

---

# Nesterov Meets Optimism: Rate-Optimal Separable Minimax Optimization

---

## Abstract

We propose a new first-order optimization algorithm—Accelerated-Gradient Optimistic Gradient (AG-OG) Descent Ascent—for separable convex-concave minimax optimization. The main idea of our algorithm is to carefully leverage the structure of the minimax problem, performing Nesterov acceleration on the individual component and optimistic gradient on the coupling component. Equipped with proper restarting, we show that AG-OG achieves the optimal convergence rate (up to a constant) for a variety of settings, including bilinearly coupled strongly convex-strongly concave minimax optimization (bi-SC-SC), bilinearly coupled convex-strongly concave minimax optimization (bi-C-SC), and bilinear games. We also extend our algorithm to the stochastic setting and achieve the optimal convergence rate in both bi-SC-SC and bi-C-SC settings. AG-OG is the first single-call algorithm with optimal convergence rates in both deterministic and stochastic settings for bilinearly coupled minimax optimization problems.

## 1 Introduction

Optimization is the workhorse for machine learning (ML) and artificial intelligence. While many ML learning tasks can be cast as a minimization problem, there is an increasing number of ML tasks, such as generative adversarial networks (GANs) (Goodfellow et al., 2020), robust/adversarial training (Bai & Jin, 2020; Madry et al., 2017), Markov games (MGs) (Shapley, 1953), and reinforcement learning (RL) (Sutton & Barto, 2018; Du et al., 2017; Dai et al., 2018), that are instead formulated as a minimax optimization problem in the following form:

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \mathcal{L}(\mathbf{x}, \mathbf{y}). \quad (1.1)$$

---

\*Equal contribution <sup>1</sup>. Correspondence to: Firstname1 Lastname1 <first1.last1@xxx.edu>, Firstname2 Lastname2 <first2.last2@www.uk>.

When  $\mathcal{L}(\mathbf{x}, \mathbf{y}) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  is a smooth function that is convex in  $\mathbf{x}$  and concave in  $\mathbf{y}$ , we refer to this problem as a *convex-concave minimax problem* (a.k.a., convex-concave saddle point problem). In this work, we focus on designing fast or even optimal deterministic and stochastic first-order algorithms for solving convex-concave minimax problems of the form (1.1).

Unlike in the convex minimization setting, where gradient descent is the method of choice, the gradient descent-ascent method can exhibit divergence on convex-concave objectives. Indeed, examples show the divergence of gradient descent ascent (GDA) on bilinear objectives (Liang & Stokes, 2019; Gidel et al., 2018). This has led to the development of extrapolation-based methods, including the extragradient (EG) method (Korpelevich, 1976) and the optimistic gradient descent ascent (OGDA) method (Popov, 1980), both of which can be shown to converge in the convex-concave setting. While the EG algorithm needs to call the gradient oracle twice at each iteration, the OGDA algorithm only needs a single call to the gradient oracle (Gidel et al., 2018; Hsieh et al., 2019) and therefore has a practical advantage when the gradient evaluation is expensive. We build on this line of research, aiming to attain improved, and even optimal, convergence rates via algorithms that retain the spirit of simplicity of OGDA.

We focus on a specific instance of the general minimax optimization problem, namely the separable minimax optimization problem, which is formulated as follows

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \mathcal{L}(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + I(\mathbf{x}, \mathbf{y}) - g(\mathbf{y}), \quad (1.2)$$

where  $f$  is  $\mu_f$ -strongly convex and  $L_f$ -smooth,  $g$  is  $\mu_g$ -strongly convex and  $L_g$ -smooth, and  $I$  is convex-concave and  $L_H$ -smooth. We refer to  $f(\mathbf{x}) - g(\mathbf{y})$  as the *individual component*, and  $I(\mathbf{x}, \mathbf{y})$  as the *coupling component* of Problem (1.2). We characterize the smoothness of  $I$  by blockwise smoothness parameters  $I_{xx}, I_{xy}, I_{yy}$  where  $\|\nabla_{xx}^2 I\|_{\text{op}} \leq I_{xx}$ ,  $\|\nabla_{xy}^2 I\|_{\text{op}} \leq I_{xy}$ , and  $\|\nabla_{yy}^2 I\|_{\text{op}} \leq I_{yy}$ . It is straightforward to observe that  $L_H = I_{xx} \vee I_{yy} + I_{xy}$ . Throughout this paper, we focus on the unconstrained problem where  $\mathcal{X} = \mathbb{R}^n$  and  $\mathcal{Y} = \mathbb{R}^m$  unless otherwise specified in certain applications.

