Last-Iterate Convergence of Saddle Point Optimizers via High-Resolution Differential Equations

Tatjana Chavdarova TATJANA.CHAVDAROVA@BERKELEY.EDU
Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Manolis Zampetakis MZAMPET@BERKELEY.EDU
University of California, Berkeley

Abstract

Several widely-used first-order saddle point optimization methods yield the same continuous-time ordinary differential equation (ODE) as that of the Gradient Descent Ascent (GDA) method when derived naively. However, their convergence properties are very different even in simple bilinear games. In this paper, we use a technique from fluid dynamics called High Resolution Differential Equations (HRDEs) to design ODEs that differentiate between popular saddle point optimizers. As our main result, we design an ODE that has last iterate convergence for monotone variational inequalities. To our knowledge, this is the first continuous-time dynamics shown to converge for such a general setting. We also provide an implicit discretization of our ODE and we show it has last iterate convergence at a rate $\mathcal{O}(1/\sqrt{T})$, which was previously shown to be optimal [Golowich et al., 2020], using *only* first-order smoothness of the monotone operator, in contrast to previous results that need second-order smoothness as well.

1. Introduction

We are interested in the convergence of optimization methods for zero-sum games, where agents $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$ are given a loss $f\colon \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ which the first agent aims to minimize and the second agent aims to maximize. Training accordingly aims at finding a *saddle point* $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$ of f:

$$f(\boldsymbol{x}^\star, \boldsymbol{y}) \leq f(\boldsymbol{x}^\star, \boldsymbol{y}^\star) \leq f(\boldsymbol{x}, \boldsymbol{y}^\star) \qquad \forall \boldsymbol{x} \in \mathcal{X}, \quad \forall \boldsymbol{y} \in \mathcal{Y}\,. \tag{SP}$$

This fundamental problem arises in various domains—such as optimization, economics, and multi-agent reinforcement learning [34]—and has recently seen renewed interest in the context of training Generative Adversarial Networks [17]. The difference of two-objective min-max training from single-objective minimization is that In contrast to minimization, in SP problems the second-order derivative matrix (defined in § 2.1) is non-symmetric in the general case, resulting in dynamics that *can* rotate around a fixed point. For example, on the simple bilinear game (BG)—which is convex in $\boldsymbol{x}$ and concave in $\boldsymbol{y}$—the last iterate of the simplest gradient descent ascent (GDA) method is oscillating around the solution for infinitesimal step size, and *diverging away* from it otherwise [10] (see also § 3.1). For this reason designing algorithms with *last-iterate* convergence has attracted significant attention [2, 7, 9, 10, 15, 16, 25, 26, 28]. Several works aim to resolve this problem, e.g., the extragradient method [EG, 21], the optimistic gradient descent ascent [OGDA, 37], and more recently the lookahead-minmax [LA, 8]. Despite their different behaviours on BG, *all* of these popular first-order min-max optimizaers–GDA, EG, OGDA, and LA–lead to the *same* system of ordinary differential equations (ODEs) in the limit when the step size goes to zero [19] (see also § 2.3), which is surprisingly *non* convergent on BG. In this paper, we apply a technique from fluid dynamics, called *High Resolution Differential Equations* (HRDEs), proposed for single objective optimization by [39]. Using this technique we derive HRDEs that differ among GDA, EG, and LA and enjoy

geometric convergence on bilinear games. Our main result shows that our HRDE for OGDA is guaranteed to converge even for general problems of monotone variational inequalities (MVI, defined in § 2) that include convex-concave min-max optimization (see §4). To our knowledge, this is the first system of differential equations that provably converges for all MVI problems.

Related works. The SP optimization problem for a convex-concave f dates to the 1960s [22, 23]. The *averaged* (ergodic) iterate of both the EG and OGDA methods achieve the optimal rate of $\mathcal{O}(\frac{1}{T})$ on general MVI problems [20, 29, 31, 40, 41]. Several papers prove the last-iterate convergence on bilinear or *strongly* monotone games: *(i)* Daskalakis et al. [10] prove last-iterate convergence of *optimistic mirror decent* (OMD) on BG, *(ii)* Azizian et al. [5] establish linear convergence on bilinear and strongly monotone games, *(iii)* [1, 24] focus on Hamiltonian gradient descent. Golowich et al. [15] prove the last-iterate convergence of EG at a rate $\mathcal{O}(\frac{1}{\sqrt{T}})$ on the more general problem of monotone VI, while relaying on the associated operator having a Λ–Lipschitz derivative [see Assumption 2 in 15]. Lyapunov stability theory is widely used in optimization. These tools are particularly attractive because they can be used for non-linear systems, and if a Lyapunov function can be found, the convergence result holds *globally*. Lastly, HRDEs were introduced in optimization by [39] in the context of minimization, to distinguish Nesterov's accelerated [32, 33] from Polyak's heavy-ball [36] method. See App. B for more elaborate overview of related works.

Contributions. We derive the HRDEs of GDA, EG, OGDA and LA, summarized in Table 1. Using these HRDEs, on bilinear games (BG), we then prove: *(i)* the divergence of GDA, *(ii)* the geometric convergence of EG and OGDA, as well as *(iii)* for the first time the convergence of LA when combined with the diverging GDA with 2 and 3 steps. Relative to ODE based analyses, these HRDE based results are more aligned with the methods' observed performances in practice when considering such simple games.

Informal Theorem *[see Theorem 4 & 7] The HRDE of the OGDA method has last iterate convergence for monotone VI problems. Additionally, there exists an implicit discretization of it that has $\mathcal{O}(\frac{1}{\sqrt{T}})$ last iterate convergence rate on MVI problems satisfying solely first order smoothness.*

We show that a simple discretization of the derived HRDE of OGDA, yields the original OGDA method. Using analogous technique as for the above result, we show the following result for OGDA.

Informal Theorem *[see Theorem 6] The best-iterate of the discrete-time OGDA converges with rate $O(1/\sqrt{t})$ for all MVI problems. Additionally, OGDA admits asymptotic last-iterate convergence for all MVI.*

Finally, we also provide the last-iterate convergence rate of an implicit discretization of our HRDE of OGDA in Theorem 7. Our results are particularly noteworthy in that they provide: *(i)* the first convergence proof in continuous time for general MVI problems, and *(ii)* more broadly the only proof that does not rely on second-order smoothness of the associated operator and does not use averaging; see Table 1.

2. Preliminaries

Problem. We are interested in the unconstrained zero-sum game (see App. A for notation):

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d_1}} \max_{\boldsymbol{y}\in\mathbb{R}^{d_2}} f(\boldsymbol{x},\boldsymbol{y})\,, \tag{ZS-G}$$

where $f : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is smooth and convex in $\boldsymbol{x}$ and concave in $\boldsymbol{y}$. We write $\boldsymbol{z} \triangleq (\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^d$ where $d = d_1 + d_2$.

Method	**High-Resolution Differential Equation** with $\dot{\boldsymbol{z}}(t)=\boldsymbol{\omega}(t)$, $\beta=2/(\text{step size})$, α is a hyperparameter of LA.
GDA	$\dot{\boldsymbol{\omega}}(t)=-\beta\cdot\boldsymbol{\omega}(t)-\beta\cdot V(\boldsymbol{z}(t))$
EG	$\dot{\boldsymbol{\omega}}(t)=-\beta\cdot\boldsymbol{\omega}(t)-\beta\cdot V(\boldsymbol{z}(t))+2\cdot J(\boldsymbol{z}(t))\cdot V(\boldsymbol{z}(t))$
OGDA	$\dot{\boldsymbol{\omega}}(t)=-\beta\cdot\boldsymbol{\omega}(t)-\beta\cdot V(\boldsymbol{z}(t))-2\cdot J(\boldsymbol{z}(t))\cdot\boldsymbol{\omega}(t)$
LA2-GDA	$\dot{\boldsymbol{\omega}}(t)=-\beta\cdot\boldsymbol{\omega}(t)-2\alpha\beta\cdot V(\boldsymbol{z}(t))+2\alpha\cdot J(\boldsymbol{z}(t))\cdot V(\boldsymbol{z}(t))$
LA3-GDA	$\dot{\boldsymbol{\omega}}(t)=-\beta\cdot\boldsymbol{\omega}(t)-2\alpha\beta\cdot V(\boldsymbol{z}(t))+2\alpha\cdot J(\boldsymbol{z}(t))\cdot V(\boldsymbol{z}(t))$ $-\alpha\beta\cdot V\big[\boldsymbol{z}(t)-4\beta\cdot V\big(\boldsymbol{z}(t)\big)\big]$

Smoothness assumptions	Discrete time	Continuous time
Last-iterate		
1st & 2nd order	EG [15]	⁄
1st–order	implicit OGDA, Thm. 7	OGDA-HRDE, Thm. 4
Best-iterate		
1st–order	OGDA, Thm. 6	⁄

Table 1: **Left.** Derived HRDEs for several saddle point optimization methods. LAk-GDA denotes LA with k steps, and GDA as its base optimizer. The map $V(\cdot)$ denotes the vector field, i.e., $(-\nabla_x f, \nabla_y f)$, whereas J is the Jacobian. **Right.** List of *last* and *best*-iterate convergence results on the general *monotone variational inequality* problem, with rate $\mathcal{O}(1/\sqrt{T})$.

2.1. Saddle-point optimization methods

Gradient Descent Ascent (GDA). The GDA operator $V:\mathbb{R}^d\to\mathbb{R}^d$ and its Jacobian J are:

$$V(\boldsymbol{z})=\begin{bmatrix}\nabla_x f(\boldsymbol{z})\\ -\nabla_y f(\boldsymbol{z})\end{bmatrix}\qquad J(\boldsymbol{z})=\begin{bmatrix}\nabla_x^2 f(\boldsymbol{z}) & \nabla_y\nabla_x f(\boldsymbol{z})\\ -\nabla_x\nabla_y f(\boldsymbol{z}) & -\nabla_y^2 f(\boldsymbol{z})\end{bmatrix}.$$

By denoting the step size with $\gamma\in[0,1]$, each GDA update at step n is then:

$$\boldsymbol{z}_{n+1}=\boldsymbol{z}_n-\gamma V(\boldsymbol{z}_n)\,. \tag{GDA}$$

Extra Gradient (EG [21]). Uses a "prediction" step to obtain an extrapolated point $\boldsymbol{z}_{t+\frac{1}{2}}\triangleq(\boldsymbol{x}_{n+\frac{1}{2}},\boldsymbol{y}_{n+\frac{1}{2}})$ using GDA: $\boldsymbol{z}_{n+\frac{1}{2}}=\boldsymbol{z}_n-\gamma V(\boldsymbol{z}_n)$–where $\gamma\in[0,1]$ denotes the step size, and the gradients at the *extrapolated* point are then applied to the current iterate $\boldsymbol{z}_t\triangleq(\boldsymbol{x}_t,\boldsymbol{y}_t)$ as follows:

$$\boldsymbol{z}_{n+1}=\boldsymbol{z}_n-\gamma V(\boldsymbol{z}_{n+\frac{1}{2}})=\boldsymbol{z}_n-\gamma V(\boldsymbol{z}_n-\gamma V(\boldsymbol{z}_n))\,. \tag{EG}$$

Optimistic Gradient Descent Ascent (OGDA). The update rule of OGDA [37] is:

$$\boldsymbol{z}_{n+1}=\boldsymbol{z}_n-2\gamma V(\boldsymbol{z}_n)+\gamma V(\boldsymbol{z}_{n-1})\,. \tag{OGDA}$$

Lookahead–Minmax (LA [8, 45]). At each step t: *(i)* a copy of the current iterate $\tilde{\boldsymbol{z}}_n$ is made: $\tilde{\boldsymbol{z}}_n\leftarrow\boldsymbol{z}_n$, *(ii)* $\tilde{\boldsymbol{z}}_n$ is then updated for $k\geq 1$ times yielding $\tilde{\boldsymbol{\omega}}_{n+k}$, and finally:

$$\boldsymbol{z}_{n+1}\leftarrow\boldsymbol{z}_n+\alpha(\tilde{\boldsymbol{z}}_{n+k}-\boldsymbol{z}_n),\qquad\text{where }\alpha\in[0,1]\,. \tag{LA}$$

2.2. Assumptions

Definition 1 *An operator $V:\mathbb{R}^d\to\mathbb{R}^d$ is monotone iff: $\langle\boldsymbol{z}-\boldsymbol{z}',V(\boldsymbol{z})-V(\boldsymbol{z}')\rangle\geq 0,\quad\forall\boldsymbol{z},\boldsymbol{z}'\in\mathbb{R}^d$. Also, V is μ-strongly monotone iff: $\langle\boldsymbol{z}-\boldsymbol{z}',V(\boldsymbol{z})-V(\boldsymbol{z}')\rangle\geq\mu\|\boldsymbol{z}-\boldsymbol{z}'\|^2$ for all $\boldsymbol{z},\boldsymbol{z}'\in\mathbb{R}^d$.*

Assumption 1 (First-order Smoothness) *Let $V : \mathbb{R}^d \to \mathbb{R}^d$ be an operator, we say that V satisfies L-first order smoothness, or just L-smoothness, if V is a L-Lipschitz function.*

Assumption 2 (Second-order Smoothness) *Let $V : \mathbb{R}^d \to \mathbb{R}^d$ be an operator and $J : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ be its Jacobian, we say that V satisfies L_2-second-order smoothness, if: J is a L_2-Lipschitz function, i.e., $\|J(\boldsymbol{z}) - J(\boldsymbol{z}')\|_F \leq L_2 \cdot \|\boldsymbol{z} - \boldsymbol{z}'\|_2$.*

The corresponding *variational inequality* (VI) problem is:

$$\text{find a point } \boldsymbol{z}^\star \quad \text{s.t.} \qquad \langle \boldsymbol{z} - \boldsymbol{z}^\star, V(\boldsymbol{z})\rangle \geq 0, \quad \forall \boldsymbol{z} \in \mathbb{R}^d\,. \tag{VI}$$

Throughout the paper we assume without loss of generality that $\boldsymbol{z}^\star = 0$, thus: $\langle V(\boldsymbol{z}), \boldsymbol{z}\rangle \geq 0, \forall \boldsymbol{z} \in \mathbb{R}^d$. Finally, one well known implication of Definition 1 is that: $J(\boldsymbol{z}) \succeq 0, \quad \forall \boldsymbol{z} \in \mathbb{R}^d$.

2.3. ODE representations

We introduce the ansatz $\boldsymbol{z}_n \approx \boldsymbol{z}(n \cdot \delta)$, for smooth curve $\boldsymbol{z}(t)$ defined for $t \geq 0$. Taylor expansion over time with time-step δ gives: $\boldsymbol{z}_{n+1} \approx \boldsymbol{z}((n+1)\delta) = \boldsymbol{z}(n\delta) + \dot{\boldsymbol{z}}(n\delta)\delta + \frac{1}{2}\ddot{\boldsymbol{z}}(n\delta)\delta^2 + \dots$, thus:

$$\boldsymbol{z}_{n+1} - \boldsymbol{z}_n \approx \dot{\boldsymbol{z}}(n\delta)\delta + \frac{1}{2}\ddot{\boldsymbol{z}}(n\delta)\delta^2 + \mathcal{O}(\delta^3)\,. \tag{1}$$

GDA. Combining GDA and (1) and setting $\delta = \gamma \to 0$ yields the classical dynamics:

$$\dot{\boldsymbol{z}}(t) = -V(\boldsymbol{z}(t)). \tag{GDA-ODE}$$

EG, OGDA, LA2-GDA, LA3-GDA. Combining the methods' discrete-time dynamics with (1) and setting $\delta = \gamma \to 0$, we recover the same (GDA-ODE), and we notice that the final step $\delta, \gamma \mapsto 0$ causes this. This raises the question if HRDEs can model better the differently performing methods.

