbf{B}^\top = \begin{bmatrix} 2\mathbf{\Lambda}^2 \circ \Sigma^2 + \Sigma^4 & 2\mathbf{\Lambda}^2 \circ \Sigma^2 + \Sigma^4 \\ 2\mathbf{\Lambda}^2 \circ \Sigma^2 + \Sigma^4 & 2\mathbf{\Lambda}^2 \circ \Sigma^2 + \Sigma^4 \end{bmatrix} \circ \begin{bmatrix} \operatorname{diag}(X_t \circ X_t) & \operatorname{diag}(X_t \circ Y_t) \\ \operatorname{diag}(X_t \circ Y_t) & \operatorname{diag}(Y_t \circ Y_t) \end{bmatrix}. \tag{272}$$

It is easy to see that

$$\frac{\|Z_t\|^2}{2} = \sum_{i=1}^d \frac{\|Z_t^i\|^2}{2}, \tag{273}$$

where $Z^i := (X^i, Y^i)$, and that

$$d\left(\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2}\right) = -2\lambda_i^2 \frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} dt + \eta\sigma_i^2(2\lambda_i^2 + \sigma_i^2)\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} dt, \tag{274}$$

which ultimately implies that

$$\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} = \frac{\|Z_0^i\|^2}{2} e^{-\left(2\lambda_i^2 - \eta\sigma_i^2\left(2\lambda_1^2+\sigma_i^2\right)\right)t}. \tag{275}$$

Empirical validation of this result is provided in Figure 7. □

F.2 SEG

Lemma F.8 (Dynamics of Hamiltonian). *Let Z_t be the solution of the SEG SDE. Then,*

$$\mathbb{E}\left[\dot{H}_t\right] = -\mathbb{E}\left[\nabla H_t^\top F_t^{SEG}\right] + \frac{\eta}{2} Tr\left(\mathbb{E}\left[\Sigma_t^{SEG} \nabla^2 H_t\right]\right). \tag{276}$$

Under the additional assumption that $\mathbb{E}_{\boldsymbol{\gamma}}[\nabla F_{\boldsymbol{\gamma}}(Z_t) F_{\boldsymbol{\gamma}}(Z_t)] = \nabla F(Z_t) F(Z_t)$, the formula simplifies and is

$$d\mathbb{E}[H_t] = -\mathbb{E}\left[F_t^\top (\nabla F_t - \rho(\nabla F_t)^2) F_t\right] dt + \frac{\eta}{2} Tr\left(\mathbb{E}\left[\Sigma_t^{SEG} \nabla^2 H_t\right]\right) dt. \tag{277}$$

Proof. The SDE for SEG is

$$dZ_t = -F^{\text{SEG}}(Z_t)\, dt + \sqrt{\eta}\sqrt{\Sigma_t^{\text{SEG}}} dW_t. \tag{278}$$

Therefore, by Itô's Lemma we have

$$d\mathbb{E}_{\boldsymbol{\gamma}}[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)] = -\nabla \mathbb{E}_{\boldsymbol{\gamma}}[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)]^\top F^{\text{SEG}} dt + \frac{\eta}{2} \text{Tr}\left(\sqrt{\Sigma_t^{\text{SEG}}} \nabla^2 \mathbb{E}_{\boldsymbol{\gamma}}[\mathcal{H}_{\boldsymbol{\gamma}}(Z_t)] \sqrt{\Sigma_t^{\text{SEG}}}\right) dt + \mathcal{O}(\text{Noise}). \tag{279}$$

Therefore,

$$\mathbb{E}\left[\dot{H}_t\right] = -\mathbb{E}\left[\nabla H_t^\top F_t^{\text{SEG}}\right] + \frac{\eta}{2} \text{Tr}\left(\mathbb{E}\left[\Sigma_t^{\text{SEG}} \nabla^2 H_t\right]\right). \tag{280}$$

□

Theorem F.9 (SEG General Convergence)**.** *Consider the solution Z_t of the SEG SDE with $\gamma^1 \neq \gamma^2$. Let $v_t := \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}\left[\|F^{SEG}(Z_t) - F^{SEG}_{\boldsymbol{\gamma}}(Z_t)\|_2^2\right]\right]$ measure the error in F^{SEG}, in expectation over the whole randomness up to time t. Suppose that:*

1. *The smallest eigenvalue (in absolute value) μ_ρ of $\mathbf{M}(z)$ is non-zero, where $\mathbf{M}(z) = \mathrm{diag}(\mathbf{M}_{1,1}(z), \mathbf{M}_{2,2}(z))$, with $\mathbf{M}_{1,1}(z) := \nabla^2 f_{xx}(z) + \rho\left(\nabla^2 f_{xy}(z)\nabla^2 f_{xy}(z)^T - (\nabla^2 f_{xx}(z))^2\right)$, and $\mathbf{M}_{2,2}(z) := -\nabla^2 f_{yy}(z) + \rho\left(\nabla^2 f_{xy}(z)\nabla^2 f_{xy}(z)^T - (\nabla^2 f_{yy}(z))^2\right)$;*
2. *$\|\nabla^2 H(z)\|_{op} < \mathcal{L}_{\mathcal{T}}$, for all $z \in \mathbb{R}^{2d}$.*

Then,

$$\mathbb{E}\left[H_t\right] \leq e^{-2\mu_\rho^2 t}\left[H_0 + \frac{\eta\mathcal{L}_{\mathcal{T}}}{2}\int_0^t v_s e^{2\mu_\rho^2 s} ds\right]. \tag{281}$$

Proof. From the previous theorem, we have that

$$d\mathbb{E}\left[H_t\right] = -\,\mathbb{E}\left[F_t^\top(\nabla F_t - \rho(\nabla F_t)^2)F_t\right]dt + \frac{\eta}{2}\mathrm{Tr}\left(\mathbb{E}\left[\Sigma_t^{\mathrm{SEG}}\nabla^2 H_t\right]\right)dt. \tag{282}$$