A notable special case of the separable minimax Problem (1.2) is the so-called *bilinearly coupled strongly convex-strongly concave minimax problem* (bi-SC-SC), which has

the following form:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{L}(x, y) \equiv f(x) + x^\top \mathbf{B}y - g(y). \quad (1.3)$$

Here  $x^\top \mathbf{B}y$  is the bilinear coupling term and is  $L_H$ -smooth with  $L_H = \sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}$ , which coincides with the operator norm of matrix  $\mathbf{B}$ .

For the general minimax optimization problem (1.1), standard algorithms such as mirror-prox (Nemirovski, 2004), EG and OGDA—when operating on the entire objective—can be shown to exhibit a complexity upper bound of  $\frac{\bar{L}}{\bar{\mu}} \log(\frac{1}{\epsilon})$  for finding an  $\epsilon$ -accurate solution (Gidel et al., 2018; Mokhtari et al., 2020a), where  $\bar{L} \equiv L_f \vee L_g \vee L_H$  and  $\bar{\mu} \equiv \mu_f \wedge \mu_g$ . Such a complexity is *optimal* when  $L_f = L_g = L_H$  and  $\mu_f = \mu_g$ , since the lower-bound complexity is  $\Omega(\frac{\bar{L}}{\bar{\mu}} \log(\frac{1}{\epsilon}))$  (Nemirovskij & Yudin (1983); Azizian et al. (2020)). However, in the general case where the strong convexity and smoothness parameters are significantly different in  $x$  and  $y$ , fine-grained convergence rates that depend on the individual strong convexity  $\mu_f, \mu_g$  and smoothness parameters  $L_f, L_g, L_H$  are more desirable. In fact, Zhang et al. (2021a) have proved the following iteration complexity lower bound for solving (1.3) by any first-order algorithm under the linear span assumption:

$$\tilde{\Omega}\left(\sqrt{\frac{L_f + I_{xx}}{\mu_f} + \frac{L_g + I_{yy}}{\mu_g} + \frac{I_{xy}^2}{\mu_f \mu_g}}\right). \quad (1.4)$$

With the goal of attaining this lower bound, several efforts have been made in the setting of bi-SC-SC (1.3) or separable SC-SC (1.2). Two notable methods are LPD (Thekumparampil et al., 2022) and PD-EG (Jin et al., 2022), which utilize techniques from primal-dual lifting and convex conjugate decomposition. Another approach is the APDG algorithm developed by Kovalev et al. (2021), which is based on adding an extrapolation step to the forward-backward algorithm. The work of Du et al. (2022) is also closely related to our work, in the sense that it uses iterate averaging and employs scaling reduction with scheduled restarting. However, these algorithms are either limited to the bi-SC-SC setting (Kovalev et al., 2021; Thekumparampil et al., 2022; Du et al., 2022), or are not single-call algorithms (Kovalev et al., 2021; Jin et al., 2022; Du et al., 2022).<sup>1</sup> In addition, only Kovalev et al. (2021) and Du et al. (2022) can be extended to the stochastic setting (Metelev et al., 2022), while the extension of Thekumparampil et al. (2022); Jin et al. (2022) to the stochastic setting remains elusive.

In this paper, we design near-optimal single-call algorithms for both deterministic and stochastic separable minimax

<sup>1</sup>By single call, we mean the algorithm only needs to call the (stochastic) gradient oracle of the coupling component once in each iteration of the algorithm. This is in accordance with the concept of single-call variants of extragradient in Hsieh et al. (2019). Previous work calls  $\nabla I(x, y)$  at least twice per iteration.

problems (1.2). We focus on accelerating OGDA because of its simplicity and because it enjoys the single-call property. We show that it achieves a *fine-grained, accelerated* convergence rate with a sharp dependency on the individual Lipschitz constants. To the best of our knowledge, this is the first presentation of a single-call algorithm that matches the best-known result for the separable minimax problem (1.2) and the lower bounds under a bi-SC-SC setting (1.3), bilinearly coupled convex-strongly concave (bi-C-SC) setting (i.e.,  $f$  is convex but not strongly convex in (1.3)), and the bilinear game setting (i.e., setting  $f = g = 0$  in (1.3)).

## 1.1 Contributions

We highlight our contributions as follows.