3. High-resolution differential equations of saddle point optimizers

GDA. Combining (GDA) with (1) we get : $\frac{\dot{\boldsymbol{z}}(n\cdot\delta)\delta + \frac{1}{2}\ddot{\boldsymbol{z}}(n\delta)\delta^2}{\gamma} = -V(\boldsymbol{z}(n\delta))$. By setting $\delta = \gamma$ and solely $\gamma^2 \to 0$ we have that: $\dot{\boldsymbol{z}}(t) + \frac{\gamma}{2}\ddot{\boldsymbol{z}}(t) = -V(\boldsymbol{z}(t))$, and by denoting $\beta = 2/\gamma$:

$$\dot{\boldsymbol{z}}(t) = \boldsymbol{\omega}(t)\,, \qquad \dot{\boldsymbol{\omega}}(t) = -\beta \cdot \boldsymbol{\omega}(t) - \beta \cdot V(\boldsymbol{z}(t)). \tag{GDA-HRDE}$$

EG. Combining (EG) with (1) we have: $\frac{\dot{\boldsymbol{z}}(n\delta)\delta + \frac{1}{2}\ddot{\boldsymbol{z}}(n\delta)\delta^2}{\gamma} = -V(\boldsymbol{z}(n\delta)) + \gamma J(\boldsymbol{z}(n\delta))V(\boldsymbol{z}(n\delta)) + \mathcal{O}(\gamma^2)$. Setting $\delta = \gamma$ and only $\gamma^2 \mapsto 0$ gives: $\dot{\boldsymbol{z}}(t) + \frac{\gamma}{2}\ddot{\boldsymbol{z}}(t) = -V(\boldsymbol{z}(t)) + \gamma \cdot J(\boldsymbol{z}(t)) \cdot V(\boldsymbol{z}(t))$:

$$\dot{\boldsymbol{z}}(t) = \boldsymbol{\omega}(t)\,, \qquad \dot{\boldsymbol{\omega}}(t) = -\beta \cdot \boldsymbol{\omega}(t) - \beta \cdot V(\boldsymbol{z}(t)) + 2 \cdot J(\boldsymbol{z}(t)) \cdot V(\boldsymbol{z}(t)). \tag{EG-HRDE}$$

OGDA. Combining (OGDA) with (1) we have: $\frac{\dot{\boldsymbol{z}}(n\delta)\delta + \frac{1}{2}\ddot{\boldsymbol{z}}(n\delta)\delta^2}{\gamma} = -V(\boldsymbol{z}(n\delta)) - \delta J\big(\boldsymbol{z}(n\delta)\big)\dot{\boldsymbol{z}}(n\delta)$. Then, setting $\delta = \gamma$ and keeping the $\mathcal{O}(\gamma)$ terms we have that: $\dot{\boldsymbol{z}}(t) + \frac{\gamma}{2}\ddot{\boldsymbol{z}}(t) = -V(\boldsymbol{z}(t)) - \gamma J(\boldsymbol{z}(t))\dot{\boldsymbol{z}}(t)$, yielding the following first-order system of HRDE in the phase-space representation:

$$\dot{\boldsymbol{z}}(t) = \boldsymbol{\omega}(t)\,, \quad \dot{\boldsymbol{\omega}}(t) = -\beta \cdot \boldsymbol{\omega}(t) - \beta \cdot V(\boldsymbol{z}(t)) - 2 \cdot J(\boldsymbol{z}(t)) \cdot \boldsymbol{\omega}(t). \tag{OGDA-HRDE}$$

LA2-GDA. Combining the equation LA2-GDA with (1) we have: $\frac{\dot{\boldsymbol{z}}(n\delta)\delta+\frac{1}{2}\ddot{\boldsymbol{z}}(n\delta)\delta^2}{\gamma} = -2\alpha V(\boldsymbol{z}(n\delta))-\alpha\gamma J(\boldsymbol{z}(n\delta))V(\boldsymbol{z}(n\delta))$. Similarly, we get: $\dot{\boldsymbol{z}}(t)+\frac{\gamma}{2}\ddot{\boldsymbol{z}}(t) = -2\alpha V(\boldsymbol{z}(t)) - \alpha\gamma J(\boldsymbol{z}(t))V(\boldsymbol{z}(t))$:

$$\dot{\boldsymbol{z}}(t) = \boldsymbol{\omega}(t)\,, \quad \dot{\boldsymbol{\omega}}(t) = -\beta\boldsymbol{\omega} - 2\alpha\beta\cdot V(\boldsymbol{z}(t)) + 2\alpha J(\boldsymbol{z}(t))V(\boldsymbol{z}(t)). \tag{LA2-GDA-HRDE}$$

LA3-GDA. Analogously as above, for LA3-GDA we have $\dot{\boldsymbol{z}}(t) = \boldsymbol{\omega}(t)$ and:

$$\dot{\boldsymbol{\omega}}(t) = -\frac{2}{\gamma}\boldsymbol{\omega}(t) - \frac{4\alpha}{\gamma}V(\boldsymbol{z}(t)) + 2\alpha\cdot J(\boldsymbol{z}(t))\cdot V(\boldsymbol{z}(t)) - \frac{2\alpha}{\gamma}V\big[\boldsymbol{z}(t) - 2\gamma V\big(\boldsymbol{z}(t)\big)\big] \tag{LA3-GDA-HRDE}$$

Proposition 2 (unique solution of the HRDEs) *The derived HRDEs have a unique solution when applied to bilinear games. Additionally, for any monotone map $V:\mathbb{R}^d\mapsto\mathbb{R}^d$ that satisfies the first and second-order smoothness as per Assumptions 1 & 2, the HRDE of OGDA has a unique solution.*

3.1. Convergence analysis of HRDEs on bilinear games

We consider the following bilinear game, with full rank $\boldsymbol{A}\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$: $\min_{\boldsymbol{x}\in\mathbb{R}^{d_1}}\max_{\boldsymbol{y}\in\mathbb{R}^{d_2}}\boldsymbol{x}^\intercal\boldsymbol{A}\boldsymbol{y}$.
Thm 9-13 summary. For *any* $\gamma>0$, the continuous-time dynamics GDA-HRDE *diverges* on BG. For sufficiently small step-size, the corresponding HRDEs of EG, OGDA and LA with $k=\{2,3\}$ converge on BG. The proofs use the Routh–Hurwitz [43] stability criteria for linear systems, see D.

4. Last iterate convergence for Monotone Variational Inequalities

To show convergence in continuous and discrete time we use Lyapunov functions, which intuitively can be seen as an energy function. For our purposes, we can guarantee convergence if we can find such a function that decreases along all possible trajectories of the described dynamical system. The corresponding proofs of the results in this section are provided in App. E.

4.1. Equivalent forms of OGDA-HRDE for MVI problems

The system of differential equations (OGDA-HRDE) is known in Physics for the special case when $V(\boldsymbol{z})=\nabla\phi(\boldsymbol{z})$, where ϕ is a potential function. The solution $\boldsymbol{z}(t)$ in this case describes the position of a system that is affected by non-elastic shocks described by the potential ϕ, see [4]. Using the techniques from [4] we can show that (OGDA-HRDE) has an equivalent formulation that does not involve the Jacobian $J(\cdot)$ and hence it is more suitable for designing discrete-time algorithms.

Theorem 3 *Let V be a continuously differentiable map, and $\boldsymbol{z}_0,\boldsymbol{\omega}_0,\boldsymbol{w}_0\in\mathbb{R}^d$, such that $\boldsymbol{w}_0 = -\frac{\beta}{2}\boldsymbol{\omega}_0 - \frac{4}{\beta}V(\boldsymbol{z}_0) - \boldsymbol{z}_0$, then these statements describe the same function $\boldsymbol{z}(t)$:*
1. the tuple $(\boldsymbol{z}(t),\boldsymbol{\omega}(t))$ is a solution of the equation (OGDA-HRDE) with initial conditions $\boldsymbol{z}(0)=\boldsymbol{z}_0$ and $\boldsymbol{\omega}(0)=\boldsymbol{\omega}_0$,
2. the tuple $(\boldsymbol{z}(t),\boldsymbol{w}(t))$ is a solution of the following system of equations

$$\begin{aligned}\dot{\boldsymbol{z}}(t) &= -\kappa\cdot\boldsymbol{z}(t) - \kappa\cdot\boldsymbol{w}(t) - 2V(\boldsymbol{z}(t))\\ \dot{\boldsymbol{w}}(t) &= -\kappa\cdot\boldsymbol{z}(t) - \kappa\cdot\boldsymbol{w}(t)\end{aligned} \tag{OGDA-HRDE-2}$$

with $\kappa=\beta/2$ and $\boldsymbol{z}(0)=\boldsymbol{z}_0$ and $\boldsymbol{w}(0)=\boldsymbol{w}_0$.

Note that is has to hold that V is continuously differentiable, and otherwise (OGDA-HRDE) and (OGDA-HRDE-2) are *not* necessarily equivalent. Thus, in the next section we show that *both* of them have last-iterate convergence for any MVI.

4.2. Last-iterate Convergence of OGDA-HRDE for MVIs

Theorem 4 *If we apply OGDA-HRDE to the MVI problem (VI) with initialization* $\boldsymbol{z}(0) = \boldsymbol{z}_0$, $\boldsymbol{\omega}(0)=0$, *we have that* $\|V(\boldsymbol{z}(t))\|_2 \leq O\left(\sqrt{\beta + \frac{L^2}{\beta}}\right) \cdot \frac{\|\boldsymbol{z}_0\|_2}{\sqrt{t}}$. *Moreover, if the variational inequality (VI) is* μ*-strongly monotone, as per Definition 1, then* $\|\boldsymbol{z}(t)\|_2 \leq O\left(\sqrt{\frac{\beta^2+L^2}{\beta^3}}\right) \cdot \|\boldsymbol{z}_0\|_2 \cdot \exp\left(-\rho \cdot t\right)$, *where* $1/\rho = \frac{1}{\mu} + \frac{9}{2\beta}$.

The strongly monotone guarantee is stronger in two ways: *(i)* the rate is geometric $\exp(-\rho \cdot t)$ instead of $1/\sqrt{t}$, *(ii)* we can bound the $\|Z(t)\|_2$ which implies a bound on $\|V(Z(t))\|_2$, due to the Lipschitzness of V, but not vice versa. We now show the convergence of (OGDA-HRDE-2).

Theorem 5 *If we apply (OGDA-HRDE-2) to the MVI problem (VI) with initialization* $\boldsymbol{z}(0) = \boldsymbol{z}_0$, $\boldsymbol{w}(0) = -\boldsymbol{z}_0$ *then we have that* $\|V(\boldsymbol{z}(t))\|_2 \leq O\left(\sqrt{\kappa + \frac{L^2}{\kappa}}\right) \cdot \frac{\|\boldsymbol{z}_0\|_2}{\sqrt{t}}$. *Moreover, if the variational inequality (VI) is* μ*-strongly monotone, as per Definition 1, then* $\|\boldsymbol{z}(t)\|_2 \leq O\left(\sqrt{\frac{\kappa^2+L^2}{\kappa^3}}\right) \cdot \|\boldsymbol{z}_0\|_2 \cdot \exp\left(-\rho \cdot t\right)$, *where* $1/\rho = O(\frac{1}{\mu} + \frac{1}{\kappa})$.

4.3. Best-iterate convergence of OGDA for MVIs

Primarily, observe that we can identify the *original* OGDA method as an *explicit* discretization of the (OGDA-HRDE-2) system of differential equations, see App. E.1.2. Since (OGDA) is strongly related with (OGDA-HRDE-2), it is natural to consider that the Lyapunov functions used to prove Thm. 5 would be useful to understand the convergence of (OGDA). Indeed, we can show the following.

Theorem 6 *If we apply the (OGDA) dynamics to the monotone variational inequality problem (VI) with* L*-Lipschitz map* V *and with initialization* $\boldsymbol{z}_1 = \boldsymbol{z}_0$ *and step size* $\gamma \leq \frac{1}{16 \cdot L}$, *then we have that* $\min_{i \in [n]} \|V(\boldsymbol{z}_i)\|_2 \leq O\left(\sqrt{\frac{1}{\gamma^2} + L^2} \cdot \frac{\|\boldsymbol{z}_0\|_2}{\sqrt{n}}\right)$. *We also have that* $\lim_{n \to \infty} \|V(\boldsymbol{z}_n)\|_2 = 0$.

The above theorem establishes that: *(i)* asymptotically the (OGDA) method converges to the solution of (VI), and *(ii)* the best iterate of the (OGDA) has similar behavior with the last-iterate convergence of EG shown in [15], the latter was shown *without* using second-order smoothness. Both of these results are new convergence properties of (OGDA) for the general case of MVIs.

4.4. Last iterate convergence of an implicit discretization of HR-OGDA

We analyze the following implicit discretization of OGDA-HRDE–see App. E.5 for its derivation:

$$\begin{aligned} \boldsymbol{z}_{n+1} &= \boldsymbol{z}_n + \frac{\gamma}{2} \cdot (\boldsymbol{\omega}_{n+1} + \boldsymbol{\omega}_n) \\ \boldsymbol{\omega}_{n+1} &= -V(\boldsymbol{z}_{n+1}) - \frac{1}{2}\left(V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\right) \end{aligned} \tag{OGDA-I}$$

Theorem 7 *If we apply the implicit discretized OGDA dynamics (OGDA-I) to the monotone variational inequality problem (VI) with initialization* $\boldsymbol{z}_0$, *and* $\boldsymbol{\omega}_0 = 0$ *then we have that* $\|V(\boldsymbol{z}_n)\|_2 \leq \mathcal{O}\left(\sqrt{\beta + \frac{L}{\beta}}\right) \cdot \frac{\|\boldsymbol{z}_0\|_2}{\sqrt{n}}$. *Moreover, if (VI) is* μ*-strongly monotone, as per Definition 1 and* $1/\rho = \frac{1}{\mu} + \frac{1}{\beta}$, *then* $\|\boldsymbol{z}_n\|_2 \leq \mathcal{O}\left(\sqrt{\frac{\beta^2+L}{\beta^3}}\right) \cdot \|\boldsymbol{z}_0\|_2 \cdot \exp\left(-\mathcal{O}(\rho) \cdot n\right)$.

5. Discussion

The fact that differently performing saddle point optimizers yield the same ODE on simple examples, motivated the use of the previously proposed HRDEs. We derived the corresponding HRDE of extragradient, optimistic gradient descent ascent (OGDA), as well as lookahead-minmax when combined with gradient descent ascent. Using these HRDEs we then proved the convergence of these methods in continuous time on bilinear games aligned with their observed performance.

Moreover, such HRDE representation allowed us to derive the first convergence proof in continuous time for the problem of monotone variational inequalities, specifically for the optimistic gradient descent ascent method. For this general problem, using only first-order smoothness assumptions we then showed *(i) last* iterate convergence of an implicit discretization of this HRDE dynamics, as well as *(ii) best* iterate convergence of an explicit discretization which corresponds to the original OGDA method.

There are several potential future directions. While HRDEs allowed for modeling the observed performance of optimistic gradient descent ascent, deriving HRDEs with an even higher resolution might allow for proving the analogous result for extragradient and lookahead-minmax. Alternatively, deriving different discretizations of the presented HRDEs could potentially allow for developing variants of the starting discrete method that may have good performances.

Acknowledgments

TC is supported by the Swiss National Science Foundation (SNSF), grant P2ELP2_199740. TC would like to thank Ya-Ping Hsieh for insightful discussions and feedback.

References

[1] Jacob Abernethy, Kevin A. Lai, and Andre Wibisono. Last-iterate convergence rates for min-max optimization: Convergence of hamiltonian gradient descent and consensus optimization. In Proceedings of the 32nd International Conference on Algorithmic Learning Theory, pages 3–47, 2021.

[2] Leonard Adolphs, Hadi Daneshmand, Aurelien Lucchi, and Thomas Hofmann. Local saddle point optimization: A curvature exploitation approach. arXiv:1805.05751, 2018.

[3] Yossi Arjevani and Ohad Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In ICML, pages 908–916, 2016.

[4] Hedy Attouch, Paul-Emile Maingé, and Patrick Redont. A second-order differential system with hessian-driven damping; application to non-elastic shock laws. Differential Equations and Applications, 4(1):27–65, 2012.

[5] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In AISTATS, pages 2863–2873, 2020.

[6] Dimitri P Bertsekas. Nonlinear programming. Athena scientific Belmont, 1999.

[7] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in GAN training with variance reduced extragradient. In NeurIPS, 2019.

[8] Tatjana Chavdarova, Matteo Pagliardini, Sebastian U Stich, François Fleuret, and Martin Jaggi. Taming GANs with Lookahead-Minmax. In ICLR, 2021.

[9] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In Advances in Neural Information Processing Systems, pages 9236–9246, 2018.

[10] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In ICLR, 2018.

[11] Jelena Diakonikolas, Constantinos Daskalakis, and Michael I. Jordan. Efficient methods for structured nonconvex-nonconcave min-max optimization. In AISTATS, 2021.

[12] Francisco Facchinei and Jong-Shi Pang. Finite-Dimensional Variational Inequalities and Complementarity Problems Vol I. Springer Series in Operations Research and Financial Engineering, Finite-Dimensional Variational Inequalities and Complementarity Problems. Springer-Verlag, 2003.

[13] Tanner Fiez and Lillian J Ratliff. Local convergence analysis of gradient descent ascent with finite timescale separation. In ICLR, 2021.

[14] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In AISTATS, 2019.

[15] Noah Golowich, Sarath Pattathil, Constantinos Daskalakis, and Asuman Ozdaglar. Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems. In Proceedings of Thirty Third Conference on Learning Theory, volume 125 of Proceedings of Machine Learning Research, pages 1758–1784, 2020.

[16] Ian Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. arXiv:1701.00160, 2016.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, 2014.

[18] Reyhane Askari Hemmat, Amartya Mitra, Guillaume Lajoie, and Ioannis Mitliagkas. Lead: Least-action dynamics for min-max optimization. arXiv:2010.13846, 2020.

[19] Ya-Ping Hsieh, Panayotis Mertikopoulos, and Volkan Cevher. The limits of min-max optimization algorithms: convergence to spurious non-critical sets. In ICML, 2021.

[20] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, NeurIPS, volume 32, 2019.

[21] Galina Michailovna Korpelevich. The extragradient method for finding saddle points and other problems. Matecon, 1976.

[22] Hans Lewy and Guido Stampacchia. On the regularity of the solution of a variational inequality. Communications on Pure and Applied Mathematics, 22(2):153–188, 1969. doi: https://doi.org/10.1002/cpa.3160220203.

[23] J. L. Lions and G. Stampacchia. Variational inequalities. Communications on Pure and Applied Mathematics, 20(3):493–519, 1967. doi: https://doi.org/10.1002/cpa.3160200302.

[24] Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic Hamiltonian gradient methods for smooth games. In ICML, pages 6370–6381, 2020.

[25] Eric Mazumdar and Lillian J Ratliff. On the convergence of gradient-based learning in continuous games. arXiv preprint arXiv:1804.05464, 2018.

[26] Panayotis Mertikopoulos, Christos H. Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2018.

[27] Panayotis Mertikopoulos, Houssam Lecouat, Bruno Zenati, , Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. 2019.

[28] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In International Conference on Machine Learning, pages 3481–3490, 2018.

[29] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. Convergence rate of $\mathcal{O}(1/k)$ for optimistic gradient and extra-gradient methods in smooth convex-concave saddle point problems. In SIAM Journal on Optimization, 2020.

[30] Renato D. C. Monteiro and Benar Fux Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. SIAM J. Optim., 20:2755–2787, 2010.

[31] A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM J. Optim., 2004.

[32] Yurii Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. Proceedings of the USSR Academy of Sciences, 269:543–547, 1983.

[33] Yurii Nesterov. Introductory lectures on convex optimization: A basic course, volume 87. Springer Science & Business Media, 2013.

[34] Shayegan Omidshafiei, Jason Pazis, Chris Amato, Jonathan P. How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In ICML, 2017.

[35] Joseph Pedlosky. Geophysical Fluid Dynamics. Springer Science & Business Media, 2013.