After observing that

$$F^\top(z)(\nabla F(z) - \rho(\nabla F(z))^2)F(z) = F(z)^\top \mathbf{M}(z)F(z). \tag{283}$$

We have that

$$\begin{aligned} d\mathbb{E}\left[H_t\right] &= -\,\mathbb{E}\left[F_t^\top(\nabla F_t - \rho(\nabla F_t)^2)F_t\right]dt + \frac{\eta}{2}\mathrm{Tr}\left(\mathbb{E}\left[\Sigma_t^{\mathrm{SEG}}\nabla^2 H_t\right]\right)dt \\ &= -\,\mathbb{E}\left[F_t^\top \mathbf{M}_t F_t\right]dt + \frac{\eta}{2}\mathrm{Tr}\left(\mathbb{E}\left[\Sigma_t^{\mathrm{SEG}}\nabla^2 H_t\right]\right)dt \\ &\leq -\,\mu_\rho^2\mathbb{E}\left[\|F_t\|_2^2\right]dt + \frac{\eta}{2}\mathbb{E}\left[\|\nabla^2 H_t\|_{\mathrm{op}}\mathrm{Tr}\left(\Sigma_t^{\mathrm{SEG}}\right)\right]dt \\ &\leq -\,\mu_\rho^2\mathbb{E}\left[\|F_t\|_2^2\right]dt + \frac{\eta\mathcal{L}_{\mathcal{T}}}{2}\mathbb{E}\left[\mathrm{Tr}\left(\Sigma_t^{\mathrm{SEG}}\right)\right]dt, \end{aligned} \tag{284}$$

where we used that $\mathrm{Tr}(\mathbf{AB}) \leq \mathrm{Tr}(\mathbf{A})\|\mathbf{B}\|_{\mathrm{op}}$ if $\mathbf{A}$ is a real positive semi-definite matrix and $\mathbf{B}$ is of the same size. Then, we observe that

$$\mathbb{E}\left[\mathrm{Tr}\left(\Sigma_t^{\mathrm{SEG}}\right)\right] = \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}\left[\|(F(Z_t) - F_{\gamma^1}(Z_t) - \rho\left(\nabla F(Z_t)F(Z_t) - \nabla F_{\gamma^1}(Z_t)F_{\gamma^2}(Z_t)\right))\|^2\right]\right] = v_t, \tag{285}$$

which implies that

$$\begin{aligned} d\mathbb{E}\left[H_t\right] &\leq -\mu_\rho^2\mathbb{E}\left[\|F_t\|^2\right]dt + \frac{\eta}{2}\mathcal{L}_{\mathcal{T}} v_t dt \\ &= -2\mu_\rho^2\mathbb{E}\left[H_t\right]dt + \frac{\eta}{2}\mathcal{L}_{\mathcal{T}} v_t dt, \end{aligned} \tag{286}$$

which in turn implies that

$$\mathbb{E}\left[H_t\right] \leq e^{-2\mu_\rho^2 t}\left[H_0 + \frac{\eta\mathcal{L}_{\mathcal{T}}}{2}\int_0^t v_s e^{2\mu_\rho^2 s} ds\right]. \tag{287}$$

□

Corollary F.10. *Under the assumptions of Theorem F.9, if for $\mathcal{L}_{\mathcal{V}} > 0$*

$$v_t \leq \mathcal{L}_{\mathcal{V}}\mathbb{E}\left[H_t\right], \tag{288}$$

the solution is

$$\mathbb{E}\left[H_t\right] \leq H_0 e^{\left(-2\mu_\rho^2 + \eta\mathcal{L}_{\mathcal{V}}\mathcal{L}_{\mathcal{T}}\right)t}. \tag{289}$$

If instead

$$v_t \leq \mathcal{L}_{\mathcal{V}}, \tag{290}$$

we have

$$\mathbb{E}\left[H_t\right] \leq H_0 e^{-2\mu_\rho^2 t} + \left(1 - e^{-2\mu_\rho^2 t}\right)\frac{\eta\mathcal{L}_{\mathcal{V}}\mathcal{L}_{\mathcal{T}}}{2\mu_\rho^2}. \tag{291}$$

More in general, if

$$v_t \leq \mathcal{L}_{\mathcal{V}} \mathbb{E}\left[H_t\right]^{\alpha}, \quad \mathcal{L}_{\mathcal{V}} > 0, \quad \alpha \in [0,1) \cup (1,\infty). \tag{292}$$

The solution is even more interesting:

1. *If* $\alpha > 1$, $\mathbb{E}\left[H_t\right] \to 0$ *as* $e^{-2\mu_\rho^2 t}$;
2. *If* $\alpha < 1$, $\mathbb{E}\left[H_t\right] \to \left(\frac{\eta \mathcal{L}_{\mathcal{T}} \mathcal{L}_{\mathcal{V}}}{2\mu_\rho^2}\right)^{\frac{1}{1-\alpha}}$.

Proof. The proof is the same as Corollary F.3 where we substitute μ with μ_ρ. □

Corollary F.11. *κ_1-Lipschitzianity on F_{γ^1}, κ_2-Lipschitzianity on $\nabla F_{\gamma^1} F_{\gamma^2}$, and β-Error Bound on F, implies that $v_t \leq 16\beta^2(\kappa_1^2 + \rho^2\kappa_2^2)\mathbb{E}\left[H_t\right]$.*

Proof.