- (i) We present a novel algorithm that blends acceleration dynamics based on the single-call OGDA algorithm for the coupling component and Nesterov’s acceleration for the individual component. We refer to this new algorithm as the *AcceleratedGradient-OptimisticGradient* (AG-OG) Descent Ascent algorithm.
- (ii) Using a scheduled restarting, we derive an *AcceleratedGradient OptimisticGradient with restarting* (AGOG with restarting) algorithm that achieves a sharp convergence rate in a variety of settings. For the setting of bilinearly coupled SC-SC in (1.3), our algorithm achieves a complexity of  $O\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_f \mu_g}}\right) \log\left(\frac{1}{\epsilon}\right)$  [Corollary 3.4], which matches the lower bound established by Zhang et al. (2021a). For bi-C-SC, we prove a  $O\left(\sqrt{\frac{L_f}{\epsilon} \vee \frac{L_g}{\mu_g}} + \frac{\|\mathbf{B}\|}{\sqrt{\epsilon \mu_g}}\right) \log\left(\frac{1}{\epsilon}\right)$  complexity [Theorem 3.5], which matches that of Thekumparampil et al. (2022) and is also optimal.
- (iii) In the stochastic setting where the algorithm can only query a stochastic gradient oracle with bounded noise, we propose a stochastic extension of AGOG with restarting and establish a sharp convergence rate. For both bi-SC-SC and bi-C-SC settings, the convergence rate of our algorithm is near-optimal in the sense that its bias error matches the respective deterministic lower bound and its variance error matches the statistical minimax rate, i.e.,  $\frac{\sigma_f^2}{\mu_f^2 \epsilon^2}$  [Corollary 4.3].
- (iv) In the special case of the bilinear game (when  $f = g = 0$  in (1.3)), our algorithm has a complexity of  $\Omega\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})}}\right) \log\left(\frac{1}{\epsilon}\right)$  [Theorem 3.6], which matches the lower bound established by Ibrahim et al. (2020). Note that prior work (Kovalev et al., 2021; Thekumparampil et al., 2022; Jin et al., 2022) cannot achieve the optimal rate when applied to bilinear games, which is an unique advantage of our algorithm.

A summary of the iteration complexity comparisons with the state-of-the-art methods can be found in Table 1.