[36] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 1964.

[37] Leonid Denisovich Popov. A modification of the arrow–hurwicz method for search of saddle points. Mathematical Notes of the Academy of Sciences of the USSR, 28(5):845–848, 1980.

[38] L. Saydy, A. Tits, and E. Abed. Guardian maps and the generalized stability of parametrized families of matrices and polynomials. Mathematics of Control, Signals and Systems, 3:345–371, 1990.

[39] Bin Shi, Simon S. Du, Michael I. Jordan, and Weijie J. Su. Understanding the acceleration phenomenon via high-resolution differential equations. arXiv:1810.08907, 2018.

[40] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. Journal of Computational and Applied Mathematics, 1995.

[41] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. In SIAM Journal on Optimization, 2008.

[42] Yuanhao Wang, Guodong Zhang, and Jimmy Ba. On solving minimax optimization locally: A follow-the-ridge approach. In ICLR, 2020.

[43] X. Xie. Stable polynomials with complex coefficients. In 1985 24th IEEE Conference on Decision and Control, pages 324–325, 1985. doi: 10.1109/CDC.1985.268856.

[44] Guodong Zhang, Xuchan Bao, Laurent Lessard, and Roger Grosse. A unified analysis of first-order methods for smooth games via integral quadratic constraints. In JMLR, 2021.

[45] Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. In NeurIPS, 2019.

Appendix A. Notation & Omitted Definitions

Notation. Vectors are denoted with bold lowercase letters, e.g., $\boldsymbol{v}$, whereas bold uppercase letters denote matrices, and with $\boldsymbol{A} \succeq 0$ we denote that A is a positive semidefinite matrix. We use indices to refer to a discrete time sequence, e.g., $\boldsymbol{z}_n$, where with $\boldsymbol{z}(t)$ is a continuous time vector valued function. For any complex number $c \in \mathbb{C}$ we have that $c = \Re(c) + i\Im(c)$. The Euclidean norm of vector $\boldsymbol{v}$ is denoted with $\|\boldsymbol{v}\|_2$, and the inner product in Euclidean space with $\langle \cdot, \cdot \rangle$. $\|\boldsymbol{A}\|_F$ denotes the Frobenius norm of the matrix $\boldsymbol{A}$, i.e. $\|\boldsymbol{A}\|_F = \sqrt{\sum_i \sum_j a_{ij}^2}$.

Lyapunov Functions. We use *Lyapunov* functions as a main tool to show convergence in both continuous and the discrete time.

Definition 8 (Lyapunov Function) *A scalar function $\mathcal{L} : \mathbb{R}^d \mapsto \mathbb{R}$ is a Lyapunov function for a dynamical system $\dot{\boldsymbol{z}} = h(\boldsymbol{z})$, and a limiting set $S \subseteq \mathbb{R}^d$ iff:*
(i) $\mathcal{L}(\boldsymbol{z}) = 0$ for all $\boldsymbol{z} \in S$,
(ii) $\mathcal{L}(\boldsymbol{z}) > 0, \quad \forall \boldsymbol{z} \in \mathbb{R}^d \setminus S$,
(iii) $\dot{\mathcal{L}}(\boldsymbol{z}) < 0, \quad \forall \boldsymbol{z} \in \mathbb{R}^d \setminus S$.

Intuitively, $\mathcal{L}$ can be seen as energy function, and the last property ensures that it decreases along all possible trajectories of the described dynamical system. Such function—if it can be found—guarantees the convergence of the continuous time dynamics.

Appendix B. Additional Related Works

The lookahead-minmax algorithm [LA, 8] algorithm discussed in the main paper, was proposed recently and it explicitly exploits the rotational game dynamics by periodically restarting the current iterate to lie on a line between two k-step separated iterates produced by a given "base" optimizer, as presented in § 2.1. LA is an attractive min-max optimizer choice due to its performance on both bilinear games and challenging real-world min-max applications such as GANs, as well as its negligible computational overhead. In this work, we proved its convergence on bilinear games when combined with the otherwise non-converging GDA–which result was not shown before. Similarly, several works provided best-iterate convergence results on *convex-concave* problems [12, 27, 30]

However, despite the above advances on a specific problem, or when assuming particular structure of f [e.g., 11], knowledge of min-max optimization on general problems remains very limited relative to minimization. Moreover, Hsieh et al. [19] showed a negative result that contradicts the celebrated Sard's theorem for single-optimization problems and thus emphasizes why min-max optimization is notably more challenging. In particular, the authors showed that although some of the above methods converge on specific setups such as bilinear or convex-concave games, there exist a large class of problems for which all the popular min-max methods almost surely get attracted by a *spurious limit cycle*–which consist of points that are *not* a solution of the problem. The core of the analysis of [19] is based on the standard (low-resolution) ODE, which for the popular min-max optimizers is *identical* to that of GDA, see § 2.3. However, the bilinear game is one counter-example that these min-max methods differ in practice, whose differing convergence is also evident on many of the toy examples in [19] where one can show that LA-GDA converges yet gives the same low-resolution ODE as for GDA, as well as in real-world applications such as GANs where Extragradient and Lookahead *consistently* outperform GDA [see e.g. 8]. This indicates a strong need of a more precise ODE that

more closely captures the method's convergence behaviour in order to theoretically understand which properties we would need to avoid such undesirable convergence behavior.

In regard to analysis frameworks, several of the above last-iterate results are established using the *(stationary) canonical linear iterative* [CLI, 3] algorithmic framework, originally proposed for minimization. Another approach relies on the magnitude of spectral radius of the linearization of the associated operator [6], commonly used to analyze the stability of a method around fixed points [14, 42]. While the former requires second-order smoothness assumption for monotone VI analyses, the latter explicitly exploits the linearity of the operator. On the other hand, the well-established Lyapunov stability theory can be directly used for possibly non-linear dynamical systems, and if a Lyapunov function can be found, the convergence result holds *globally*. Several works make use of Lyapunov theory in the context of games, for example: *(i)* [18] view the iterate as a particle in a dynamical system while modelling also its rotational force and adding a compensating force to guarantee convergence on quadratic games, resulting in second-order update rule, and *(ii)* [13] establish local convergence guarantees of GDA when using timescale separation, by combining Lyapunov stability and *guard map* [38]. Zhang et al. [44] propose a standardized way of finding the parameters of a *quadratic* Lyapunov function of a given *first-order* saddle point optimizer in *discrete* time, using the theory of integral quadratic constraints, but can be applied solely to strongly-monotone operators.

As we pointed out in the main paper, the use of HRDEs in optimization was introduced by [39] in the context of single-objective minimization. The motivation for the use of HRDEs in this case was that the classical ODEs *cannot* distinguish between Nesterov's accelerated gradient method [32, 33] and Polyak's heavy-ball method [36]. Inspired from analysis used in fluid mechanics where the physical properties are investigated at different scales using various orders of perturbations [35], Shi et al., 2018 use the framework of *High-Resolution Differential Equations* (HRDEs) to distinguish between these two methods.

Appendix C. Proof of Proposition 2

The uniqueness of the solutions for our HRDEs for the case of bilinear games follows from standard results on the uniqueness of the solutions in the case linear dynamical systems.

For the more general case of Proposition 2, and for the HRDE of OGDA we follow the proof of Proposition 2.1 of [39]. The differential equations of OGDA have the form $\dot{\boldsymbol{u}} = \mathbf{G}(\boldsymbol{u})$, where $\boldsymbol{u} = (\boldsymbol{z}, \boldsymbol{\omega})$ and $\mathbf{G}$ is the vector field defined in (OGDA-HRDE). If we show that $\mathbf{G}$ is Lipschitz then we can apply Theorem 10 of [39] and Proposition 2 follows. Now to show the Lipschitzness of $\mathbf{G}$ we have from our assumptions the following:

$$\|V(\boldsymbol{z}_1) - V(\boldsymbol{z}_2)\|_2 \leq L_1 \|\boldsymbol{z}_2 - \boldsymbol{z}_1\|_2 . \tag{L1}$$

$$\|J(\boldsymbol{z}_1) - J(\boldsymbol{z}_2)\|_2 \leq L_2 \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 \tag{L2}$$

Also, from the two Lyapunov functions $\mathcal{L}_1, \mathcal{L}_2$ that we find in Section E.2 we have that for any initial condition $(\boldsymbol{z}_0, \boldsymbol{\omega}_0)$ both the norm of $\boldsymbol{\omega}(t)$ and the norm of $\boldsymbol{z}(t)$ will be upper bounded for all future times by a constant $\mathcal{C}_1$ that depends only on $\boldsymbol{z}_0$ and $\boldsymbol{\omega}_0$. Because of the Lipschitzness of V and J and the boundness of $\boldsymbol{\omega}(t)$ we hence get that the norm of $\dot{w}(t)$ is also bounded by a constant $\mathcal{C}_2$ that also only depends on $\boldsymbol{z}_0$ and $\boldsymbol{\omega}_0$. Therefore we have the following.

$$\sup_{0 \leq t \leq \infty} ||\boldsymbol{\omega}(t)|| \leq \mathcal{C}_1, \tag{C1}$$

$$\sup_{0\leq t\leq\infty} ||\dot{\boldsymbol{\omega}}(t)|| \leq \mathcal{C}_2. \tag{C2}$$

From OGDA we have:

$$\frac{d}{dt}\begin{bmatrix}\boldsymbol{z}_s\\ \boldsymbol{\omega}_s\end{bmatrix} = \begin{bmatrix}\boldsymbol{\omega}_s\\ -\frac{2}{\gamma}\boldsymbol{\omega}_s - \frac{2}{\gamma}V(\boldsymbol{z}_s) - 2J(\boldsymbol{z}_s)\boldsymbol{\omega}_s\end{bmatrix}.$$

Thus, for any $[\boldsymbol{z}_1, \boldsymbol{\omega}_1]^\intercal$, $[\boldsymbol{z}_2, \boldsymbol{\omega}_2]^\intercal \in$ with the initial condition bounds on the norm of $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$, we have:

$$\begin{aligned}
&\left\|\begin{bmatrix}\boldsymbol{\omega}_1\\ -\frac{2}{\gamma}\boldsymbol{\omega}_1 - \frac{2}{\gamma}V(\boldsymbol{z}_1) - 2J(\boldsymbol{z}_1)\boldsymbol{\omega}_1\end{bmatrix} - \begin{bmatrix}\boldsymbol{\omega}_2\\ -\frac{2}{\gamma}\boldsymbol{\omega}_2 - \frac{2}{\gamma}V(\boldsymbol{z}_2) - 2J(\boldsymbol{z}_2)\boldsymbol{\omega}_2\end{bmatrix}\right\| \\
&\leq \left\|\begin{matrix}\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2\\ -\frac{2}{\gamma}(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2)\end{matrix}\right\| + \frac{2}{\gamma}\left\|\begin{matrix}0\\ (V(\boldsymbol{z}_1) - V(\boldsymbol{z}_2))\end{matrix}\right\| + 2\left\|\begin{matrix}0\\ J(\boldsymbol{z}_1)(\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2)\end{matrix}\right\| + 2\left\|\begin{matrix}0\\ \boldsymbol{\omega}_2(J(\boldsymbol{z}_1) - J(\boldsymbol{z}_2))\end{matrix}\right\| \\
&\leq \sqrt{1+\frac{4}{\gamma^2}}||\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2|| + \frac{2}{\gamma}L_1||\boldsymbol{z}_1 - \boldsymbol{z}_2|| + 2\underbrace{||J(\boldsymbol{z}_1)||}_{\leq L_2||\boldsymbol{z}_1||}||\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2|| + 2\underbrace{||\boldsymbol{\omega}_2||}_{\leq\mathcal{C}_1}\underbrace{||J(\boldsymbol{z}_1) - J(\boldsymbol{z}_2)||}_{\leq L_2||\boldsymbol{z}_1 - \boldsymbol{z}_2||} \\
&\leq (\frac{2}{\gamma}L_1 + 2\mathcal{C}_1L_2)||\boldsymbol{z}_1 - \boldsymbol{z}_2|| + (\sqrt{1+\frac{4}{\gamma^2}} + (L_2||\boldsymbol{z}_1||))||\boldsymbol{\omega}_1 - \boldsymbol{\omega}_2|| \\
&\leq 2\max\{\frac{2}{\gamma}L_1 + 2\mathcal{C}_1L_2, \sqrt{1+\frac{4}{\gamma^2}} + (L_2||\boldsymbol{z}_1||)\}\left\|\begin{bmatrix}\boldsymbol{z}_1\\ \boldsymbol{\omega}_1\end{bmatrix} - \begin{bmatrix}\boldsymbol{z}_2\\ \boldsymbol{\omega}_2\end{bmatrix}\right\|
\end{aligned}$$

and finally we can use the initial bounds that we have on the norms of $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, $\boldsymbol{\omega}_1$, $\boldsymbol{\omega}_2$ and we get that the vector field $\mathbf{G}$ of OGDA is Lipschitz and hence we can apply Theorem 10 of [39] and Proposition 2 follows.

Appendix D. Analysis on Bilinear Games

In this section, we consider the following bilinear game, with full rank $\boldsymbol{A} \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$:

$$\min_{\boldsymbol{x}\in\mathbb{R}^{d_1}} \max_{\boldsymbol{y}\in\mathbb{R}^{d_2}} \boldsymbol{x}^\intercal \boldsymbol{A}\boldsymbol{y}\,. \tag{BG}$$

The joint vector field of BG and its Jacobian are:

$$V_{\text{BG}}(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{A}\boldsymbol{y} \\ -\boldsymbol{A}^\intercal \boldsymbol{x} \end{bmatrix} \tag{BG:JVF}$$

$$J_{\text{BG}}(\boldsymbol{z}) = \begin{bmatrix} 0 & \boldsymbol{A} \\ -\boldsymbol{A}^\intercal & 0 \end{bmatrix} \tag{BG:Jac-JVF}$$

By replacing Eq. BG:JVF and BG:Jac-JVF in the derived HRDEs we analyze the convergence of the corresponding method on BG. Moreover, as the obtained system is linear this can be done using dynamical systems tools, *without* designing Lyapunov functions, as we describe next. A dynamical system $\begin{bmatrix}\dot{\boldsymbol{z}}^\intercal & \dot{\boldsymbol{\omega}}^\intercal\end{bmatrix}^\intercal = \boldsymbol{C}\begin{bmatrix}\boldsymbol{z}^\intercal & \boldsymbol{\omega}^\intercal\end{bmatrix}^\intercal$ is *stable* iff the real part of the eigenvalues of $\boldsymbol{C}$ is always *negative*: $\Re(\lambda_i) < 0, \quad \forall \lambda_i \in \text{Sp}(\boldsymbol{C})$. The *Routh–Hurwitz stability criterion* provides a necessary and sufficient condition for the stability of the linear system, and allows for determining if the system is stable, without explicitly deriving the eigenvalues of the matrix $\boldsymbol{C}$. Using the coefficients of its associated characteristic polynomial, the test is performed on the so called *Hurwitz* array, or its generalized form [43] when the coefficients are complex numbers. Proofs are in App. D.1–D.5.

Theorem 9 (divergence of GDA on BG) *For any step size $\gamma > 0$, there exist an eigenvalue of $\boldsymbol{C}_{GDA}$ whose real part is non-negative, $\exists \lambda_i \in Sp(\boldsymbol{C}_{GDA}), \quad$ s.t. $\Re(\lambda_i) \geq 0$. Thus the Gradient Descent Ascent method diverges on the BG problem for any choice of nonzero step-size (and any $\boldsymbol{A}$).*

Theorem 10 (convergence of EG on BG) *For sufficiently small γ, the real part of the eigenvalues of $\boldsymbol{C}_{EG}$ is always negative, $\Re(\lambda_i) < 0, \quad \forall \lambda_i \in Sp(\boldsymbol{C}_{EG})$, thus the ExtraGradient method converges on the BG problem for any such step-size.*

Theorem 11 (convergence of OGDA on BG) *For sufficiently small γ, the real part of the eigenvalues of $\boldsymbol{C}_{OGDA}$ is always negative, $\Re(\lambda_i) < 0, \quad \forall \lambda_i \in Sp(\boldsymbol{C}_{OGDA})$, thus the Optimistic Gradient Ascent Descent method converges on the BG problem for any such step-size.*

Theorem 12 (convergence of LA2-GDA on BG) *For sufficiently small γ, the real part of the eigenvalues of $\boldsymbol{C}_{LA2\text{-}GDA}$ is always negative, $\Re(\lambda_i) < 0, \quad \forall \lambda_i \in Sp(\boldsymbol{C}_{LA2\text{-}GDA})$, thus the LA2-GDA method converges on the BG problem for any such step-size.*

Theorem 13 (convergence of LA3-GDA on BG) *For γ, α that satisfy:*

$$(3+2\gamma)^2 + 2\gamma^2(3+2\gamma)\frac{|\bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{A}^\intercal \boldsymbol{A}\boldsymbol{y}|}{|\bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{y}|} + \gamma^4 \frac{|\bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{A}^\intercal \boldsymbol{A}\boldsymbol{y}|^2}{|\bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{y}|^2} \leq \frac{1}{\alpha}$$

the real part of the eigenvalues of $\boldsymbol{C}_{LA2\text{-}GDA}$ is always negative, $\Re(\lambda_i) < 0, \quad \forall \lambda_i \in Sp(\boldsymbol{C}_{LA3\text{-}GDA})$ and the LA3-GDA method converges on the BG problem.