$$\begin{aligned} v_t &= \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}\left[\left\|F\left(Z_t\right) - F_{\gamma^1}\left(Z_t\right) - \rho\left(\nabla F\left(Z_t\right) F\left(Z_t\right) - \nabla F_{\gamma^1}\left(Z_t\right) F_{\gamma^2}\left(Z_t\right)\right)\right\|^2\right]\right] \\ &\leq 2\mathbb{E}\left[\mathbb{E}_{\gamma^1}\left[\left\|F\left(Z_t\right) - F_{\gamma^1}\left(Z_t\right)\right\|^2\right]\right] + 2\rho^2\mathbb{E}\left[\mathbb{E}_{\boldsymbol{\gamma}}\left[\left\|\nabla F\left(Z_t\right) F\left(Z_t\right) - \nabla F_{\gamma^1}\left(Z_t\right) F_{\gamma^2}\left(Z_t\right)\right\|^2\right]\right] \\ &\leq 8(\kappa_1^2 + \rho^2\kappa_2^2)\mathbb{E}[\|Z_t - Z^*\|_2^2] \leq 8\beta^2(\kappa_1^2 + \rho^2\kappa_2^2)\mathbb{E}[\|F(Z_t)\|_2^2] = 16\beta^2(\kappa_1^2 + \rho^2\kappa_2^2)\mathbb{E}\left[H_t\right]. \end{aligned} \tag{293}$$

□

The following corollary exemplifies the case where v_t is bounded and we achieve convergence only up to a certain ball, and that selecting a proper ρ has a crucial role: If it is too large, it might increase the suboptimality of the algorithm.

Corollary F.12. *Under the assumptions of Theorem D.1, for $f(x,y) := x^\top \boldsymbol{\Lambda} y$, we have:*

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{=} \eta \sum_{i=1}^{d} \sigma_i^2 \frac{1 + \rho^2\lambda_i^2}{2\rho\lambda_i^2} > 0. \tag{294}$$

Proof. It is easy to see that:

$$\frac{\|Z_t\|^2}{2} = \sum_{i=1}^{d} \frac{\|Z_t^i\|^2}{2}, \tag{295}$$

where $Z^i := (X^i, Y^i)$, and that

$$d\left(\frac{\|Z_t^i\|^2}{2}\right) = -2\rho\lambda_i^2 \frac{\|Z_t^i\|^2}{2} dt + \eta\sigma_i^2(1 + \rho^2\lambda_i^2)dt + \mathcal{O}(\text{Noise}). \tag{296}$$

This implies that

$$d\left(\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2}\right) = -2\rho\lambda_i^2 \frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} dt + \eta\sigma_i^2(1 + \rho^2\lambda_i^2)dt. \tag{297}$$

Which implies that

$$\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} = \frac{\|Z_0^i\|^2}{2} e^{-2\rho\lambda_i^2 t} + (1 - e^{-2\lambda_i^2 t}) \frac{\eta\sigma_i^2}{2} \frac{(1 + \rho^2\lambda_i^2)}{\rho\lambda_i^2} \overset{t\to\infty}{\to} \eta\sigma_i^2 \frac{(1 + \rho^2\lambda_i^2)}{2\rho\lambda_i^2}. \tag{298}$$

□

Interestingly, one can recover convergence by allowing stepsize schedulers. In the following result, we derive a necessary and sufficient condition to craft such schedulers. Then, we provide two concrete examples.

Corollary F.13 (SEG Insights)**.** *Under the assumptions of Theorem D.1, for $f(x,y) := x^\top \mathbf{\Lambda} y$, for any positive schedulers η_t and ρ_t we have*

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} = \sum_{i=1}^{d} e^{-2\lambda_i^2 \rho \int_0^t \eta_s \rho_s ds} \left(\frac{\|Z_0^i\|^2}{2} + \eta \sigma_i^2 \int_0^t e^{2\lambda_i^2 \rho \int_0^s \eta_r \rho_r dr} \eta_s^2 (1 + \lambda_i^2 \rho^2 \rho_s^2) ds \right). \tag{299}$$

Therefore,

$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{\to} 0 \iff \int_0^\infty \eta_s \rho_s ds = \infty \text{ and } \lim_{t\to\infty} \eta_t \rho_t = \lim_{t\to\infty} \frac{\eta_t}{\rho_t} = 0. \tag{300}$$

In particular, consistently with (Hsieh et al., 2020),

1. *$\eta_t = \rho_t = 1$ implies that*
$$\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \overset{t\to\infty}{=} \eta \sum_{i=1}^{d} \sigma_i^2 \frac{1 + \rho^2 \lambda_i^2}{2\rho \lambda_i^2} > 0; \tag{301}$$
2. *$\eta_t = \frac{1}{(t+1)^\gamma}$ and $\rho_t = 1$, $\gamma \in \{0.5, 1\}$, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \to 0$;*
3. *$\eta_t = \frac{1}{(t+1)^2}$ and $\rho_t = 1$, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \nrightarrow 0$.*

Proof. In this case, the SDE when we include the schedulers η_t and ρ_t is

$$dZ_t = \mathbf{A} Z_t \eta_t dt + \sqrt{\eta} \eta_t \sigma \mathbf{B} dW_t \tag{302}$$

where

$$\mathbf{A} = \begin{bmatrix} -\rho \rho_t \mathbf{\Lambda}^2 & -\mathbf{\Lambda} \\ \mathbf{\Lambda} & -\rho \rho_t \mathbf{\Lambda}^2 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{I}_d & -\rho \rho_t \mathbf{\Lambda} \\ \rho \rho_t \mathbf{\Lambda} & \mathbf{I}_d \end{bmatrix}. \tag{303}$$

Therefore,

$$d\left(\frac{\|Z_t^i\|^2}{2}\right) = -2\rho \lambda_i^2 \rho_t \eta_t \frac{\|Z_t^i\|^2}{2} dt + \eta \sigma_i^2 \left(1 + \lambda_i^2 \rho^2 \rho_t^2\right) \eta_t^2 dt + \mathcal{O}(\text{Noise}), \tag{304}$$

which implies that

$$d\left(\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2}\right) = -2\rho \lambda_i^2 \rho_t \eta_t \frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} dt + \eta \sigma_i^2 \left(1 + \lambda_i^2 \rho^2 \rho_t^2\right) \eta_t^2 dt, \tag{305}$$

which implies that

$$\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} = e^{-2\lambda_i^2 \rho \int_0^t \eta_s \rho_s ds} \left(\frac{\|Z_0^i\|^2}{2} + \eta \sigma_i^2 \int_0^t e^{2\lambda_i^2 \rho \int_0^s \eta_r \rho_r dr} \eta_s^2 (1 + \lambda_i^2 \rho^2 \rho_s^2) ds \right). \tag{306}$$

With arguments similar to Corollary F.6, the necessary and sufficient conditions for convergence are the following:

1. For the first part $e^{-2\lambda_i^2 \int_0^t \rho_s \eta_s ds} \frac{\|Z_0^i\|^2}{2}$ to go to 0, we need $\int_0^\infty \eta_s \rho_s ds = \infty$;
2. For the second part to go to 0, we need both $\frac{\eta_t}{\rho_t}$ and $\eta_t \rho_t$ to go to 0.