D.1. Proof of Theorem 9: Divergence of GDA on BG

Recall that for the GDA optimizer we have the following (Eq. GDA-HRDE):

$$\begin{aligned}\dot{\boldsymbol{z}}(t) &= \boldsymbol{\omega}(t)\\ \dot{\boldsymbol{\omega}}(t) &= -\beta\cdot\boldsymbol{\omega}(t) - \beta\cdot V(\boldsymbol{z}(t)).\end{aligned}$$

By denoting $\dot{\boldsymbol{x}}(t) = \boldsymbol{\omega}_x(t)$, $\dot{\boldsymbol{y}}(t) = \boldsymbol{\omega}_y(t)$, and $b = \frac{2}{\gamma}$, for GDA we have the following:

$$\begin{bmatrix}\dot{\boldsymbol{x}}(t)\\ \dot{\boldsymbol{y}}(t)\\ \dot{\boldsymbol{\omega}}_x(t)\\ \dot{\boldsymbol{\omega}}_y(t)\end{bmatrix} = \underbrace{\begin{bmatrix}0 & 0 & \boldsymbol{I} & 0\\ 0 & 0 & 0 & \boldsymbol{I}\\ 0 & -\beta\boldsymbol{A} & -\beta\boldsymbol{I} & 0\\ \beta\boldsymbol{A}^\intercal & 0 & 0 & -\beta\boldsymbol{I}\end{bmatrix}}_{\triangleq \boldsymbol{C}_{\text{GDA}}} \cdot \begin{bmatrix}\boldsymbol{x}(t)\\ \boldsymbol{y}(t)\\ \boldsymbol{\omega}_x(t)\\ \boldsymbol{\omega}_y(t)\end{bmatrix}.$$

Proof [Proof of Thm. 9] To obtain the eigenvalues $\lambda \in \mathbb{C}$ of $\boldsymbol{C}_{\text{GDA}}$ we have:

$$\begin{aligned}\det(\boldsymbol{C}_{\text{GDA}} - \lambda\boldsymbol{I}) &= \det\Big(\begin{bmatrix}-\lambda\boldsymbol{I} & 0 & \boldsymbol{I} & 0\\ 0 & -\lambda\boldsymbol{I} & 0 & \boldsymbol{I}\\ 0 & -\beta\boldsymbol{A} & -(\beta+\lambda)\boldsymbol{I} & 0\\ \beta\boldsymbol{A}^\intercal & 0 & 0 & -(\beta+\lambda)\boldsymbol{I}\end{bmatrix}\Big)\\ &= \det\Big(\lambda(\beta+\lambda)\boldsymbol{I} + \beta\begin{bmatrix}0 & \boldsymbol{A}\\ -\boldsymbol{A}^\intercal & 0\end{bmatrix}\Big)\\ &= \det\Big(\begin{bmatrix}\lambda(\beta+\lambda)\boldsymbol{I} & \beta\boldsymbol{A}\\ -\beta\boldsymbol{A}^\intercal & \lambda(\beta+\lambda)\boldsymbol{I}\end{bmatrix}\Big)\\ &= \det\Big(\Big[\lambda^2(\beta+\lambda)^2\boldsymbol{I} + \beta^2\boldsymbol{A}\boldsymbol{A}^\intercal\Big]\Big)\\ &= \beta^{2n}\det\Big(\Big[\tfrac{\lambda^2}{\beta^2}(\beta+\lambda)^2\boldsymbol{I} + \boldsymbol{A}\boldsymbol{A}^\intercal\Big]\Big)\ ,\end{aligned}$$

where we successively used $\det(\begin{bmatrix}\boldsymbol{B}_1 & \boldsymbol{B}_2\\ \boldsymbol{B}_3 & \boldsymbol{B}_4\end{bmatrix}) = \det(\boldsymbol{B}_1\boldsymbol{B}_4 - \boldsymbol{B}_2\boldsymbol{B}_3)$.

Let κ denote an eigenvalue of $-\boldsymbol{A}\boldsymbol{A}^\intercal$, and since $-\boldsymbol{A}\boldsymbol{A}^\intercal$ is symmetric and negative definite and $\boldsymbol{A}$ is full rank, we have that $\kappa \in \mathbb{R}$, $\kappa < 0$. From the last expression we observe that $\kappa = \frac{\lambda^2}{\beta^2}(\beta+\lambda)^2$ is the eigenvalue of $-\boldsymbol{A}\boldsymbol{A}^\intercal$. Thus, we have the following polynomial with real coefficients:

$$\begin{aligned}\lambda^2(\beta+\lambda)^2 &= \kappa\beta^2\,, \quad \text{or}\\ \lambda^4 + 2\beta\lambda^3 + \beta^2\lambda^2 + 0\lambda - \kappa\beta^2 &= 0\,.\end{aligned}$$

The Routh Array is then:

1	β^2	$-\kappa\beta^2$
2β	0	0
β^2	$-\kappa\beta^2$	0
$2\kappa\beta$	0	0
$-\kappa\beta^2$	0	0

where recall that $\beta = \frac{2}{\gamma} > 0$ and $\kappa < 0$. We observe that the first column of the Routh Array for GDA has change of signs, indicating that there is an eigenvalue $\lambda_i \in \text{Spec}\{\boldsymbol{C}_{\text{GDA}}\}$ whose real part is positive $\Re\{\lambda_i\} > 0$. Due to the Routh Hurwitz criterion the system is unstable.

Thus, for any choice of step size γ the GDA method *diverges* on the BG problem. ■

D.2. Proof of Theorem 10: Convergence of EG on BG

Recall that for the EG optimizer we have the following (Eq. EG-HRDE):

$$\begin{aligned}\dot{\boldsymbol{z}}(t) &= \boldsymbol{\omega}(t)\\ \dot{\boldsymbol{\omega}}(t) &= -\beta\cdot\boldsymbol{\omega}(t) - \beta\cdot V(\boldsymbol{z}(t)) + 2\cdot J(\boldsymbol{z}(t))\cdot V(\boldsymbol{z}(t))\,,\end{aligned}$$

where $\beta = \frac{2}{\gamma}$.

By denoting $\dot{\boldsymbol{x}}(t) = \boldsymbol{\omega}_x(t)$ and $\dot{\boldsymbol{y}}(t) = \boldsymbol{\omega}_y(t)$ for EG we have the following:

$$\begin{bmatrix}\dot{\boldsymbol{x}}(t)\\ \dot{\boldsymbol{y}}(t)\\ \dot{\boldsymbol{\omega}}_x(t)\\ \dot{\boldsymbol{\omega}}_y(t)\end{bmatrix} = \underbrace{\begin{bmatrix}0 & 0 & \boldsymbol{I} & 0\\ 0 & 0 & 0 & \boldsymbol{I}\\ -2\boldsymbol{A}\boldsymbol{A}^\intercal & -\beta\boldsymbol{A} & -\beta\boldsymbol{I} & 0\\ \beta\boldsymbol{A}^\intercal & -2\boldsymbol{A}^\intercal\boldsymbol{A} & 0 & -\beta\boldsymbol{I}\end{bmatrix}}_{\triangleq\boldsymbol{C}_{\text{EG}}}\cdot\begin{bmatrix}\boldsymbol{x}(t)\\ \boldsymbol{y}(t)\\ \boldsymbol{\omega}_x(t)\\ \boldsymbol{\omega}_y(t)\end{bmatrix}.$$

Proof [Proof of Thm. 10] To obtain the eigenvalues $\lambda \in \mathbb{C}$ of $\boldsymbol{C}_{\text{EG}}$ we have:

$$\begin{aligned}\det(\boldsymbol{C}_{\text{EG}} - \lambda\boldsymbol{I}) &= \det\Big(\begin{bmatrix}-\lambda\boldsymbol{I} & 0 & \boldsymbol{I} & 0\\ 0 & -\lambda\boldsymbol{I} & 0 & \boldsymbol{I}\\ -2\boldsymbol{A}\boldsymbol{A}^\intercal & -\beta\boldsymbol{A} & -(\beta+\lambda)\boldsymbol{I} & 0\\ \beta\boldsymbol{A}^\intercal & -2\boldsymbol{A}^\intercal\boldsymbol{A} & 0 & -(\beta+\lambda)\boldsymbol{I}\end{bmatrix}\Big)\\ &= \det\big(\lambda(\beta+\lambda)\boldsymbol{I} - \underbrace{\begin{bmatrix}-2\boldsymbol{A}\boldsymbol{A}^\intercal & -\beta\boldsymbol{A}\\ \beta\boldsymbol{A}^\intercal & -2\boldsymbol{A}^\intercal\boldsymbol{A}\end{bmatrix}}_{\triangleq\boldsymbol{D}}\big)\,,\end{aligned}$$

where we used $\det(\begin{bmatrix}\boldsymbol{B}_1 & \boldsymbol{B}_2\\ \boldsymbol{B}_3 & \boldsymbol{B}_4\end{bmatrix}) = \det(\boldsymbol{B}_1\boldsymbol{B}_4 - \boldsymbol{B}_2\boldsymbol{B}_3)$.

Let $\mu = \mu_1 + \mu_2 i \in \mathbb{C}$ denote the eigenvalues of $\boldsymbol{D}$. We have: $\lambda(\beta + \lambda) - \mu = 0$. Using the generalized Hurwitz theorem for polynomials with complex coefficients [43], we obtain the following generalized Hurwitz array:

$$\begin{array}{cccc} \lambda^2 & 1 & 0 & \mu_1 \\ \lambda^1 & \beta & \mu_2 & 0 \\ & -\mu_2 & \beta\mu_2 & 0 \\ \lambda^0 & -\mu_2^2 - \beta^2\mu_1 & 0 & 0 \end{array}$$

where the terms whose change of sign determine the stability of the polynomial are highlighted. As $\beta > 0$, it follows that the system is stable iff $\mu_1 < -\frac{1}{\beta^2}\mu_2^2$.

Thus, it suffices to show that:

$$\Re(\mu(\boldsymbol{z})) < -\frac{1}{\beta^2}\big(\Im(\mu(\boldsymbol{z}))\big)^2 . \tag{2}$$

We have that:

$$\begin{aligned} \mu(\boldsymbol{z}) = \bar{\boldsymbol{z}}^\intercal \boldsymbol{D}\boldsymbol{z} &= \begin{bmatrix} \bar{\boldsymbol{x}}^\intercal & \bar{\boldsymbol{y}}^\intercal \end{bmatrix} \begin{bmatrix} -2\boldsymbol{A}\boldsymbol{A}^\intercal & -\beta\boldsymbol{A} \\ \beta\boldsymbol{A}^\intercal & -2\boldsymbol{A}^\intercal\boldsymbol{A} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} \\ &= -2||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 - 2||\boldsymbol{A}\boldsymbol{y}||_2^2 + \beta\big(\bar{\boldsymbol{y}}^\intercal\boldsymbol{A}^\intercal\boldsymbol{x} - \bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}\big) \\ &= \underbrace{-2||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 - 2||\boldsymbol{A}\boldsymbol{y}||_2^2}_{\Re\left(\mu(\boldsymbol{z})\right)} + \underbrace{2\beta \cdot \Im(\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y})}_{\Im\left(\mu(\boldsymbol{z})\right)} \cdot i \,, \end{aligned}$$

where the last equality follows from the fact that $\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}$ is a complex conjugate of $\bar{\boldsymbol{y}}^\intercal\boldsymbol{A}^\intercal\boldsymbol{x}$, thus $\bar{\boldsymbol{y}}^\intercal\boldsymbol{A}^\intercal\boldsymbol{x} - \bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y} = 2\Im\big(\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}\big) \cdot i$. By replacing this in Eq. 2, it follows that we need to show:

$$-2\big(||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2\big) \leq -4\Im^2(\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}) \,, \quad \text{or:}$$

$$2\big(||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2\big) \geq 4\Im^2(\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}) \,.$$

Thus, it suffices to show that:

$$||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2 \geq 2|\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}|^2 \,. \tag{3}$$

Consider the case $||\boldsymbol{x}||_2 \leq ||\boldsymbol{y}||_2 \leq 1$. We have two sub-cases.

1. $||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 \leq ||\boldsymbol{A}\boldsymbol{y}||_2^2$**.** We can set $||\boldsymbol{A}\boldsymbol{y}||_2 = ||\boldsymbol{y}||_2$, and we have:

$$||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 + ||\boldsymbol{A}^\intercal\boldsymbol{y}||_2^2 = ||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 + ||\boldsymbol{y}||_2^2 \geq \frac{1}{2}(||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 + ||\boldsymbol{y}||_2^2)^2 \geq 2||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2||\boldsymbol{y}||_2^2 \geq 2|\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}|_2^2 \,, \tag{4}$$

where the last inequality follows from Cauchy–schwarz inequality.

2. $||\boldsymbol{A}\boldsymbol{y}||_2^2 \leq ||\boldsymbol{A}^\intercal \boldsymbol{x}||_2^2$**.** We can set $||\boldsymbol{A}^\intercal \boldsymbol{x}||_2 = ||\boldsymbol{x}||_2$, and we have:

$$||\boldsymbol{A}^\intercal \boldsymbol{x}||_2^2 + ||\boldsymbol{A}^\intercal \boldsymbol{y}||_2^2 = ||\boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2 \geq \frac{1}{2}(||\boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2)^2 \geq 2||\boldsymbol{x}||_2^2||\boldsymbol{A}\boldsymbol{y}||_2^2 \geq 2|\bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{y}|_2^2\,, \tag{5}$$

where the last inequality follows from Cauchy–schwarz inequality.

The case $||\boldsymbol{y}||_2 \leq ||\boldsymbol{x}||_2 \leq 1$ can be shown analogously. ■

D.3. Proof of Theorem 11: Convergence of OGDA on BG

Recall that for the OGDA optimizer we have the following (Eq. OGDA-HRDE):

$$\begin{aligned}\dot{\boldsymbol{z}}(t) &= \boldsymbol{\omega}(t)\\ \dot{\boldsymbol{\omega}}(t) &= -\beta \cdot \boldsymbol{\omega}(t) - \beta \cdot V(\boldsymbol{z}(t)) - 2 \cdot J(\boldsymbol{z}(t)) \cdot \boldsymbol{\omega}(t)\,,\end{aligned}$$

where $\beta = \frac{2}{\gamma}$.

By denoting $\dot{\boldsymbol{x}}(t) = \boldsymbol{\omega}_x(t)$ and $\dot{\boldsymbol{y}}(t) = \boldsymbol{\omega}_y(t)$, for OGDA we have the following:

$$\begin{bmatrix}\dot{\boldsymbol{x}}(t)\\ \dot{\boldsymbol{y}}(t)\\ \dot{\boldsymbol{\omega}}_x(t)\\ \dot{\boldsymbol{\omega}}_y(t)\end{bmatrix} = \underbrace{\begin{bmatrix}0 & 0 & \boldsymbol{I} & 0\\ 0 & 0 & 0 & \boldsymbol{I}\\ 0 & -\beta\boldsymbol{A} & -\beta\boldsymbol{I} & -2\boldsymbol{A}\\ \beta\boldsymbol{A}^\intercal & 0 & 2\boldsymbol{A}^\intercal & -\beta\boldsymbol{I}\end{bmatrix}}_{\triangleq \boldsymbol{C}_{\text{OGDA}}} \begin{bmatrix}\boldsymbol{x}(t)\\ \boldsymbol{y}(t)\\ \boldsymbol{\omega}_x(t)\\ \boldsymbol{\omega}_y(t)\end{bmatrix}. \tag{6}$$

Proof [Proof of Thm. 11] To obtain the eigenvalues $\lambda \in \mathbb{C}$ of $\boldsymbol{C}_{\text{OGDA}}$ we have:

$$\begin{aligned}\det(\boldsymbol{C}_{\text{OGDA}} - \lambda\boldsymbol{I}) &= \det\Big(\begin{bmatrix}-\lambda\boldsymbol{I} & 0 & \boldsymbol{I} & 0\\ 0 & -\lambda\boldsymbol{I} & 0 & \boldsymbol{I}\\ 0 & -\beta\boldsymbol{A} & -(\beta+\lambda)\boldsymbol{I} & -2\boldsymbol{A}\\ \beta\boldsymbol{A}^\intercal & 0 & 2\boldsymbol{A}^\intercal & -(\beta+\lambda)\boldsymbol{I}\end{bmatrix}\Big)\\ &= \det\Big(\begin{bmatrix}\lambda(\beta+\lambda)\boldsymbol{I} & 2\lambda\boldsymbol{A}\\ -2\lambda\boldsymbol{A}^\intercal & \lambda(\beta+\lambda)\boldsymbol{I}\end{bmatrix} - \begin{bmatrix}0 & -\beta\boldsymbol{A}\\ \beta\boldsymbol{A}^\intercal & 0\end{bmatrix}\Big)\\ &= \det\Big(\begin{bmatrix}\lambda(\beta+\lambda)\boldsymbol{I} & (2\lambda+\beta)\boldsymbol{A}\\ -(2\lambda+\beta)\boldsymbol{A}^\intercal & \lambda(\beta+\lambda)\boldsymbol{I}\end{bmatrix}\Big)\,,\end{aligned}$$

where we used $\det(\begin{bmatrix}\boldsymbol{B}_1 & \boldsymbol{B}_2\\ \boldsymbol{B}_3 & \boldsymbol{B}_4\end{bmatrix}) = \det(\boldsymbol{B}_1\boldsymbol{B}_4 - \boldsymbol{B}_2\boldsymbol{B}_3)$.

Square $\boldsymbol{A}$. For simplicity, we will first consider the case when $\boldsymbol{A}$ is square matrix. For the eigenvalues of $\boldsymbol{C}_{\text{OGDA}}$ we have:

$$\begin{aligned}\det(\boldsymbol{C}_{\text{OGDA}} - \lambda \boldsymbol{I}) &= \det\left(\left[\lambda^2(\beta+\lambda)^2\boldsymbol{I} + (2\lambda+\beta)^2\boldsymbol{A}\boldsymbol{A}^\intercal\right]\right) \\ &= (2\lambda+\beta)^{2n}\det\left(\left[\tfrac{\lambda^2(\beta+\lambda)^2}{(2\lambda+\beta)^2}\boldsymbol{I} + \boldsymbol{A}\boldsymbol{A}^\intercal\right]\right),\end{aligned}$$

where we used $\det(\begin{bmatrix}\boldsymbol{B}_1 & \boldsymbol{B}_2 \\ \boldsymbol{B}_3 & \boldsymbol{B}_4\end{bmatrix}) = \det(\boldsymbol{B}_1\boldsymbol{B}_4 - \boldsymbol{B}_2\boldsymbol{B}_3)$.