For the schedulers above, the proofs of their convergence or divergence are the same as Corollary F.6 with different constants. □

Now we study a case where the noise structure itself is enough to guarantee the convergence. In this case, v_t scales with H_t.

Corollary F.14 (SEG Insights)**.** *For $f(x,y) = x^\top \mathbb{E}_\xi\left[\mathbf{\Lambda}_\xi\right] y$ such that $\mathbf{\Lambda}_\xi$ is diagonal, we have*

1. *$\mathbb{E}\left[Z_t\right] = \tilde{\mathbf{E}}(t) \tilde{\mathbf{R}}(t) z \overset{t\to\infty}{=} 0$;*

2. $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} = \sum_{i=1}^{d} \frac{\|Z_0^i\|^2}{2} e^{-\left(2\rho\lambda_i^2 - \eta\sigma_i^2\left(1+\rho^2\left(2\lambda_1^2+\sigma_i^2\right)\right)\right)t}$.

In particular, $\frac{\mathbb{E}\left[\|Z_t\|^2\right]}{2} \to 0$ *if* $2\rho\lambda_i^2 - \eta\sigma_i^2\left(1+\rho^2\left(2\lambda_1^2+\sigma_i^2\right)\right) > 0, \forall i$.

Proof. The derivation of the SDE is straightforward and is

$$dZ_t = \mathbf{A}Z_t dt + \sqrt{\eta}\mathbf{B}dW_t, \tag{307}$$

where

$$\mathbf{A} = \begin{bmatrix} -\rho\mathbf{\Lambda}^2 & -\mathbf{\Lambda} \\ \mathbf{\Lambda} & -\rho\mathbf{\Lambda}^2 \end{bmatrix} \text{ and } \mathbf{B}\mathbf{B}^\top = \begin{bmatrix} \Sigma^2 & \Sigma^2 \\ \Sigma^2 & \Sigma^2 \end{bmatrix} \circ \begin{bmatrix} D_{1,1} & D_{1,2} \\ D_{2,1} & D_{2,2} \end{bmatrix}, \tag{308}$$

where

$$D_{1,1} := \operatorname{diag}((Y_t + \rho\mathbf{\Lambda}X_t) \circ (Y_t + \rho\mathbf{\Lambda}X_t) + \rho^2\Sigma^2 \circ (\mathbf{\Lambda}^2 + \Sigma^2) \circ Y_t \circ Y_t), \tag{309}$$

$$D_{2,2} := \operatorname{diag}((X_t - \rho\mathbf{\Lambda}Y_t) \circ (X_t - \rho\mathbf{\Lambda}Y_t) + \rho^2\Sigma^2 \circ (\mathbf{\Lambda}^2 + \Sigma^2) \circ X_t \circ X_t), \tag{310}$$

and $D_{1,2}$ and $D_{2,1}$ do not matter for this calculation.

It is easy to see that

$$\frac{\|Z_t\|^2}{2} = \sum_{i=1}^{d} \frac{\|Z_t^i\|^2}{2}, \tag{311}$$

where $Z^i := (X^i, Y^i)$, and that

$$d\left(\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2}\right) = -2\rho\lambda_i^2 \frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} dt + \eta\sigma_i^2(1+\rho(2\lambda_i^2+\sigma_i^2)) \frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} dt, \tag{312}$$

which ultimately implies that

$$\frac{\mathbb{E}\left[\|Z_t^i\|^2\right]}{2} = \frac{\|Z_0^i\|^2}{2} e^{-\left(2\rho\lambda_i^2 - \eta\sigma_i^2\left(1+\rho^2\left(2\lambda_1^2+\sigma_i^2\right)\right)\right)t}. \tag{313}$$

Empirical validation of this result is provided in Figure 7. □

G EXPERIMENTS

In this section, we provide the details of the experiments we carried out to validate the theoretical results derived in the paper. We highlight that since we always use diagonal matrices, it is enough to validate our results in two dimensions. In the following, the choice of the initialization points does not have a special reason. When there is one, it is explained in the respective paragraphs.

Computational Infrastructure All experiments have been performed on Google Colaboratory without any premium subscription. The code to replicate the experiments is available at `https://github.com/eneamc/MinimaxSDEs`

G.1 SDE Validation: Figure 1

In this subsection, we provide the details to replicate the experiments shown in Figure 1. The objective is to provide empirical validation to Theorem 3.4 and 3.6: The trajectories of the simulated SDEs match that of the respective algorithms averaged over 5 runs.

SGDA This paragraph refers to the *top left* of Figure 1. Inspired by **Example 5.2** in (Hsieh et al., 2019), we study **Nonbilinear Game # 1** $f(x,y) := x(y-0.45) + \phi(x) - \phi(y)$ where $\phi(z) := \frac{1}{4}z^2 - \frac{1}{2}z^4 + \frac{1}{6}z^6$. In the figure, we show the comparison between the average of 5 realizations of the trajectories of SGDA with the average of 5 simulations of the trajectories of the SDE of SGDA.

For each of the 5 trajectories of SGDA, we initialize each trajectory at $(x_0, y_0) = (2.0, 2.0)$ because this point is outside of the limit cycle that surrounds the optimal saddle point. We use a stepsize $\eta = 0.01$, and run the optimizer for $N = 10000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 1.0$. Each trajectory is run with a different random seed.