Let κ denote the eigenvalue of $-\boldsymbol{A}\boldsymbol{A}^\intercal$, thus $\kappa \in \mathbb{R}, \kappa < 0$. From the last expression we observe that $\kappa = \frac{\lambda^2(\beta+\lambda)^2}{(2\lambda+\beta)^2}$ is the eigenvalue of $-\boldsymbol{A}\boldsymbol{A}^\intercal$. Thus, we have the following polynomial with real coefficients:

$$\begin{aligned}\lambda^2(\beta+\lambda)^2 &= \kappa(2\lambda+\beta)^2\,, \quad \text{or} \\ \lambda^4 + 2\beta\lambda^3 + (\beta^2 - 4\kappa)\lambda^2 - 4\beta\kappa\lambda - \kappa\beta^2 &= 0\,.\end{aligned}$$

The Routh Array is then:

1	$\beta^2 - 4\kappa$	$-\kappa\beta^2$
2β	$-4\beta\kappa$	0
$\beta^2 - 2\kappa$	$-\kappa\beta^2$	0
$\frac{(-2\beta\kappa)(3\beta^2-4\kappa)}{(\beta^2-2\kappa)} > 0$	0	0
$\frac{(-2\beta\kappa)(3\beta^2-4\kappa)(-\kappa\beta^2)}{(\beta^2-2\kappa)} > 0$	0	0

where recall that $\beta = \frac{2}{\gamma} > 0$ and $\kappa < 0$. We observe that the first column of the Routh Array has only positive elements change of signs, implying that all eigenvalues $\Re\{\lambda_i\} < 0, \quad \forall \lambda_i \in Sp\{\boldsymbol{C}_{\text{OGDA}}\}$– due to the Routh Hurwitz criterion the system is stable.

General $\boldsymbol{A}$. When $\boldsymbol{A}$ is not necessarily square, using the fact that $\lambda(\beta+\lambda)\boldsymbol{I}$ is invertable—since $\boldsymbol{C}_{\text{OGDA}}$ is full rank—and using $\det(\begin{bmatrix}\boldsymbol{B}_1 & \boldsymbol{B}_2 \\ \boldsymbol{B}_3 & \boldsymbol{B}_4\end{bmatrix}) = \det(\boldsymbol{B}_4)\det(\boldsymbol{B}_1 - \boldsymbol{B}_2\boldsymbol{B}_4^{-1}\boldsymbol{B}_3)$, for the eigenvalues of $\boldsymbol{C}_{\text{OGDA}}$ we have:

$$\begin{aligned}\det(\boldsymbol{C}_{\text{OGDA}} - \lambda\boldsymbol{I}) &= \det\left(\lambda(\beta+\lambda)\boldsymbol{I}\right)\det\left(\lambda(\beta+\lambda)\boldsymbol{I} + (2\lambda+\beta)^2(\lambda(\beta+\lambda))^{-1}\boldsymbol{A}\boldsymbol{A}^\intercal\right) \\ &= \underbrace{\det\left(\lambda(\beta+\lambda)\boldsymbol{I}\right)}_{1^\circ}\underbrace{\det\left(\frac{\lambda^2(\beta+\lambda)^2}{(2\lambda+\beta)^2}\boldsymbol{I} + \boldsymbol{A}\boldsymbol{A}^\intercal\right)}_{2^\circ}\end{aligned}$$

As $1^\circ \neq 0$, it has to hold that $2^\circ = 0$. Let $\kappa \in Sp\{-\boldsymbol{A}^\intercal\boldsymbol{A}\}$ (thus $\kappa \in \mathbb{R}$ and $\kappa < 0$), and thus we get the same polynomial with real coefficients, as for the case when $\boldsymbol{A}$ is square:

$$\begin{aligned}\lambda^2(\beta+\lambda)^2 &= \kappa(2\lambda+\beta)^2\,, \quad \text{or} \\ \lambda^4 + 2\beta\lambda^3 + (\beta^2 - 4\kappa)\lambda^2 - 4\beta\kappa\lambda - \kappa\beta^2 &= 0\,.\end{aligned}$$

Hence, we get the same Routh Array as above, and the same proof follows.

Thus, for any choice of step size γ the OGDA method *converges* on the BG problem. ■

D.4. Proof of Theorem 12: Convergence of LA2-GDA on BG

Recall that for the LA2-GDA optimizer we have the following (Eq. LA2-GDA-HRDE):

$$\begin{aligned}\dot{\boldsymbol{z}}(t) &= \boldsymbol{\omega}(t)\\ \dot{\boldsymbol{\omega}}(t) &= -\beta\boldsymbol{\omega}(t) - 2\alpha\beta\cdot V(\boldsymbol{z}(t)) + 2\alpha J(\boldsymbol{z}(t))V(\boldsymbol{z}(t))\,,\end{aligned}$$

where $\beta = \frac{2}{\gamma}$.

By denoting $\dot{\boldsymbol{x}}(t) = \boldsymbol{\omega}_x(t)$ and $\dot{\boldsymbol{y}}(t) = \boldsymbol{\omega}_y(t)$ for LA2-GDA we have the following:

$$\begin{bmatrix}\dot{\boldsymbol{x}}(t)\\ \dot{\boldsymbol{y}}(t)\\ \dot{\boldsymbol{\omega}}_x(t)\\ \dot{\boldsymbol{\omega}}_y(t)\end{bmatrix} = \underbrace{\begin{bmatrix}0 & 0 & \boldsymbol{I} & 0\\ 0 & 0 & 0 & \boldsymbol{I}\\ -2\alpha\boldsymbol{A}\boldsymbol{A}^\intercal & -\frac{4\alpha}{\gamma}\boldsymbol{A} & -\frac{2}{\gamma}\boldsymbol{I} & 0\\ \frac{4\alpha}{\gamma}\boldsymbol{A}^\intercal & -2\alpha\boldsymbol{A}^\intercal\boldsymbol{A} & 0 & -\frac{2}{\gamma}\boldsymbol{I}\end{bmatrix}}_{\triangleq \boldsymbol{C}_{\text{LA2-GDA}}} \begin{bmatrix}\boldsymbol{x}(t)\\ \boldsymbol{y}(t)\\ \boldsymbol{\omega}_x(t)\\ \boldsymbol{\omega}_y(t)\end{bmatrix}.$$

Proof [Proof of Thm. 12] As in § D.2, to obtain the eigenvalues $\lambda \in \mathbb{C}$ of $\boldsymbol{C}_{\text{LA2-GDA}}$ we have:

$$\begin{aligned}\det(\boldsymbol{C}_{\text{LA2-GDA}} - \lambda\boldsymbol{I}) &= \det\Big(\begin{bmatrix}-\lambda\boldsymbol{I} & 0 & \boldsymbol{I} & 0\\ 0 & -\lambda\boldsymbol{I} & 0 & \boldsymbol{I}\\ -2\alpha\boldsymbol{A}\boldsymbol{A}^\intercal & -\frac{4\alpha}{\gamma}\boldsymbol{A} & -(\frac{2}{\gamma}+\lambda)\boldsymbol{I} & 0\\ \frac{4\alpha}{\gamma}\boldsymbol{A}^\intercal & -2\alpha\boldsymbol{A}^\intercal\boldsymbol{A} & 0 & -(\frac{2}{\gamma}+\lambda)\boldsymbol{I}\end{bmatrix}\Big)\\ &= \det\Big(\lambda(\frac{2}{\gamma}+\lambda)\boldsymbol{I} - \underbrace{\begin{bmatrix}-2\boldsymbol{A}\boldsymbol{A}^\intercal & -\frac{2}{\gamma}\boldsymbol{A}\\ \frac{2}{\gamma}\boldsymbol{A}^\intercal & -2\boldsymbol{A}^\intercal\boldsymbol{A}\end{bmatrix}}_{\triangleq \boldsymbol{D}}\Big)\,.\end{aligned}$$

We observe that only the lower left block matrix $\boldsymbol{D}$ differs from the proof for EG in § D.2, thus the inequality Eq. 2 needs to be proven for this optimizer too.

However, due to the different lower left block matrices for LA2-GDA we have:

$$\begin{aligned}\mu(\boldsymbol{z}) = \bar{\boldsymbol{z}}^\intercal\boldsymbol{D}\boldsymbol{z} &= \begin{bmatrix}\bar{\boldsymbol{x}}^\intercal & \bar{\boldsymbol{y}}^\intercal\end{bmatrix}\begin{bmatrix}-2\alpha\boldsymbol{A}\boldsymbol{A}^\intercal & -\frac{4\alpha}{\gamma}\boldsymbol{A}\\ \frac{4\alpha}{\gamma}\boldsymbol{A}^\intercal & -2\alpha\boldsymbol{A}^\intercal\boldsymbol{A}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{bmatrix}\\ &= -2\alpha||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 - 2\alpha||\boldsymbol{A}\boldsymbol{y}||_2^2 + \frac{4\alpha}{\gamma}\big(\bar{\boldsymbol{y}}^\intercal\boldsymbol{A}^\intercal\boldsymbol{x} - \bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}\big)\\ &= \underbrace{-2\alpha\big(||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 - ||\boldsymbol{A}\boldsymbol{y}||_2^2\big)}_{\Re\left(\mu(\boldsymbol{z})\right)} + \underbrace{\frac{4\alpha}{\gamma}\cdot\Im(\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y})\cdot i}_{\Im\left(\mu(\boldsymbol{z})\right)}\,,\end{aligned}$$

where the last equality follows from the fact that $\bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{y}$ is a complex conjugate of $\bar{\boldsymbol{y}}^\intercal \boldsymbol{A}^\intercal \boldsymbol{x}$, thus $\bar{\boldsymbol{y}}^\intercal \boldsymbol{A}^\intercal \boldsymbol{x} - \bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{y} = 2\Im\big(\bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{y}\big)\cdot i$. By replacing this in Eq. 2, we need to show that:

$$-2\alpha\big(||\boldsymbol{A}^\intercal \boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2\big) \leq -\frac{\gamma^2}{4}\frac{4^2\alpha^2}{\gamma^2}\Im^2(\bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{y})\,, \quad \text{or:}$$

$$2\alpha\big(||\boldsymbol{A}^\intercal \boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2\big) \geq 4\alpha^2\Im^2(\bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{y})\,.$$

Thus, it suffices to show that (assum. $\alpha > 0$):

$$||\boldsymbol{A}^\intercal \boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2 \geq 2\underbrace{\alpha}_{\text{only difference with EG, and } \alpha \in (0,1)}|\bar{\boldsymbol{x}}^\intercal \boldsymbol{A}\boldsymbol{y}|^2\,.$$

As $\alpha \in (0,1)$ and the inequality 3 is tighter, the same proof for EG (§ D.2) follows. ■

D.5. Proof of Theorem. 13: Convergence of LA3-GDA on BG

Recall that for the LA3-GDA optimizer we have the following HRDE (Eq. LA3-GDA-HRDE):

$$\begin{aligned}
\dot{\boldsymbol{z}}(t) &= \boldsymbol{\omega}(t)\\
\dot{\boldsymbol{\omega}}(t) &= -\frac{2}{\gamma}\boldsymbol{\omega}(t) - \frac{4\alpha}{\gamma}V(\boldsymbol{z}(t)) + 2\alpha\cdot J(\boldsymbol{z}(t))\cdot V(\boldsymbol{z}(t)) - \frac{2\alpha}{\gamma}V\big[\boldsymbol{z}(t) - 2\gamma V\big(\boldsymbol{z}(t)\big)\big].
\end{aligned}$$

By denoting $\dot{\boldsymbol{x}}(t) = \boldsymbol{\omega}_x(t)$ and $\dot{\boldsymbol{y}}(t) = \boldsymbol{\omega}_y(t)$ for LA3-GDA we have the following:

$$\begin{bmatrix} \dot{\boldsymbol{x}}(t) \\ \dot{\boldsymbol{y}}(t) \\ \dot{\boldsymbol{\omega}}_x(t) \\ \dot{\boldsymbol{\omega}}_y(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & \boldsymbol{I} & 0 \\ 0 & 0 & 0 & \boldsymbol{I} \\ -6\alpha\boldsymbol{A}\boldsymbol{A}^\intercal & -\frac{6\alpha}{\gamma}\boldsymbol{A} & -\frac{2}{\gamma}\boldsymbol{I} & 0 \\ \frac{6\alpha}{\gamma}\boldsymbol{A}^\intercal & -6\alpha\boldsymbol{A}^\intercal\boldsymbol{A} & 0 & -\frac{2}{\gamma}\boldsymbol{I} \end{bmatrix}}_{\triangleq \boldsymbol{C}_{\text{LA3-GDA}}} \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{y}(t) \\ \boldsymbol{\omega}_x(t) \\ \boldsymbol{\omega}_y(t) \end{bmatrix}.$$

Proof [**Proof of Thm. 13**] As in § D.2, to obtain the eigenvalues $\lambda \in \mathbb{C}$ of $\boldsymbol{C}_{\text{LA3-GDA}}$ we have:

$$\begin{aligned}
\det(\boldsymbol{C}_{\text{LA2-GDA}} - \lambda\boldsymbol{I}) &= \det\Bigg(\begin{bmatrix} -\lambda\boldsymbol{I} & 0 & \boldsymbol{I} & 0 \\ 0 & -\lambda\boldsymbol{I} & 0 & \boldsymbol{I} \\ -6\alpha\boldsymbol{A}\boldsymbol{A}^\intercal & -\frac{6\alpha}{\gamma}\boldsymbol{A} & -(\frac{2}{\gamma}+\lambda)\boldsymbol{I} & 0 \\ \frac{6\alpha}{\gamma}\boldsymbol{A}^\intercal & -6\alpha\boldsymbol{A}^\intercal\boldsymbol{A} & 0 & -(\frac{2}{\gamma}+\lambda)\boldsymbol{I} \end{bmatrix}\Bigg)\\
&= \det\Big(\lambda(\frac{2}{\gamma}+\lambda)\boldsymbol{I} - \underbrace{\begin{bmatrix} -6\alpha\boldsymbol{A}\boldsymbol{A}^\intercal & -\frac{6\alpha}{\gamma}\boldsymbol{A} \\ \frac{6\alpha}{\gamma}\boldsymbol{A}^\intercal & -6\alpha\boldsymbol{A}^\intercal\boldsymbol{A} \end{bmatrix}}_{\triangleq \boldsymbol{D}}\Big)\,.
\end{aligned}$$

We observe that only the lower left block matrix $\boldsymbol{D}$ differs from the proof for EG in § D.2, thus the inequality Eq. 2 needs to be proven for this optimizer too.

However, due to the different lower left block matrices for LA2-GDA we have:

$$\begin{aligned}\mu(\boldsymbol{z}) = \bar{\boldsymbol{z}}^\intercal \boldsymbol{D}\boldsymbol{z} &= \begin{bmatrix}\bar{\boldsymbol{x}}^\intercal & \bar{\boldsymbol{y}}^\intercal\end{bmatrix}\begin{bmatrix}-6\alpha\boldsymbol{A}\boldsymbol{A}^\intercal & -\frac{6\alpha}{\gamma}\boldsymbol{A} \\ \frac{6\alpha}{\gamma}\boldsymbol{A}^\intercal & -6\alpha\boldsymbol{A}^\intercal\boldsymbol{A}\end{bmatrix}\begin{bmatrix}\boldsymbol{x} \\ \boldsymbol{y}\end{bmatrix} \\ &= -6\alpha||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 - 6\alpha||\boldsymbol{A}\boldsymbol{y}||_2^2 + \frac{6\alpha}{\gamma}\big(\bar{\boldsymbol{y}}^\intercal\boldsymbol{A}^\intercal\boldsymbol{x} - \bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}\big) \\ &= \underbrace{-6\alpha\big(||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 - ||\boldsymbol{A}\boldsymbol{y}||_2^2\big)}_{\Re\big(\mu(\boldsymbol{z})\big)} + \underbrace{\frac{6\alpha}{\gamma}\cdot\Im(\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y})}_{\Im\big(\mu(\boldsymbol{z})\big)}\cdot i\,,\end{aligned}$$

where the last equality follows from the fact that $\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}$ is a complex conjugate of $\bar{\boldsymbol{y}}^\intercal\boldsymbol{A}^\intercal\boldsymbol{x}$, thus $\bar{\boldsymbol{y}}^\intercal\boldsymbol{A}^\intercal\boldsymbol{x} - \bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y} = 2\Im\big(\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}\big)\cdot i$. By replacing this in Eq. 2, we need to show that:

$$-6\alpha\big(||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2\big) \leq -\frac{\gamma^2}{4}\frac{6^2\alpha^2}{\gamma^2}\Im^2(\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y})\,, \quad \text{or:}$$

$$6\alpha\big(||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2\big) \geq \frac{6^2}{4}\alpha^2\Im^2(\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y})\,.$$

Thus, it suffices to show that (assum. $\alpha > 0$):

$$||\boldsymbol{A}^\intercal\boldsymbol{x}||_2^2 + ||\boldsymbol{A}\boldsymbol{y}||_2^2 \geq 2 \underbrace{\frac{3}{4}\cdot\alpha}_{\substack{\text{only difference with EG, and } \alpha\in(0,1)}} |\bar{\boldsymbol{x}}^\intercal\boldsymbol{A}\boldsymbol{y}|^2\,.$$

As $\alpha \in (0, 1)$ and the inequality (3) is tighter, the same proof as for EG (§ D.2) follows. ■

Appendix E. Analysis on MVI problems

In this section we show the convergence of the high resolution continuous time dynamics of the OGDA method for the general problem of solving monotone variational inequalities. Primarily we show the result that under some assumptions (OGDA-HRDE) and (OGDA-HRDE-2) are equivalent. We continue with the analysis in continuous time using the derived HRDEs. Using the insights of these continuous time analysis, we then show convergence of the best iterate of the discrete time OGDA method. Finally, we present an implicit discretization of OGDA-HRDE for which we prove that it converges as well.