For each of the 5 trajectories of the SDE of SGDA, we initialize each trajectory at $(x_0, y_0) = (2.0, 2.0)$, use a discretization step $dt = \frac{\eta}{10} = 0.001$ for the Euler–Maruyama method, and integrate the system for $N = 100000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 1.0$. Each trajectory is run with a different random seed.

SHGD This paragraph refers to the *bottom left* of Figure 1. As a variation on **Example 5.1** and **Example 5.2** in (Hsieh et al., 2019), we study **Nonbilinear Game # 3** $f(x,y) := xy + \phi(x) - \phi(y)$ where $\phi(z) := \frac{1}{2}z^2 - \frac{1}{4}z^4 + \frac{1}{6}z^6 - \frac{1}{8}z^8$. In the figure, we show the comparison between the average of 5 realizations of the trajectories of SHGD with the average of 5 simulations of the trajectories of the SDE of SHGD.

For each of the 5 trajectories of SHGD, we initialize each trajectory at $(x_0, y_0) = (0.7, 0.7)$. Given the extreme nonlinearity of this landscape, this initial point allows the use of sizeable stepsizes: we use a stepsize $\eta = 0.0001$. Going further away from $(0, 0)$ would require extremely smaller stepsizes to avoid numerical instabilities. We run the optimizer for $N = 100000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 1.0$. Each trajectory is run with a different random seed.

For each of the 5 trajectories of the SDE of SHGD, we initialize each trajectory at $(x_0, y_0) = (0.7, 0.7)$, use a discretization step $dt = \frac{\eta}{10} = 0.00001$ for the Euler–Maruyama method, and integrate the system for $N = 1000000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 1.0$. Each trajectory is run with a different random seed.

SEG This paragraph refers to the *top right* and *bottom right* of Figure 1. Inspired by **Example 5.1** in (Hsieh et al., 2019), we study **Nonbilinear Game # 2** $f(x,y) := xy - \epsilon\phi(y)$ where $\phi(z) := \frac{1}{2}z^2 - \frac{1}{4}z^4$ and $\epsilon = 0.01$. For two different values of ρ, we show the comparison between the average of 5 realizations of the trajectories of SEG with the average of 5 simulations of the trajectories of the SDE of SEG. Additionally, we show the comparison with the average of 5 simulations of the trajectories of the SDE of SGDA.

For each $\rho \in \{0.1, 1\}$, we repeat the following procedure: For each of the 5 trajectories of SEG, we initialize each trajectory at $(x_0, y_0) = (1.0, 1.0)$, use a stepsize $\eta = 0.01$, and run the optimizer for $N = 10000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 1.0$. Each trajectory is run with a different random seed.

For each $\rho \in \{0.1, 1\}$, we repeat the following procedure: For each of the 5 trajectories of the SDE of SEG, we initialize each trajectory at $(x_0, y_0) = (1.0, 1.0)$, use a discretization step $dt = \frac{\eta}{10} = 0.001$ for the Euler–Maruyama method, and integrate the system for $N = 100000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 1.0$. Each trajectory is run with a different random seed.

For each of the 5 trajectories of the SDE of SGDA, we initialize each trajectory at $(x_0, y_0) = (1.0, 1.0)$, use a discretization step $dt = \frac{\eta}{10} = 0.001$ for the Euler–Maruyama method, and integrate the system for $N = 100000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 1.0$. Each trajectory is run with a different random seed.

For the *top right* figure, we plot the average of the trajectories. For the *bottom right* figure, we plot the average norm of the iterates $\mathbb{E}\left[\|x_k\|^2 + \|y_k\|^2\right]$ for the SEG and $\mathbb{E}\left[\|X_{\eta k}\|^2 + \|Y_{\eta k}\|^2\right]$ for the SDEs. We did not report the average norm of the SDE of SGDA as it diverges and would spoil the informativeness of the figure.

G.2 Schedulers Validation: Figure 3

In this subsection, we provide the details to replicate the experiments shown in Figure 3. The objective is to provide empirical validation to Prop. 4.4, Prop. 4.5, Prop. 4.9, and Prop. 4.10. Consistently with the assumptions

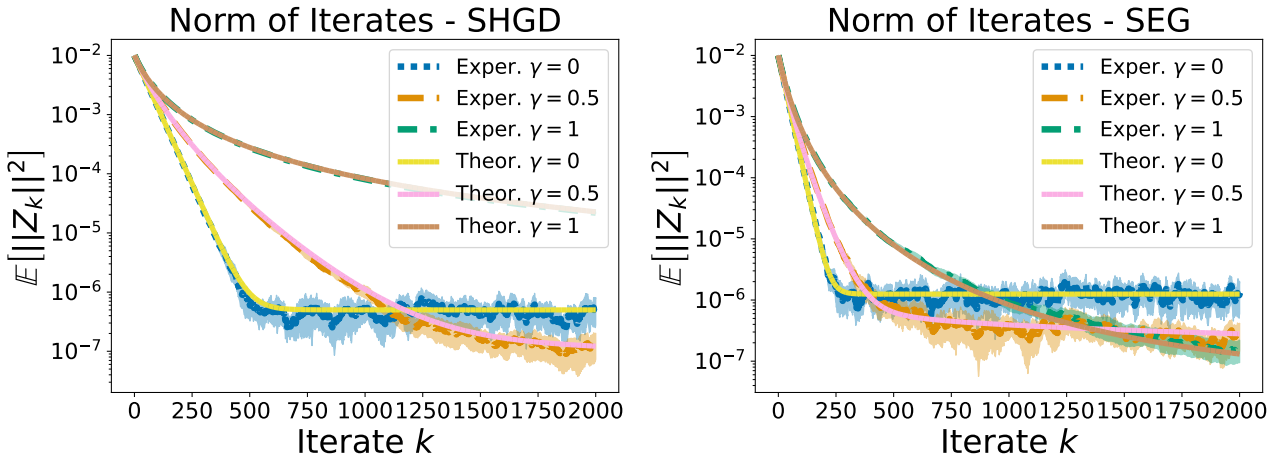

Figure 5: Empirical validation of Prop. 4.4 and Prop. 4.5 (Left); Prop. 4.9 and Prop. 4.10 (Right): The dynamics of $\mathbb{E}\left[\|Z_t\|^2\right]$ averaged across 5 runs perfectly matches that prescribed by our results for all schedulers. Both for SEG and SHGD, $\eta = 0.01$, while $\rho = 2$.

of these theorems, the landscape is that of the Bilinear Game $f(x, y) = 2xy$.