Our main tool for analyzing the convergence of both the continuous and the discrete time is using Lyapunov Functions, defined in App. A. Recall that the OGDA-HRDE is:

$$\begin{aligned} \dot{\boldsymbol{z}}(t) &= \boldsymbol{\omega}(t) \\ \dot{\boldsymbol{\omega}}(t) &= -\beta \cdot \boldsymbol{\omega}(t) - 2J(\boldsymbol{z}(t)) \cdot \boldsymbol{\omega}(t) - \beta \cdot V(\boldsymbol{z}(t)) \end{aligned} \tag{HR-OGDA}$$

where $\beta = 2/\gamma$ is a positive real constant.

E.1. Equivalent forms of OGDA-HRDE

In this section, we first prove Thm 3, and then we show that the explicit discretization of (OGDA-HRDE-2) yields the original OGDA method.

E.1.1. PROOF OF THEOREM 3: EQUIVALENT FORMS OF OGDA-HRDE

Proof [Proof of Thm. 3] We first show that any $\boldsymbol{z}(t)$ that is a solution to (OGDA-HRDE) is also a solution to (OGDA-HRDE-2). To do this we define the function:

$$\boldsymbol{w}(t) = -\frac{2}{\beta}\dot{\boldsymbol{z}}(t) - \frac{4}{\beta}V(\boldsymbol{z}(t)) - \boldsymbol{z}(t) \Rightarrow \ddot{\boldsymbol{z}}(t) = -\frac{\beta}{2}\dot{\boldsymbol{w}}(t) - 2J(\boldsymbol{z}(t))\dot{\boldsymbol{z}}(t) - \frac{\beta}{2}\dot{\boldsymbol{z}}(t)\,.$$

Observe that $\boldsymbol{w}(t)$ is a continuous differentiable function and using (OGDA-HRDE) we have that

$$\dot{\boldsymbol{w}}(t) = \dot{\boldsymbol{z}}(t) + 2V(\boldsymbol{z}(t)). \tag{7}$$

Now if we substitute that to the definition of $\boldsymbol{w}$ then we get that:

$$\boldsymbol{w}(t) = -\frac{2}{\beta}\dot{\boldsymbol{w}}(t) - \boldsymbol{z}(t) \Rightarrow \dot{\boldsymbol{w}}(t) = -\frac{\beta}{2}\boldsymbol{z}(t) - \frac{\beta}{2}\boldsymbol{w}(t)$$

Combining the above with (7) we have that

$$\dot{\boldsymbol{z}}(t) = -\frac{\beta}{2}\boldsymbol{z}(t) - \frac{\beta}{2}\boldsymbol{w}(t) - 2V(\boldsymbol{z}(t))$$

hence the tuple $(\boldsymbol{z}(t), \boldsymbol{w}(t))$ also satisfies the system of equations (OGDA-HRDE-2).

Now for the other direction we first differentiate the first equation of (OGDA-HRDE-2) and we get that

$$\ddot{\boldsymbol{z}}(t) = -\frac{\beta}{2}\dot{\boldsymbol{z}}(t) - \frac{\beta}{2}\dot{\boldsymbol{w}}(t) - 2J(\boldsymbol{z}(t)) \cdot \dot{\boldsymbol{z}}(t)$$

then observe subtracting the two equations of (OGDA-HRDE-2) it holds that

$$\dot{\boldsymbol{w}}(t) = \dot{\boldsymbol{z}}(t) + 2V(\boldsymbol{z}(t)).$$

finally substituting the first equation to the latter one and setting $\boldsymbol{\omega}(t) = \dot{\boldsymbol{z}}(t)$ we get the equations (OGDA-HRDE). ■

E.1.2. OGDA AS AN EXPLICIT DISCRETIZATION OF OGDA-HRDE-2

We can identify the *original* OGDA method as an *explicit* discretization of the (OGDA-HRDE-2) system of differential equations. It particular, if we apply the Euler method to (OGDA-HRDE-2) with $\kappa = 1/2\gamma$ and γ to be equal to the step size of the Euler discretization then we have that

$$\begin{aligned} \boldsymbol{z}_{n+1} &= \frac{1}{2}(\boldsymbol{z}_n - \boldsymbol{w}_n) - 2\gamma V(\boldsymbol{z}_n) \\ \boldsymbol{w}_{n+1} &= \frac{1}{2}(\boldsymbol{w}_n - \boldsymbol{z}_n) \end{aligned} \tag{OGDA-S}$$

where if we eliminate $\boldsymbol{w}$ from the above system we get:

$$\boldsymbol{z}_{n+1} = \boldsymbol{z}_n - 2\gamma V(\boldsymbol{z}_n) + \gamma V(\boldsymbol{z}_{n-1}). \tag{OGDA}$$

E.2. Proof of Theorem 4: Last-iterate Convergence of OGDA-HRDE for MVIs

We will use the following Lyapunov function:

$$\mathcal{L}_{\text{OGDA}}(\boldsymbol{z}, \boldsymbol{\omega}) = \|b\boldsymbol{z} + \boldsymbol{\omega}\|_2^2 + \|\boldsymbol{\omega}\|_2^2 + 4b\boldsymbol{z}^\intercal V(\boldsymbol{z}) + \|V(\boldsymbol{z}) + \boldsymbol{\omega}\|_2^2 + \|V(\boldsymbol{z})\|_2^2. \tag{OGDA-L}$$

In particular, we are going to split the above Lyapunov function in the following two parts

$$\mathcal{L}_1(\boldsymbol{z}, \boldsymbol{\omega}) = \frac{1}{2}\left(\|b\boldsymbol{z} + \boldsymbol{\omega}\|_2^2 + \|\boldsymbol{\omega}\|_2^2 + 4b\boldsymbol{z}^\intercal V(\boldsymbol{z})\right) \tag{OGDA-L1}$$

$$\mathcal{L}_2(\boldsymbol{z}, \boldsymbol{\omega}) = \frac{1}{2}\left(\|V(\boldsymbol{z}) + \boldsymbol{\omega}\|_2^2 + \|V(\boldsymbol{z})\|_2^2\right). \tag{OGDA-L2}$$

From the definition of the monotone variational inequality problem and from the positivity of the norms we have that $\mathcal{L}_1(\boldsymbol{z}, \boldsymbol{\omega}) \geq 0$ and $\mathcal{L}_2(\boldsymbol{z}, \boldsymbol{\omega}) \geq 0$, and also $\mathcal{L}_1(0, 0) = \mathcal{L}_2(0, 0) = 0$. Next we are going to explore $\dot{\mathcal{L}}_1$ and $\dot{\mathcal{L}}_2$.

$$\dot{\mathcal{L}}_1(\boldsymbol{z}, \boldsymbol{\omega}) = \underbrace{\langle b\boldsymbol{z} + \boldsymbol{\omega}, b\dot{\boldsymbol{z}} + \dot{\boldsymbol{\omega}}\rangle}_{1^\circ} + \underbrace{\langle \boldsymbol{\omega}, \dot{\boldsymbol{\omega}}\rangle}_{2^\circ} + \underbrace{2b\dot{\boldsymbol{z}}^\intercal V(\boldsymbol{z})}_{3^\circ} + \underbrace{2b\boldsymbol{z}^\intercal \dot{\big(V(\boldsymbol{z})\big)}}_{4^\circ} \tag{8}$$

$$\dot{\mathcal{L}}_2(\boldsymbol{z}, \boldsymbol{\omega}) = \underbrace{\langle V(\boldsymbol{z}) + \boldsymbol{\omega}, \dot{\big(V(\boldsymbol{z})\big)} + \dot{\boldsymbol{\omega}}\rangle}_{5^\circ} + \underbrace{\langle V(\boldsymbol{z}), \dot{\big(V(\boldsymbol{z})\big)}\rangle}_{6^\circ} \tag{9}$$

We are now going to analyze each of these terms separately, using the following equality

$$\dot{\big(V(\boldsymbol{z})\big)} = J(\boldsymbol{z})\dot{\boldsymbol{z}} = J(\boldsymbol{z}) \cdot \boldsymbol{\omega}. \tag{10}$$

1°) Replacing $\dot{\boldsymbol{\omega}}$ of OGDA-HRDE, we get $\langle b\boldsymbol{z} + \boldsymbol{\omega}, b\dot{\boldsymbol{z}} + \dot{\boldsymbol{\omega}}\rangle = \langle b\boldsymbol{z} + \boldsymbol{\omega}, -bV(\boldsymbol{z}) - 2J(\boldsymbol{z})\boldsymbol{\omega}\rangle$. Simplifying it then gives: $-b^2\boldsymbol{z}^\intercal V(\boldsymbol{z}) - b\boldsymbol{\omega}^\intercal V(\boldsymbol{z}) - 2b\boldsymbol{z}^\intercal J(\boldsymbol{z})\boldsymbol{\omega} - 2\boldsymbol{\omega}^\intercal J(\boldsymbol{z})\boldsymbol{\omega}$;

2°) Similarly, by replacing $\dot{\boldsymbol{\omega}}$ of OGDA-HRDE, we get $\langle \boldsymbol{\omega}, \dot{\boldsymbol{\omega}}\rangle = -b\|\boldsymbol{\omega}\|_2^2 - b\boldsymbol{\omega}^\intercal V(\boldsymbol{z}) - 2\boldsymbol{\omega}^\intercal J(\boldsymbol{z})\boldsymbol{\omega}$;

3°) $2b\dot{\boldsymbol{z}}^\intercal V(\boldsymbol{z}) = 2b\boldsymbol{\omega}^\intercal V(\boldsymbol{z})$;

4°) Using the fact (10) we get: $2b\boldsymbol{z}^\intercal \dot{\big(V(\boldsymbol{z})\big)} = 2b\boldsymbol{z}^\intercal J(\boldsymbol{z})\boldsymbol{\omega}$;

5°) Using (OGDA-HRDE) and 10 we get: $\langle V(\boldsymbol{z}) + \boldsymbol{\omega}, \dot{\big(V(\boldsymbol{z})\big)} + \dot{\boldsymbol{\omega}} \rangle = \langle V(\boldsymbol{z}) + \boldsymbol{\omega}, -b\big(V(\boldsymbol{z}) + \boldsymbol{\omega}\big) - J(\boldsymbol{z})\boldsymbol{\omega} \rangle = -b||V(\boldsymbol{z}) + \boldsymbol{\omega}||_2^2 - V^\intercal(\boldsymbol{z})J(\boldsymbol{z})\boldsymbol{\omega} - \boldsymbol{\omega}^\intercal J(\boldsymbol{z})\boldsymbol{\omega}$; and

6°) Using Eq. OGDA-HRDE we get: $\langle V(\boldsymbol{z}), \dot{\big(V(\boldsymbol{z})\big)} \rangle = V^\intercal(\boldsymbol{z})J(\boldsymbol{z})\boldsymbol{\omega}$.

Using all the above we get that

$$\dot{\mathcal{L}}_1(\boldsymbol{z}, \boldsymbol{\omega}) = -b \left\|\boldsymbol{\omega}\right\|_2^2 - b^2 \boldsymbol{z}^\intercal V(\boldsymbol{z}) - 4\boldsymbol{\omega}^\intercal J(\boldsymbol{z})\boldsymbol{\omega}. \tag{11}$$

$$\dot{\mathcal{L}}_2(\boldsymbol{z}, \boldsymbol{\omega}) = -b \left\|V(\boldsymbol{z}) + \boldsymbol{\omega}\right\|_2^2 - \boldsymbol{\omega}^\intercal J(\boldsymbol{z})\boldsymbol{\omega}. \tag{12}$$

Using the monotonicity of V and the positive definiteness of J we have that $\dot{\mathcal{L}}_1(\boldsymbol{z}, \boldsymbol{\omega}) \leq 0$ and $\dot{\mathcal{L}}_2(\boldsymbol{z}, \boldsymbol{\omega}) \leq 0$. Hence, both $\mathcal{L}_1$ and $\mathcal{L}_2$ are Lyapunov functions for our problem. Therefore by observing that $\mathcal{L}_{\text{OGDA}}(\boldsymbol{z}, \boldsymbol{\omega}) = 2\mathcal{L}_1(\boldsymbol{z}, \boldsymbol{\omega}) + 2\mathcal{L}_2(\boldsymbol{z}, \boldsymbol{\omega})$, we have that $\mathcal{L}_{\text{OGDA}}$ is also a Lyapunov function of our problem. To prove the convergence rate we start with the following lemma.

Lemma 14 *We assume the initial conditions $\boldsymbol{z}(0) = \boldsymbol{z}_0$ and $\boldsymbol{\omega}(0) = 0$. If for all $t \in [0, T]$ it holds that $\max\{\left\|\boldsymbol{\omega}(t)\right\|_2, \left\|V(\boldsymbol{z}(t)\right\|_2\} \geq \epsilon$ then we have that*

$$T \leq \left(b + 8 + \left(8 + \frac{2}{b}\right) \cdot L\right) \cdot \frac{\left\|\boldsymbol{z}_0\right\|_2^2}{\epsilon^2}.$$

Proof With abuse of notation we use $\mathcal{L}_{\text{OGDA}}(t) \triangleq \mathcal{L}_{\text{OGDA}}(\boldsymbol{z}(t), \boldsymbol{\omega}(t))$ and we have that

$$\begin{aligned}
\mathcal{L}_{\text{OGDA}}(0) &= b^2 \left\|\boldsymbol{z}_0\right\|_2^2 + 4b\boldsymbol{z}_0^\intercal V(\boldsymbol{z}_0) + 2\left\|V(\boldsymbol{z}_0)\right\|_2^2 \\
&\leq (b^2 + 8b)\left\|\boldsymbol{z}_0\right\|_2^2 + (2 + 8b)\left\|V(\boldsymbol{z}_0)\right\|_2^2 \\
&\leq (b^2 + 8b + (8b + 2) \cdot L)\left\|\boldsymbol{z}_0\right\|_2^2
\end{aligned} \tag{13}$$

where in the last inequality we used the Lipschitzness of V and the fact that $\boldsymbol{z}^\star = 0$ is a solution and hence $V(0) = 0$.

Now using (11), (12) together with the monotonicity of V and the positive definiteness of J we get that

$$\begin{aligned}
\dot{\mathcal{L}}_{\text{OGDA}}(t) &\leq -2b\left(\left\|\boldsymbol{\omega}(t)\right\|_2^2 + \left\|V(\boldsymbol{z}(t)) + \boldsymbol{\omega}\right\|_2^2\right) \\
&\leq -b\left(\left\|V(\boldsymbol{z}(t))\right\|_2^2\right)
\end{aligned}$$

where we have used the fact that for any vectors $\boldsymbol{x}$, $\boldsymbol{y}$ it holds that $\left\|\boldsymbol{x} + \boldsymbol{y}\right\|_2^2 \leq 2\left\|\boldsymbol{x}\right\|_2^2 + 2\left\|\boldsymbol{y}\right\|_2^2$. Also, we have that

$$\dot{\mathcal{L}}_{\text{OGDA}}(t) \leq -2b\left(\left\|\boldsymbol{\omega}(t)\right\|_2^2\right)$$

hence overall we have that

$$\dot{\mathcal{L}}_{\text{OGDA}}(t) \leq -b(\max\{\left\|\boldsymbol{\omega}(t)\right\|_2^2, \left\|V(\boldsymbol{z}(t))\right\|_2^2\}) \tag{14}$$

Now if for every $t \in [0, T]$ it holds that $\max\{\left\|\boldsymbol{\omega}(t)\right\|_2, \left\|V(\boldsymbol{z}(t)\right\|_2\} \geq \epsilon$ then we have that for all $t \in [0, T]$ it holds

$$\dot{\mathcal{L}}_{\text{OGDA}}(t) \leq -b \cdot \epsilon^2.$$

Now since $\mathcal{L}_{\text{OGDA}}(t) \geq 0$ and using the above upper bound on the time derivative of $\mathcal{L}$ together with the Mean Value Theorem we have that

$$(\mathcal{L}_{\text{OGDA}}(0) - \mathcal{L}_{\text{OGDA}}(t)) \geq (b \cdot \epsilon^2) \cdot T.$$

Finally using the bound on the initial conditions (13) the lemma follows. ■

Lemma 14 shows that after a finite time both $\|\boldsymbol{\omega}(t)\|_2$ and $\|V(\boldsymbol{z}(t))\|_2$ will become less than ϵ. What remains to show the last iterate convergence is to show that this upper bound will remain for later times. For this we show the following lemma.

Lemma 15 *Let $t^\star > 0$ a time such that both $\|\boldsymbol{\omega}(t^\star)\|_2 \leq \epsilon$ and $\|V(\boldsymbol{z}(t^\star))\|_2 \leq \epsilon$, then we have that for all $t > t^\star$ it holds that*

$$\|V(\boldsymbol{z}(t))\|_2 \leq \sqrt{5}\epsilon.$$

Proof Using the assumptions of the lemma we have the following

1. $\mathcal{L}_2(t^\star) \leq \frac{5}{2}\epsilon^2$,
2. $\dot{\mathcal{L}}_2(t) \leq 0$, and
3. $\|V(\boldsymbol{z}(t))\|_2 \leq \sqrt{2\mathcal{L}_2(t)}$

Combining the first two properties we have that for all $t > t^\star$ it holds that $\mathcal{L}_2(t) \leq \frac{5}{2}\epsilon^2$ and if we apply property 3. to this the lemma follows. ■

Now if we combine Lemma 14 and Lemma 15, then the first part of Theorem 4 follows.