SHGD This paragraph refers to the *left* of Figure 3. As we use the stepsize scheduler $\eta_t := \frac{1}{(t+1)^\gamma}$, for $\gamma \in \{0, 0.5, 1.0\}$, we compare the average norm of the iterates across 5 realizations of the trajectories of SHGD with the exact dynamics of such a quantity prescribed in Prop. 4.4 and Prop. 4.5. For each of the 5 trajectories of SHGD, we initialize each trajectory at $(x_0, y_0) = (0.1, 0.1)$, use a stepsize $\eta = 0.01$, and run the optimizer for $N = 2000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 0.001$. Each trajectory is run with a different random seed.

The left of Figure 5 reports an additional experiment with the same setup apart from $f(x, y) = xy$, $\sigma = 0.01$, and $\rho = 2$.

SEG This paragraph refers to the *right* of Figure 3. As we use the stepsize scheduler $\eta_t := \frac{1}{(t+1)^\gamma}$, for $\gamma \in \{0, 0.5, 1.0\}$, we compare the average norm of the iterates across 5 realizations of the trajectories of SEG with the exact dynamics of such a quantity prescribed in Prop. 4.4 and Prop. 4.5. For each of the 5 trajectories of SEG, we initialize each trajectory at $(x_0, y_0) = (0.1, 0.1)$, use a stepsize $\eta = 0.01$, extra stepsize $\rho = 1.0$, and run the optimizer for $N = 2000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 0.001$. Each trajectory is run with a different random seed.

The right of Figure 5 reports an additional experiment with the same setup apart from $f(x, y) = xy$, $\sigma = 0.01$, and $\rho = 2$.

G.3 Role of ρ: *Left* of Figure 4

In this subsection, we provide the details to replicate the experiments shown on the *left* of Figure 4. The objective is to provide empirical validation to the insights derived in Paragraph E.1 and Paragraph E.2. Consistently with the assumptions, the landscape is that of the Quadratic Game $f(x, y) = \frac{3}{2}x^2 + xy - \frac{3}{2}y^2$.

Let us remember that ρ^{H} is meant to replicate the speed of the exponential decay of SHGD while ρ^{V} is meant to achieve the lowest possible asymptotic variance of SEG. Finally, this is a case where **negative** ρ converges to the optimum faster than SGDA and than any positive ρ. Of course, this particular choice confirms that large (absolute) values of ρ result in larger suboptimality. It is key to notice that we indeed verify all these insights clearly in this Figure.

The left of Figure 6 reports an additional experiment with the same setting but $f(x, y) = x^2 + 3xy - y^2$. In these cases, positive ρ are the ones inducing fast convergence.

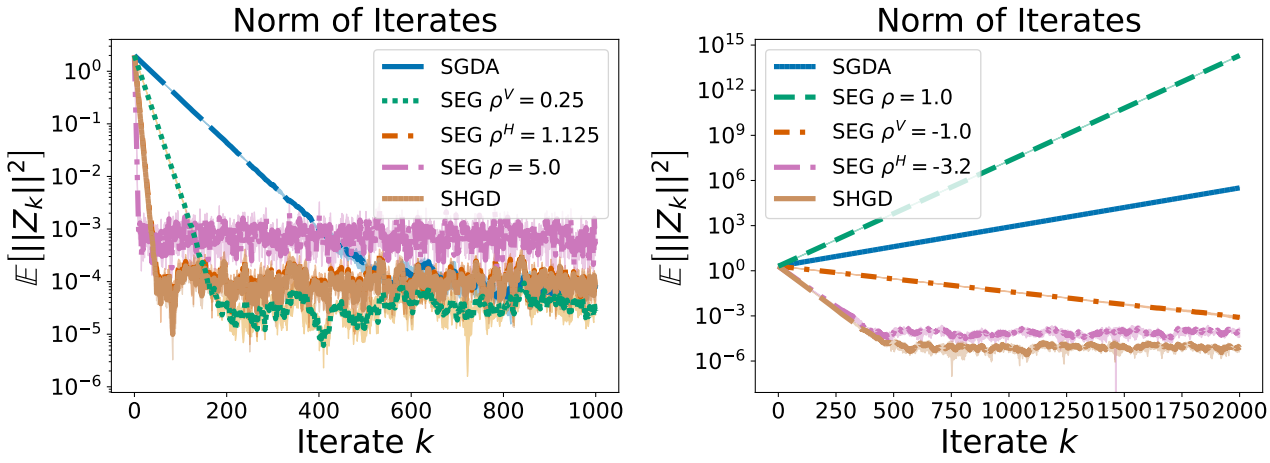

Figure 6: Empirical validation of the comparison between SEG and SHGD on Quadratic Games: (Left), ρ^V and ρ^H clearly meet the designated goals. Large $|\rho|$ induces faster convergence which in turn results in larger suboptimality. (Right), positive ρ escapes the *bad saddle* faster than SGDA, negative ones induce convergence, and ρ^H even matches the decay of SHGD. In both experiments, $\eta = 0.01$.

SHGD We plot the average norm of the iterates across 5 realizations of the trajectories of SHGD. For each of the 5 trajectories of SHGD, we initialize each trajectory at $(x_0, y_0) = (1.0, 1.0)$, use a stepsize $\eta = 0.01$, and run the optimizer for $N = 1000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 0.1$. Each trajectory is run with a different random seed.

SEG For different values of ρ, we repeat the following procedure: We plot the average norm of the iterates across 5 realizations of the trajectories of SEG. For each of the 5 trajectories of SEG, we initialize each trajectory at $(x_0, y_0) = (1.0, 1.0)$, use a stepsize $\eta = 0.01$, and run the optimizer for $N = 1000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 0.1$. Each trajectory is run with a different random seed. The values of ρ we used are $\rho = -5$, $\rho^{\mathrm{H}} = \frac{a^2+\lambda^2-a}{\lambda^2-a^2} = -0.875$, $\rho^{\mathrm{V}} = \frac{1}{a+\lambda} = \frac{1}{4}$, and $\rho = 0$ which corresponds to SGDA in the Figure.

G.4 Escape from *Bad Saddles*: *Right* of Figure 4

In this subsection, we provide the details to replicate the experiments shown on the *right* of Figure 4. The objective is to provide empirical validation to the insights derived in Paragraph E.1 and Paragraph E.2, with a special focus on the ability of SEG to escape *bad saddles* compared to SHGD that gets trapped. Consistently with the assumptions, the landscape is that of the Quadratic Game $f(x, y) = -\frac{1}{2}x^2 + 2xy + \frac{1}{2}y^2$. We indeed observe that consistently with the theory, SHGD is attracted by such undesirable saddle points while suitable choices of ρ allow SEG to escape the saddle. Interestingly, unfortunate choices of ρ replicate the regrettable behavior of SHGD.

The right of Figure 6 reports an additional experiment with the same setting but $f(x, y) = -3x^2 + 2xy + 3y^2$. In these cases, positive ρ are the ones inducing fast divergence while negative ones induce (undesirable) convergence to the saddle.

SHGD We plot the average norm of the iterates across 5 realizations of the trajectories of SHGD. For each of the 5 trajectories of SHGD, we initialize each trajectory at $(x_0, y_0) = (1.0, 1.0)$, use a stepsize $\eta = 0.001$, and run the optimizer for $N = 2000$ iterations. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 0.1$. Each trajectory is run with a different random seed.

SEG For different values of ρ, we repeat the following procedure: We plot the average norm of the iterates across 5 realizations of the trajectories of SEG. For each of the 5 trajectories of SEG, we initialize each trajectory at $(x_0, y_0) = (1.0, 1.0)$, use a stepsize $\eta = 0.001$, and run the optimizer for $N = 2000$ iterations. The noise used to

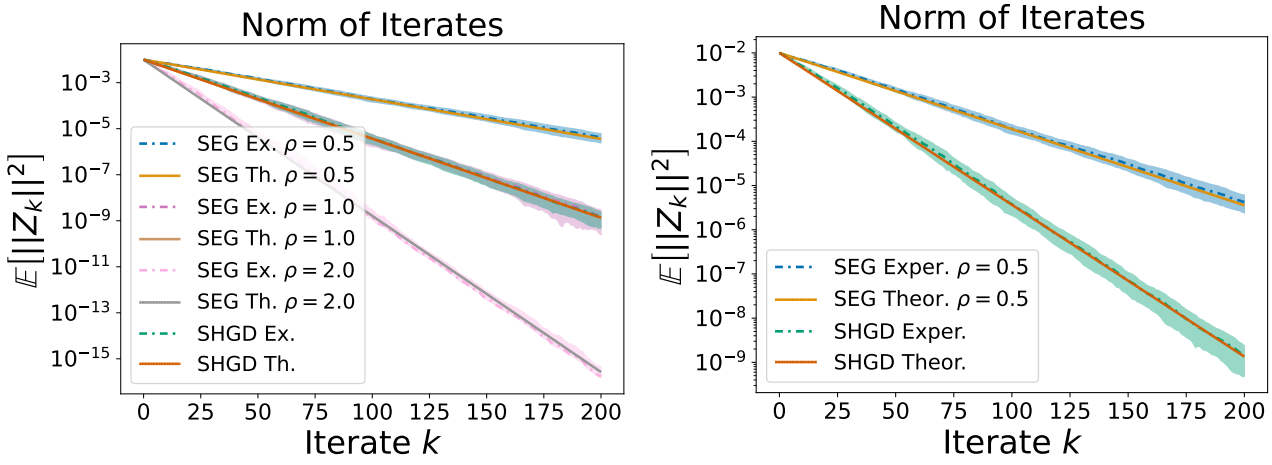

Figure 7: Empirical validation: Corollary F.7 and Corollary F.14 (Left) and detail of SEG vs SHGD when their convergence speed matches (Right).

perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$ and $\sigma = 0.1$. Each trajectory is run with a different random seed. The values of ρ we used are $\rho = -1$, $\rho^{\mathrm{H}} = \frac{a^2+\lambda^2-a}{\lambda^2-a^2} = 2$, $\rho^{\mathrm{V}} = \frac{1}{a+\lambda} = 1$, and $\rho = 0$ which corresponds to SGDA in the Figure.

G.5 Empirical Validation of Figure 7

In this subsection, we provide the details to replicate the experiments shown in Figure 7. The objective is to provide empirical validation to Corollary F.7 and Corollary F.14. Consistently with the assumptions, the landscape is that of the Stochastic Bilinear Game $f(x, y) = x^\top \mathbb{E}_\xi [\mathbf{\Lambda}_\xi] y$ such that $\mathbf{\Lambda}_\xi$, where $\mathbf{\Lambda} = 2\mathbf{I}_2$ and $\xi \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_2)$, and $\sigma = 1.0$. We indeed observe that the average behavior of the norm of the iterates of SEG and SHGD matches that prescribed by Corollary F.7 and Corollary F.14.

SHGD We compare the average norm of the iterates across 5 realizations of the trajectories of SHGD with the exact dynamics prescribed by Corollary F.7. For each of the 5 trajectories of SHGD, we initialize each trajectory at $(x_0, y_0) = (0.1, 0.1)$, use a stepsize $\eta = 0.01$, and run the optimizer for $N = 200$ iterations. Each trajectory is run with a different random seed.