Strongly Monotone Variational Inequalities. For the second part of Theorem 4 we observe that

$$\mathcal{L}_{\text{OGDA}}(t) \leq 2b^2 \|Z(t)\|_2^2 + 5\|W(t)\|_2^2 + 2\|V(\boldsymbol{z}(t))\|_2^2 + 4b(\boldsymbol{z}(t))^\intercal V(\boldsymbol{z}(t)). \tag{15}$$

Also, using the strong monotonicity of V and the fact that $V(0) = 0$, which we have assumed without loss of generality, then we have that $(\boldsymbol{z}(t))^\intercal V(\boldsymbol{z}(t)) \geq \mu\|\boldsymbol{z}(t)\|_2^2$. Applying this together with (11) and (12) we get all the following upper bounds for the time derivative of $\mathcal{L}_{\text{OGDA}}$

$$\frac{1}{\mu} \cdot \dot{\mathcal{L}}_{\text{OGDA}}(t) \leq -\frac{2}{\mu} b^2 (\boldsymbol{z}(t))^\intercal V(\boldsymbol{z}(t)) \leq -2b^2 \cdot \|\boldsymbol{z}(t)\|_2^2 \tag{16}$$

$$\frac{5}{2b} \cdot \dot{\mathcal{L}}_{\text{OGDA}}(t) \leq -5\|\boldsymbol{\omega}(t)\|_2^2 \tag{17}$$

$$\begin{aligned}\frac{2}{b} \cdot \dot{\mathcal{L}}_{\text{OGDA}}(t) &\leq -4\|\boldsymbol{\omega}(t)\|_2^2 - 4\|V(\boldsymbol{z}(t)) + \boldsymbol{\omega}(t)\|_2^2 - 4b(\boldsymbol{z}(t))^\intercal V(\boldsymbol{z}(t)) \\ &\leq -2\|V(\boldsymbol{z}(t))\|_2^2 - 4b(\boldsymbol{z}(t))^\intercal V(Z(t))\end{aligned} \tag{18}$$

Combining these with (15) and setting $\kappa = \frac{1}{\mu} + \frac{9}{2b}$ we get that

$$\mathcal{L}_{\text{OGDA}}(t) \leq -\kappa \cdot \dot{\mathcal{L}}_{\text{OGDA}}(t)$$

where if we use the fact that $\mathcal{L}_{\text{OGDA}}(t) \geq 0$ we get that

$$-\kappa \cdot \frac{\dot{\mathcal{L}}_{\text{OGDA}}(t)}{\mathcal{L}_{\text{OGDA}}(t)} \geq 1$$

and hence

$$-\kappa \frac{d}{dt} \left(\ln(\mathcal{L}_{\text{OGDA}}(t)) \right) \geq 1$$

now integrating both parts from 0 to t we get

$$\ln(\mathcal{L}_{\text{OGDA}}(t)) \leq \ln(\mathcal{L}_{\text{OGDA}}(0)) - \frac{1}{\kappa} t$$

then applying the exponential function we get

$$\mathcal{L}_{\text{OGDA}}(t) \leq \mathcal{L}_{\text{OGDA}}(0) \cdot \exp\left(-\frac{1}{\kappa} \cdot t \right)$$

and using (13) together with the fact that $\mathcal{L}_{\text{OGDA}}(t) \geq \|V(\boldsymbol{z}(t)\|_2^2 + 4b\mu \|Z(t)\|_2^2$ the second part of Theorem 4 follows.

E.3. Proof of Theorem 5: Last-iterate Convergence of OGDA-HRDE-2 for MVIs

We are going to use the following Lyapunov function

$$\mathcal{L}_{\text{OGDA2}}(\boldsymbol{z}, \boldsymbol{w}) = \kappa^2 \|\boldsymbol{z} + \boldsymbol{w}\|_2^2 + \kappa^2 \|\boldsymbol{z} - \boldsymbol{w}\|_2^2 + \|\kappa(\boldsymbol{z} + \boldsymbol{w}) + V(\boldsymbol{z})\|_2^2 + \|V(\boldsymbol{z})\|_2^2 . \quad \text{(OGDA2-L)}$$

In particular, we are going to split the above Lyapunov function in the following two parts

$$\mathcal{L}_3(\boldsymbol{z}, \boldsymbol{w}) = \frac{1}{2} \left(\|\boldsymbol{z} + \boldsymbol{w}\|_2^2 + \|\boldsymbol{z} - \boldsymbol{w}\|_2^2 \right) \quad \text{(OGDA2-L3)}$$

$$\mathcal{L}_4(\boldsymbol{z}, \boldsymbol{\omega}) = \frac{1}{2} \left(\|\kappa(\boldsymbol{z} + \boldsymbol{w}) + V(\boldsymbol{z})\|_2^2 + \|V(\boldsymbol{z})\|_2^2 \right) . \quad \text{(OGDA2-L4)}$$

From the definition of the monotone variational inequality problem and from the positivity of the norms we have that $\mathcal{L}_3(\boldsymbol{z}, \boldsymbol{w}) > 0$ and $\mathcal{L}_4(\boldsymbol{z}, \boldsymbol{\omega}) > 0$ for any $(\boldsymbol{z}, \boldsymbol{w}) \neq (0, 0)$, and also $\mathcal{L}_3(0, 0) = \mathcal{L}_4(0, 0) = 0$. Next we are going to explore $\dot{\mathcal{L}}_3$ and $\dot{\mathcal{L}}_4$.

$$\dot{\mathcal{L}}_3(\boldsymbol{z}, \boldsymbol{\omega}) = \underbrace{\langle \beta \boldsymbol{z} + \boldsymbol{w}, \beta \dot{\boldsymbol{z}} + \dot{\boldsymbol{w}} \rangle}_{1^\circ} + \underbrace{\langle \beta \boldsymbol{z} - \boldsymbol{w}, \beta \dot{\boldsymbol{z}} - \dot{\boldsymbol{w}} \rangle}_{2^\circ} \quad (19)$$

$$\dot{\mathcal{L}}_4(\boldsymbol{z}, \boldsymbol{\omega}) = \underbrace{\langle \kappa(\boldsymbol{z} + \boldsymbol{w}) + V(\boldsymbol{z}), \kappa(\dot{\boldsymbol{z}} + \dot{\boldsymbol{w}}) + \dot{(V(\boldsymbol{z}))} \rangle}_{3^\circ} + \underbrace{\langle V(\boldsymbol{z}), \dot{(V(\boldsymbol{z}))} \rangle}_{4^\circ} \quad (20)$$

We are now going to analyze each of these terms separately, using the following inequality, which straightforwardly follows from the monotonicity of V

$$\langle \dot{\boldsymbol{z}}, \dot{(V(\boldsymbol{z}))} \rangle \geq 0. \quad (21)$$

1°) Replacing $\dot{\boldsymbol{z}}$ and $\dot{\boldsymbol{w}}$ from OGDA-HRDE-2, we get $\langle \beta \boldsymbol{z} + \boldsymbol{w}, \beta \dot{\boldsymbol{z}} + \dot{\boldsymbol{w}} \rangle = -\kappa \|\boldsymbol{z} + \boldsymbol{w}\|_2^2 - 2\langle \boldsymbol{z}, V(\boldsymbol{z}) \rangle - 2\langle \boldsymbol{w}, V(\boldsymbol{z}) \rangle$;

2°) Replacing $\dot{\boldsymbol{z}}$ and $\dot{\boldsymbol{w}}$ from OGDA-HRDE-2, we get $\langle \beta \boldsymbol{z} - \boldsymbol{w}, \beta \dot{\boldsymbol{z}} - \dot{\boldsymbol{w}} \rangle = -2\langle \boldsymbol{z}, V(\boldsymbol{z}) \rangle + 2\langle \boldsymbol{w}, V(\boldsymbol{z}) \rangle$;

3°) Similarly we have that the time derivative of the term 3° is $-2\kappa \left\|\kappa(\boldsymbol{z}+\boldsymbol{w})+V(\boldsymbol{z})\right\|_2^2 - 2\langle \dot{\boldsymbol{z}}, \dot{\big(V(\boldsymbol{z})\big)}\rangle - \langle V(\boldsymbol{z}), \dot{\big(V(\boldsymbol{z})\big)}\rangle$;

4°) The derivative is directly equal to $\langle V(\boldsymbol{z}), \dot{\big(V(\boldsymbol{z})\big)}\rangle$.

Using all the above we get that

$$\dot{\mathcal{L}}_3(\boldsymbol{z},\boldsymbol{w}) = -\kappa \left\|\boldsymbol{z}+\boldsymbol{w}\right\|_2^2 - 4\langle \boldsymbol{z}, V(\boldsymbol{z})\rangle \tag{22}$$

$$\dot{\mathcal{L}}_4(\boldsymbol{z},\boldsymbol{w}) = -2\kappa \left\|\kappa(\boldsymbol{z}+\boldsymbol{w})+V(\boldsymbol{z})\right\|_2^2 - 2\langle \dot{\boldsymbol{z}}, \dot{\big(V(\boldsymbol{z})\big)}\rangle. \tag{23}$$

We define the set $S = \{(\boldsymbol{z},\boldsymbol{w}) \in \mathbb{R}^{2d} : V(\boldsymbol{z}) = 0\}$. Using the monotonicity of V and (21) we have that $\dot{\mathcal{L}}_1(\boldsymbol{z},\boldsymbol{\omega}) \leq 0$, $\dot{\mathcal{L}}_2(\boldsymbol{z},\boldsymbol{\omega}) \leq 0$ for all $(\boldsymbol{z},\boldsymbol{w}) \in \mathbb{R}^{2d}$ and also $\dot{\mathcal{L}}_{\text{OGDA2}} < 0$ for all $(\boldsymbol{z},\boldsymbol{w}) \notin S$. Hence, both $\mathcal{L}_1$ and $\mathcal{L}_2$ are Lyapunov functions for our problem. To prove the convergence rate we start with the following lemma.

Lemma 16 *We assume the initial conditions $\boldsymbol{z}(0) = \boldsymbol{z}_0$ and $\boldsymbol{w}(0) = -\boldsymbol{z}_0$. If for all $t \in [0,T]$ it holds that $\max\{\left\|\kappa(\boldsymbol{z}(t)+\boldsymbol{w}(t))\right\|_2, \left\|V(\boldsymbol{z}(t)\right\|_2\} \geq \epsilon$ then we have that*

$$T \leq 2\left(2\kappa + \frac{L^2}{\kappa}\right) \cdot \frac{\left\|\boldsymbol{z}_0\right\|_2^2}{\epsilon^2}.$$

Proof With abuse of notation we use $\mathcal{L}_{\text{OGDA2}}(t) \triangleq \mathcal{L}_{\text{OGDA2}}(\boldsymbol{z}(t),\boldsymbol{\omega}(t))$ and we have that

$$\begin{aligned}\mathcal{L}_{\text{OGDA2}}(0) &= 4\cdot\kappa^2 \left\|\boldsymbol{z}_0\right\|_2^2 + 2\left\|V(\boldsymbol{z}_0)\right\|_2^2 \\ &\leq (4\cdot\kappa^2 + 2\cdot L^2)\left\|\boldsymbol{z}_0\right\|_2^2\end{aligned} \tag{24}$$

where in the last inequality we used the Lipschitzness of V and the fact that $\boldsymbol{z}^\star = 0$ is a solution and hence $V(0) = 0$.

Now using (22), (23) together with (21) we have that

$$\begin{aligned}\dot{\mathcal{L}}_{\text{OGDA2}}(t) &\leq -\kappa\left(\left\|\kappa(\boldsymbol{z}(t)+\boldsymbol{w}(t))\right\|_2^2 + \left\|\kappa(\boldsymbol{z}(t)+\boldsymbol{w}(t)) + V(\boldsymbol{z}(t))\right\|_2^2\right) \\ &\leq -\kappa\left(\left\|V(\boldsymbol{z}(t))\right\|_2^2\right)\end{aligned}$$

where we have used the fact that for any vectors $\boldsymbol{x}$, $\boldsymbol{y}$ it holds that $\left\|\boldsymbol{x}+\boldsymbol{y}\right\|_2^2 \leq 2\left\|\boldsymbol{x}\right\|_2^2 + 2\left\|\boldsymbol{y}\right\|_2^2$. Also, we have that

$$\dot{\mathcal{L}}_{\text{OGDA2}}(t) \leq -\kappa\left(\left\|\kappa(\boldsymbol{z}(t)+\boldsymbol{w}(t))\right\|_2^2\right)$$

hence overall we have that

$$\dot{\mathcal{L}}_{\text{OGDA2}}(t) \leq -\kappa\left(\max\{\left\|\kappa(\boldsymbol{z}(t)+\boldsymbol{w}(t))\right\|_2, \left\|V(\boldsymbol{z}(t)\right\|_2\}\right) \tag{25}$$

Now if for every $t \in [0,T]$ it holds that $\max\{\left\|\kappa(\boldsymbol{z}(t)+\boldsymbol{w}(t))\right\|_2, \left\|V(\boldsymbol{z}(t)\right\|_2\} \geq \epsilon$ then we have that for all $t \in [0,T]$ it holds

$$\dot{\mathcal{L}}_{\text{OGDA2}}(t) \leq -\kappa\cdot\epsilon^2.$$

Now since $\mathcal{L}_{\text{OGDA2}}(t) \geq 0$ and using the above upper bound on the time derivative of $\mathcal{L}$ together with the Mean Value Theorem we have that

$$(\mathcal{L}_{\text{OGDA2}}(0) - \mathcal{L}_{\text{OGDA2}}(t)) \geq (\kappa \cdot \epsilon^2) \cdot T.$$

Finally using the bound on the initial conditions (24) the lemma follows. ■

What remains is to show the last-iterate convergence is to show that this upper bound will remain for later times. For this we show the following lemma.

Lemma 17 *Let $t^\star > 0$ a time such that both $\|\kappa(\boldsymbol{z}(t^\star + \boldsymbol{w}(t^\star))\|_2 \leq \epsilon$ and $\|V(\boldsymbol{z}(t^\star))\|_2 \leq \epsilon$, then we have that for all $t > t^\star$ it holds that*

$$\|V(\boldsymbol{z}(t))\|_2 \leq \sqrt{3}\epsilon.$$

Proof Using the assumptions of the lemma we have the following

1. $\mathcal{L}_4(t^\star) \leq \frac{3}{2}\epsilon^2$,
2. $\dot{\mathcal{L}}_4(t) \leq 0$, and
3. $\|V(\boldsymbol{z}(t))\|_2 \leq \sqrt{2\mathcal{L}_4(t)}$

Combining the first two properties we have that for all $t > t^\star$ it holds that $\mathcal{L}_4(t) \leq \frac{3}{2}\epsilon^2$ and if we apply property 3. to this the lemma follows. ■

Now if we combine Lemma 16 and Lemma 17, then the first part of Theorem 5 follows and we omit the details of the second part because it is almost identical with the corresponding argument for proving the second part of Theorem 4.

E.4. Proof of Theorem 6: Best-iterate convergence of OGDA for MVIs

Inspired by the analysis of the continuous-time dynamics that we presented in the previous section we will use the following Lyapunov function

$$\mathcal{L}_5(\boldsymbol{z}, \boldsymbol{w}) = \|\boldsymbol{z} - \boldsymbol{w}\|_2^2 + \|\boldsymbol{z} + \boldsymbol{w} + 2\gamma V(\boldsymbol{z})\|_2^2. \tag{OGDA-L5}$$

Before analyzing the above Lyapunov function we show the following useful lemma.

Lemma 18 *Let $\boldsymbol{z}_0 = \boldsymbol{z}_1$ and $\boldsymbol{z}_n$ follows the (OGDA) dynamics for $n \geq 2$, also assume that V is L-Lipschitz and that $\gamma \leq 1/8L$ then for all $n \geq 0$ it holds that*

$$\frac{1}{2} \leq \frac{\|V(\boldsymbol{z}_{n+1})\|_2}{\|V(\boldsymbol{z}_n)\|_2} \leq \frac{3}{2}.$$

Proof Using the Lipschitzness of V we have that

$$\begin{aligned}
\|V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\|_2 &\leq L \cdot \|\boldsymbol{z}_{n+1} - \boldsymbol{z}_n\|_2 \\
&= L \cdot \gamma \cdot \|2V(\boldsymbol{z}_n) - V(\boldsymbol{z}_{n-1})\|_2 \\
&\leq L \cdot \gamma \cdot (2\|V(\boldsymbol{z}_n)\|_2 + \|V(\boldsymbol{z}_{n-1})\|_2). \qquad (26)
\end{aligned}$$

Using the triangle inequality together with the above we have that

$$\begin{aligned}\|V(\boldsymbol{z}_{n+1})\|_2 &\leq \|V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\|_2 + \|V(\boldsymbol{z}_n)\|_2 \\ &= (1 + 2 \cdot L \cdot \gamma) \cdot \|V(\boldsymbol{z}_n)\|_2 + L \cdot \gamma \|V(\boldsymbol{z}_{n-1})\|_2 \end{aligned} \tag{27}$$

and also that

$$\begin{aligned}\|V(\boldsymbol{z}_{n+1})\|_2 &\geq \|V(\boldsymbol{z}_n)\|_2 - \|V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\|_2 \\ &= (1 - 2 \cdot L \cdot \gamma) \cdot \|V(\boldsymbol{z}_n)\|_2 - L \cdot \gamma \|V(\boldsymbol{z}_{n-1})\|_2 . \end{aligned} \tag{28}$$

Now let $b_n = \frac{\|V(\boldsymbol{z}_n)\|_2}{\|V(\boldsymbol{z}_{n-1})\|_2}$ then we can re-write (27) and (28) as follows

$$b_{n+1} \leq (1 + 2 \cdot L \cdot \gamma) + \frac{L \cdot \gamma}{b_n} \tag{29}$$

$$b_{n+1} \geq (1 - 2 \cdot L \cdot \gamma) - \frac{L \cdot \gamma}{b_n} \tag{30}$$