SEG We compare the average norm of the iterates across 5 realizations of the trajectories of SEG with the exact dynamics prescribed by Corollary F.7. For each of the 5 trajectories of SHGD, we initialize each trajectory at $(x_0, y_0) = (0.1, 0.1)$, use a stepsize $\eta = 0.01$, extra stepsize $\rho \in \{0.5, 1, 2\}$, and run the optimizer for $N = 200$ iterations. Each trajectory is run with a different random seed.

G.6 Empirical Validation of Figure 8: The asymptotic variance of SEG is influenced by ρ

In this subsection, we provide the details to replicate the experiments shown in Figure 8. The objective is to provide empirical validation to Eq. 211. Consistently with the assumptions, the landscape is that of the Quadratic Game $f(x, y) = x^2 + xy - y^2$. We indeed observe that the experimental average asymptotic variance of SEG matches the one prescribed by Eq. 211.

For different values of ρ, we repeat the following procedure: We compare the average asymptotic variance of the iterates across 5 realizations of the trajectories of SEG with the exact formula prescribed by Eq. 211. For each of the 5 trajectories of SEG, we initialize each trajectory at $(x_0, y_0) = (0.01, 0.01)$, use a stepsize $\eta = 0.01$, and run the optimizer for $N = 200000$ iterations. Each trajectory is run with a different random seed. The values of ρ used are $\rho \in \left\{-\frac{1}{6}, 0, \frac{1}{6}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}\right\}$.

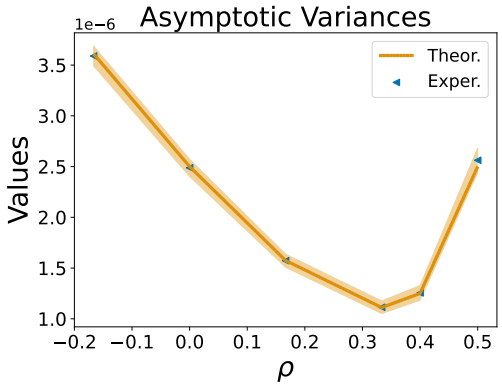

Figure 8: Empirical validation of Equation (211).

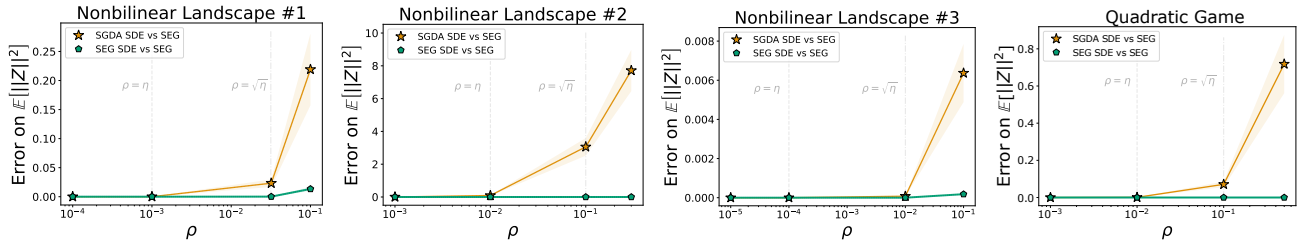

Figure 9: Comparison in terms of $\mathbb{E}\left[\|z\|^2\right]$ with respect to ρ - Nonbilinear Game # 1 (Left); Nonbilinear Game # 2 (Center Left); Nonbilinear Game # 3 (Center Right); Quadratic Game (Right).

G.7 Empirical Validation of SDEs: Figure 9

In this subsection, we provide the details to replicate the experiments shown in Figure 9. The objective is to show that if $\rho = \mathcal{O}(\eta)$ or even smaller, the SDE of SGDA models the dynamics of SEG accurately. However, once $\rho = \mathcal{O}(\sqrt{\eta})$ or even larger, the SDE of SGDA no longer models the dynamics of SEG correctly while the SDE of SEG does so. To simulate the SDEs, we use Algorithm 1.

Nonbilinear Game # 1 In this paragraph, we provide the details of the Nonbilinear Game # 1 experiment. We optimize the loss function $f(x, y) := x(y - 0.45) + \phi(x) - \phi(y)$ where $\phi(z) := \frac{1}{4}z^2 - \frac{1}{2}z^4 + \frac{1}{6}z^6$. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2\mathbf{I}_2)$ and $\sigma = 1.00$. We use $\eta = 0.001$, $\rho \in \{0.0001, 0.001, 0.0316, 0.1\}$. The results are averaged over 5 experiments.

Nonbilinear Game # 2 In this paragraph, we provide the details of the Nonbilinear Game # 2 experiment. We optimize the loss function $f(x, y) := xy - \epsilon\phi(y)$ where $\phi(z) := \frac{1}{2}z^2 - \frac{1}{4}z^4$ and $\epsilon = 0.01$. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2\mathbf{I}_2)$ and $\sigma = 1.00$. We use $\eta = 0.01$, $\rho \in \{0.001, 0.01, 0.01, 0.3\}$. The results are averaged over 5 experiments.

Nonbilinear Game # 3 In this paragraph, we provide the details of the Nonbilinear Game # 3 experiment. We optimize the loss function $f(x, y) := xy + \phi(x) - \phi(y)$ where $\phi(z) := \frac{1}{2}z^2 - \frac{1}{4}z^4 + \frac{1}{6}z^6 - \frac{1}{8}z^8$. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2\mathbf{I}_2)$ and $\sigma = 1.00$. We use $\eta = 0.0001$, $\rho \in \{0.00001, 0.0001, 0.01, 0.1\}$. The results are averaged over 5 experiments.

Quadratic Game In this paragraph, we provide the details of the Quadratic Game experiment. We optimize the loss function $f(x, y) := x^2 + 2xy - y^2$. The noise used to perturb the gradients is $Z \sim \mathcal{N}(0, \sigma^2\mathbf{I}_2)$ and $\sigma = 1.00$. We use $\eta = 0.01$, $\rho \in \{0.001, 0.01, 0.1, 0.5\}$. The results are averaged over 5 experiments.