We will use induction to show that $b_n \in [1/2, 3/2]$. The base step holds since $\boldsymbol{z}_0 = \boldsymbol{z}_1$ and hence $b_1 = 1$. For the inductive step we assume that $b_n \in [1/2, 3/2]$ and from (29) we have that

$$b_{n+1} \leq 1 + 4 \cdot L \cdot \gamma$$

whereas from (30) we have that

$$b_{n+1} \geq 1 - 4 \cdot L \cdot \gamma$$

so it suffices to have that $\gamma \leq 1/(8 \cdot L)$. ■

We continue with the analysis of the Lyapunov function $\mathcal{L}_5$. It is a matter of algebraic calculations to verify that

$$\begin{aligned}\mathcal{L}_5(\boldsymbol{z}_{n+1}, \boldsymbol{w}_{n+1}) - \mathcal{L}_5(\boldsymbol{z}_n, \boldsymbol{w}_n) = &- \|\boldsymbol{z}_n + \boldsymbol{w}_n\|_2^2 - 8\gamma\langle \boldsymbol{z}_n, V(\boldsymbol{z}_n)\rangle \\ &+ 4\gamma^2 \|V(\boldsymbol{z}_{n+1})\|_2^2 + 4\gamma^2 \|V(\boldsymbol{z}_n)\|_2^2 - 8\gamma^2 \langle V(\boldsymbol{z}_n), V(\boldsymbol{z}_{n+1})\rangle\end{aligned}$$

adding the two equations of (OGDA-S) we can see that $\boldsymbol{z}_{n+1} + \boldsymbol{w}_{n+1} = 2\gamma V(\boldsymbol{z}_n)$, from which we can get that

$$\begin{aligned}\mathcal{L}_5(\boldsymbol{z}_{n+1}, \boldsymbol{w}_{n+1}) - \mathcal{L}_5(\boldsymbol{z}_n, \boldsymbol{w}_n) &= -8\gamma\langle \boldsymbol{z}_n, V(\boldsymbol{z}_n)\rangle - 4\gamma^2 \Big(\|V(\boldsymbol{z}_{n-1})\|_2^2 - \|V(\boldsymbol{z}_{n+1})\|_2^2 - \|V(\boldsymbol{z}_n)\|_2^2 + 2\langle V(\boldsymbol{z}_n), V(\boldsymbol{z}\\ &= -8\gamma\langle \boldsymbol{z}_n, V(\boldsymbol{z}_n)\rangle - 4\gamma^2 \Big(\|V(\boldsymbol{z}_{n-1})\|_2^2 - \|V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\|_2^2 \Big) \\ &= -8\gamma\langle \boldsymbol{z}_n, V(\boldsymbol{z}_n)\rangle - 2\gamma^2 \|V(\boldsymbol{z}_{n-1})\|_2^2 - 2\gamma^2 \Big(\|V(\boldsymbol{z}_{n-1})\|_2^2 - 2\|V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\| \end{aligned}$$

now using (26) we have that

$$\|V(\boldsymbol{z}_{n-1})\|_2^2 - 2\|V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\|_2^2 \geq (1 - 2 \cdot L \cdot \gamma) \cdot \|V(\boldsymbol{z}_{n-1})\|_2^2 - 4 \cdot L \cdot \gamma \|V(\boldsymbol{z}_n)\|_2^2$$

now assuming that $\gamma \leq 1/(4 \cdot L)$ we have that

$$\|V(\boldsymbol{z}_{n-1})\|_2^2 - 2\|V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\|_2^2 \geq \frac{1}{2} \cdot \|V(\boldsymbol{z}_{n-1})\|_2^2 - 4 \cdot L \cdot \gamma \|V(\boldsymbol{z}_n)\|_2^2$$

now using Lemma 18 we get that

$$\|V(\boldsymbol{z}_{n-1})\|_2^2 - 2\|V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\|_2^2 \geq \left(\frac{1}{4} - 4 \cdot L \cdot \gamma\right) \cdot \|V(\boldsymbol{z}_{n-1})\|_2^2$$

and hence since $\gamma \leq 1/(16 \cdot L)$ this implies that

$$\|V(\boldsymbol{z}_{n-1})\|_2^2 - 2\|V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\|_2^2 \geq 0$$

If we know substitute the latter in the last expression about the difference $\mathcal{L}_5(\boldsymbol{z}_{n+1}, \boldsymbol{w}_{n+1}) - \mathcal{L}_5(\boldsymbol{z}_n, \boldsymbol{w}_n)$ then we get the following

$$\mathcal{L}_5(\boldsymbol{z}_{n+1}, \boldsymbol{w}_{n+1}) - \mathcal{L}_5(\boldsymbol{z}_n, \boldsymbol{w}_n) \leq -8\gamma\langle \boldsymbol{z}_n, V(\boldsymbol{z}_n)\rangle - 2\gamma^2\|V(\boldsymbol{z}_{n-1})\|_2^2$$

but due to the monotonicity of V we have that $\langle \boldsymbol{z}_n, V(\boldsymbol{z}_n)\rangle \geq 0$ and therefore we have that

$$\mathcal{L}_5(\boldsymbol{z}_{n+1}, \boldsymbol{w}_{n+1}) - \mathcal{L}_5(\boldsymbol{z}_n, \boldsymbol{w}_n) \leq -2\gamma^2\|V(\boldsymbol{z}_{n-1})\|_2^2. \tag{31}$$

We next bound the initial value of $\mathcal{L}_5$. The initial conditions that $\boldsymbol{z}_0 = \boldsymbol{z}_1$ are equivalent with $\boldsymbol{w}_0 = -\boldsymbol{z}_0 - 4\gamma V(\boldsymbol{z}_0)$ and hence we have that

$$\begin{aligned}\mathcal{L}_5(\boldsymbol{z}_0, \boldsymbol{w}_0) &= \|2\boldsymbol{z}_0 + 4\gamma V(\boldsymbol{z}_0)\|_2^2 + 4\gamma^2\|V(\boldsymbol{z}_0)\|_2^2 \\ &\leq 8\|\boldsymbol{z}_0\|_2^2 + 36\gamma^2\|V(\boldsymbol{z}_0)\|_2^2 \\ &\leq (8 + 36\gamma^2 L^2)\|\boldsymbol{z}_0\|_2^2 \end{aligned} \tag{32}$$

Now combining (32) and (31) the Theorem 6 follows.

E.5. Proof of Theorem 7: Last-iterate convergence of an implicitdiscretization of OGDA-HRDE

In this section, we analyze the implicit discretization of the OGDA continuous-time dynamics that we presented in the previous section, for with implicit discretization we integrate the (OGDA-HRDE) from $t = n \cdot \gamma$ to $t' = (n+1) \cdot \gamma$, we use $\boldsymbol{z}_n = \boldsymbol{z}(n \cdot \gamma)$ and we make use of the following approximations:

- ▷ $\int_t^{t'} \boldsymbol{\omega}(\tau)d\tau \approx \frac{\gamma}{2}(\boldsymbol{\omega}_{n+1} + \boldsymbol{\omega}_n)$,
- ▷ $\int_t^{t'} J(\boldsymbol{z}(\tau))) \cdot \boldsymbol{\omega}(\tau)d\tau = \int_t^{t'} \frac{d}{d\tau}(V(\boldsymbol{z}(\tau)))d\tau = V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}(n))$,
- ▷ $\int_t^{t'} V(\boldsymbol{z}(\tau))d\tau \approx \frac{\gamma}{2}(V(\boldsymbol{z}_{n+1}) + V(\boldsymbol{z}_n))$, and
- ▷ $\int_t^{t'} \dot{\boldsymbol{z}}(\tau)d\tau = \boldsymbol{z}_{n+1} - \boldsymbol{z}_n$, $\int_t^{t'} \dot{\boldsymbol{\omega}}(\tau)d\tau = \boldsymbol{\omega}_{n+1} - \boldsymbol{\omega}_n$, $\beta = 2/\gamma$.

Using all the above we get the following implicit discrete time dynamics:

$$\begin{aligned}\boldsymbol{z}_{n+1} &= \boldsymbol{z}_n + \frac{\gamma}{2} \cdot (\boldsymbol{\omega}_{n+1} + \boldsymbol{\omega}_n) \\ \boldsymbol{\omega}_{n+1} &= -V(\boldsymbol{z}_{n+1}) - \frac{1}{2}(V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n))\end{aligned} \tag{OGDA-I}$$

We are going to use again two Lyapunov function to complete our last-iterate argument

$$\mathcal{L}_1(\boldsymbol{z},\boldsymbol{\omega}) = \|\beta\boldsymbol{z}+\boldsymbol{\omega}\|_2^2 + \|\boldsymbol{\omega}\|_2^2 + 2\beta\boldsymbol{z}^\intercal V(\boldsymbol{z}) \tag{OGDA-I-L1}$$

$$\mathcal{L}_2(\boldsymbol{z},\boldsymbol{\omega}) = \|V(\boldsymbol{z})+\boldsymbol{\omega}\|_2^2 + \|V(\boldsymbol{z})\|_2^2. \tag{OGDA-I-L2}$$

We use $(\mathcal{L}_1)_n$ to denote the value $\mathcal{L}_1(\boldsymbol{z}_n,\boldsymbol{\omega}_n)$ and correspondingly for $\mathcal{L}_2$. The discrete time differences of these functions are :

$$\begin{aligned}(\mathcal{L}_1)_{n+1} - (\mathcal{L}_1)_n =& \underbrace{\langle \beta(\boldsymbol{z}_{n+1}+\boldsymbol{z}_n) + (\boldsymbol{\omega}_n+\boldsymbol{\omega}_{n+1}), \beta(\boldsymbol{z}_{n+1}-\boldsymbol{z}_n) + (\boldsymbol{\omega}_{n+1}-\boldsymbol{\omega}_n)\rangle}_{1^\circ} \\ &+ \underbrace{\langle \boldsymbol{\omega}_n+\boldsymbol{\omega}_{n+1}, \boldsymbol{\omega}_{n+1}-\boldsymbol{\omega}_n\rangle}_{2^\circ} + \underbrace{\beta\cdot\boldsymbol{z}_{n+1}^\intercal V(\boldsymbol{z}_{n+1}) - \beta\cdot\boldsymbol{z}_n^\intercal V(\boldsymbol{z}_n)}_{3^\circ} \\ (\mathcal{L}_2)_{n+1} - (\mathcal{L}_2)_n =& \underbrace{\langle (V(\boldsymbol{z}_{n+1})+V(\boldsymbol{z}_n)) + (\boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n), (V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n)) + (\boldsymbol{\omega}_{n+1}-\boldsymbol{\omega}_n)\rangle}_{4^\circ} \\ &+ \underbrace{\langle V(\boldsymbol{z}_{n+1})+V(\boldsymbol{z}_n), V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n)\rangle}_{5^\circ}\end{aligned}$$

Before analyzing each of these terms we observe that we can write the difference

$$\boldsymbol{\omega}_{n+1}-\boldsymbol{\omega}_n = -(\boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n) - (V(\boldsymbol{z}_{n+1})+V(\boldsymbol{z}_n)) - 2\left(V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n)\right) \tag{33}$$

Separately for the above terms we get:

1°) Replacing the difference $\boldsymbol{z}_{n+1}-\boldsymbol{z}_n$ with $\frac{\gamma}{2}\cdot(\boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n)$ and using the fact that $\beta=\frac{2}{\gamma}$ we get that this term 1° is equal to

$$\begin{aligned}&\langle \beta(\boldsymbol{z}_{n+1}+\boldsymbol{z}_n) + (\boldsymbol{\omega}_n+\boldsymbol{\omega}_{n+1}), -(V(\boldsymbol{z}_{n+1})+V(\boldsymbol{z}_n)) - 2\left(V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n)\right)\rangle \\ &\quad= -\beta\cdot\langle \boldsymbol{z}_{n+1}+\boldsymbol{z}_n, V(\boldsymbol{z}_{n+1})+V(\boldsymbol{z}_n)\rangle - 2\cdot\beta\cdot\langle \boldsymbol{z}_{n+1}+\boldsymbol{z}_n, V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n)\rangle \\ &\qquad -\langle \boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n, V(\boldsymbol{z}_{n+1})+V(\boldsymbol{z}_n)\rangle - 2\cdot\langle \boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n, V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n)\rangle.\end{aligned}$$

2°) Using (33) we get that the term 2°

$$\begin{aligned}&\langle \boldsymbol{\omega}_n+\boldsymbol{\omega}_{n+1}, \boldsymbol{\omega}_{n+1}-\boldsymbol{\omega}_n\rangle \\ &\quad= -\|\boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n\|_2^2 - \langle \boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n, V(\boldsymbol{z}_{n+1})+V(\boldsymbol{z}_n)\rangle \\ &\qquad -2\langle \boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n, V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n)\rangle.\end{aligned}$$

3°) For this term we have two different expressions. The first one is

$$\begin{aligned}\beta\cdot\boldsymbol{z}_{n+1}^\intercal V(\boldsymbol{z}_{n+1}) - \beta\cdot\boldsymbol{z}_n^\intercal V(\boldsymbol{z}_n) &= \beta\cdot(\boldsymbol{z}_{n+1}-\boldsymbol{z}_n)^\intercal V(\boldsymbol{z}_{n+1}) - \beta\cdot\boldsymbol{z}_n^\intercal(V(\boldsymbol{z}_n)-V(\boldsymbol{z}_{n+1})) \\ &= \langle \boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n, V(\boldsymbol{z}_{n+1})\rangle + \beta\cdot\boldsymbol{z}_n^\intercal(V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n))\end{aligned}$$

and the other one is

$$\begin{aligned}\beta\cdot\boldsymbol{z}_{n+1}^\intercal V(\boldsymbol{z}_{n+1}) - \beta\cdot\boldsymbol{z}_n^\intercal V(\boldsymbol{z}_n) &= \beta\cdot\boldsymbol{z}_{n+1}^\intercal(V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n)) - \beta\cdot(\boldsymbol{z}_n-\boldsymbol{z}_{n+1})^\intercal V(\boldsymbol{z}_n) \\ &= \beta\cdot\boldsymbol{z}_{n+1}^\intercal(V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n)) + \langle \boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n, V(\boldsymbol{z}_n)\rangle.\end{aligned}$$

If we combine the two above expressions then we have then

$$\begin{aligned}&4\beta\cdot\boldsymbol{z}_{n+1}^\intercal V(\boldsymbol{z}_{n+1}) - 4\beta\cdot\boldsymbol{z}_n^\intercal V(\boldsymbol{z}_n) = \\ &\qquad 2\cdot\beta\cdot\langle \boldsymbol{z}_{n+1}+\boldsymbol{z}_n, V(\boldsymbol{z}_{n+1})-V(\boldsymbol{z}_n)\rangle + 2\cdot\langle \boldsymbol{\omega}_{n+1}+\boldsymbol{\omega}_n, V(\boldsymbol{z}_{n+1})+V(\boldsymbol{z}_n)\rangle.\end{aligned}$$

4°) Using the dynamics for $\boldsymbol{\omega}_{n+1} - \boldsymbol{\omega}_n$ we have that fact this fourth term is equal to

$$\begin{aligned}
&\langle (V(\boldsymbol{z}_{n+1}) + V(\boldsymbol{z}_n)) + (\boldsymbol{\omega}_{n+1} + \boldsymbol{\omega}_n), \\
&\quad - (\boldsymbol{\omega}_{n+1} + \boldsymbol{\omega}_n) - (V(\boldsymbol{z}_{n+1}) + V(\boldsymbol{z}_n)) - (V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n))\rangle \\
&= - \|V(\boldsymbol{z}_{n+1}) + V(\boldsymbol{z}_n) + \boldsymbol{\omega}_{n+1} + \boldsymbol{\omega}_n\|_2^2 - \langle V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n), V(\boldsymbol{z}_{n+1}) + V(\boldsymbol{z}_n)\rangle \\
&\quad - \langle V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n), \boldsymbol{\omega}_{n+1} + \boldsymbol{\omega}_n\rangle.
\end{aligned}$$

5°) For this term we do not make any changes

$$\langle V(\boldsymbol{z}_{n+1}) + V(\boldsymbol{z}_n), V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\rangle.$$

Therefore we have that

$$\begin{aligned}
(\mathcal{L}_1)_{n+1} - (\mathcal{L}_1)_n &= - \beta \cdot \langle \boldsymbol{z}_{n+1} + \boldsymbol{z}_n, V(\boldsymbol{z}_{n+1}) + V(\boldsymbol{z}_n)\rangle - \|\boldsymbol{\omega}_{n+1} + \boldsymbol{\omega}_n\|_2^2 \\
&\quad - 8 \cdot \beta \cdot \langle \boldsymbol{z}_{n+1} - \boldsymbol{z}_n, V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\rangle \qquad (34) \\
(\mathcal{L}_2)_{n+1} - (\mathcal{L}_2)_n &= - \|V(\boldsymbol{z}_{n+1}) + V(\boldsymbol{z}_n) + \boldsymbol{\omega}_{n+1} + \boldsymbol{\omega}_n\|_2^2 \\
&\quad - 2 \cdot \beta \cdot \langle V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n), \boldsymbol{z}_{n+1} - \boldsymbol{z}_n\rangle. \qquad (35)
\end{aligned}$$

The differences of $\mathcal{L}_2$ are clearly negative. For the differences of $\mathcal{L}_1$ the only thing that remains is to show that

$$\langle \boldsymbol{z}_{n+1} + \boldsymbol{z}_n, V(\boldsymbol{z}_{n+1}) + V(\boldsymbol{z}_n)\rangle + \langle \boldsymbol{z}_{n+1} - \boldsymbol{z}_n, V(\boldsymbol{z}_{n+1}) - V(\boldsymbol{z}_n)\rangle \geq 0$$

but the left hand side is equal to

$$2\boldsymbol{z}_{n+1}^\intercal V(\boldsymbol{z}_{n+1}) + 2\boldsymbol{z}_n^\intercal V(\boldsymbol{z}_n)$$

which is clearly positive due to the monotonicity of V and the without loss of generality assumption of the optimality of $\boldsymbol{z} = 0$ which implies $V(0) = 0$.

The proofs of the convergence rates are almost identical to the case of continuous time since the Lyapunov functions are almost the same expect maybe some constants in some of the terms. For this reason we refer to Appendix E.2 for a proof of the rates in Theorem 7.