## Rate-Optimal Separable Minimax Optimization

Setting Method	SC-SC	bi-SC-SC	Bilinear	bi-C-SC	Stochastic Rate	Single Call
OGDA (Mokhtari et al., 2020b)	$\tilde{O}\left(\frac{L'_f \vee L'_g \vee I_{xy}}{\mu_f \wedge \mu_g}\right)$	$\tilde{O}\left(\frac{L_f \vee L_g \vee \ \mathbf{B}\ }{\mu_f \wedge \mu_g}\right)$	$\tilde{O}\left(\frac{\ \mathbf{B}\ ^2}{\lambda_{\min}}\right)$	$\mathcal{O}\left(\frac{L_f \vee L_g \vee \ \mathbf{B}\ }{\epsilon}\right)$	✓	✓
Proximal Best Response (Wang & Li, 2020)	$\tilde{O}\left(\sqrt{\frac{L'_f}{\mu_f}} \vee \sqrt{\frac{L'_g}{\mu_g}} + \sqrt{\frac{I_{xy}(L'_f \vee L'_g \vee I_{xy})}{\mu_f \mu_g}}\right)$	$\tilde{O}\left(\sqrt{\frac{L_f}{\mu_f}} \vee \sqrt{\frac{L_g}{\mu_g}} + \sqrt{\frac{\ \mathbf{B}\ (L_f \vee L_g \vee \ \mathbf{B}\ )}{\mu_f \mu_g}}\right)$	—	—	✗	✗
DIPPA (Xie et al., 2021)	—	$\tilde{O}\left(\left(\frac{L_f L_g}{\mu_f} \vee \frac{L_g}{\mu_g}\right)^{\frac{1}{2}} + \frac{\ \mathbf{B}\ }{\sqrt{\mu_f \mu_g}}\right)$	—	—	✗	✗
LPD (Thekumparampil et al., 2022)	—	$\tilde{O}\left(\sqrt{\frac{L_f}{\mu_f}} \vee \sqrt{\frac{L_g}{\mu_g}} + \frac{\ \mathbf{B}\ }{\sqrt{\mu_f \mu_g}}\right)$	$\tilde{O}\left(\frac{\ \mathbf{B}\ ^2}{\lambda_{\min}}\right)$	$\tilde{O}\left(\sqrt{\frac{L_f}{\epsilon}} \vee \sqrt{\frac{L_g}{\epsilon}} + \frac{\ \mathbf{B}\ }{\sqrt{\epsilon \mu_g}}\right)$	✗	✓
APDG (Kovalev et al., 2021) (Metelev et al., 2022)	—	$\tilde{O}\left(\sqrt{\frac{L_f}{\mu_f}} \vee \sqrt{\frac{L_g}{\mu_g}} + \frac{\ \mathbf{B}\ }{\sqrt{\mu_f \mu_g}}\right)$	$\tilde{O}\left(\frac{\ \mathbf{B}\ ^2}{\lambda_{\min}}\right)$	$\tilde{O}\left(\sqrt{\frac{L_f L_g}{\lambda_{\min}}} \vee \frac{\ \mathbf{B}\ }{\sqrt{\lambda_{\min}}} \sqrt{\frac{L_g}{\mu_g}} \vee \frac{\ \mathbf{B}\ ^2}{\lambda_{\min}}\right)$	✓	✗
PD-EG (Jin et al., 2022)	$\tilde{O}\left(\sqrt{\frac{L'_f}{\mu_f}} \vee \sqrt{\frac{L'_g}{\mu_g}} + \frac{I_{xx}}{\mu_f} \vee \sqrt{\frac{I_{xy}}{\mu_f \mu_g}} \vee \sqrt{\frac{I_{yy}}{\mu_g}}\right)$	$\tilde{O}\left(\sqrt{\frac{L_f}{\mu_f}} \vee \sqrt{\frac{L_g}{\mu_g}} + \frac{\ \mathbf{B}\ }{\sqrt{\mu_f \mu_g}}\right)$	$\tilde{O}\left(\frac{\ \mathbf{B}\ ^2}{\lambda_{\min}}\right)$	—	✗	✗
EG+Momentum (Azizian et al., 2020)	—	—	$\tilde{O}\left(\frac{\ \mathbf{B}\ }{\sqrt{\lambda_{\min}}}\right)$	—	✗	✗
SEG with Restarting (Li et al., 2022)	—	—	$\tilde{O}\left(\frac{\ \mathbf{B}\ }{\sqrt{\lambda_{\min}}}\right)$	—	✓	✗
AG-EG with Restarting (Du et al., 2022)	—	$\tilde{O}\left(\sqrt{\frac{L_f}{\mu_f}} \vee \sqrt{\frac{L_g}{\mu_g}} + \frac{\ \mathbf{B}\ }{\sqrt{\mu_f \mu_g}}\right)$	$\tilde{O}\left(\frac{\ \mathbf{B}\ }{\sqrt{\lambda_{\min}}}\right)$	—	✓	✗
AG-OG with Restarting (this work)	$\tilde{O}\left(\sqrt{\frac{L'_f}{\mu_f}} \vee \sqrt{\frac{L'_g}{\mu_g}} + \frac{I_{xx}}{\mu_f} \vee \sqrt{\frac{I_{xy}}{\mu_f \mu_g}} \vee \sqrt{\frac{I_{yy}}{\mu_g}}\right)$	$\tilde{O}\left(\sqrt{\frac{L_f}{\mu_f}} \vee \sqrt{\frac{L_g}{\mu_g}} + \frac{\ \mathbf{B}\ }{\sqrt{\mu_f \mu_g}}\right)$	$\tilde{O}\left(\frac{\ \mathbf{B}\ }{\sqrt{\lambda_{\min}}}\right)$	$\tilde{O}\left(\sqrt{\frac{L_f}{\epsilon}} \vee \sqrt{\frac{L_g}{\epsilon}} + \frac{\ \mathbf{B}\ }{\sqrt{\epsilon \mu_g}}\right)$	✓	✓
Lower Bound (Zhang et al., 2021a) (Ibrahim et al., 2020)	$\tilde{\Omega}\left(\sqrt{\frac{L'_f}{\mu_f}} \vee \sqrt{\frac{L'_g}{\mu_g}} + \frac{I_{xy}}{\sqrt{\mu_f \mu_g}}\right)$	$\tilde{\Omega}\left(\sqrt{\frac{L_f}{\mu_f}} \vee \sqrt{\frac{L_g}{\mu_g}} + \frac{\ \mathbf{B}\ }{\sqrt{\mu_f \mu_g}}\right)$	$\tilde{\Omega}\left(\frac{\ \mathbf{B}\ }{\sqrt{\lambda_{\min}}}\right)$	—	—	—

Table 1. We present a comparison of the first-order gradient complexities of our proposed algorithm with selected prevailing algorithms for solving bilinearly-coupled minimax problems. The comparison includes several cases such as general SC-SC, bilinear games, bi-SC-SC (bilinearly-coupled SC-SC), and the bi-C-SC cases. We denote  $\lambda_{\min} \equiv \lambda_{\min}(\mathbf{B}^\top \mathbf{B})$ ,  $L'_f \equiv L_f + I_{xx}$  and  $L'_g \equiv L_g + I_{yy}$ . We focus on comparing the gradient complexities of deterministic algorithms, and include a column to indicate whether the stochastic case has been discussed. The row in blue background is the convergence result presented in this paper. The “—” indicates that the complexity does not apply to the given case.

### 1.2 More Related Work

**Deterministic Case.** Much attention has been paid to obtaining linear convergence rates for gradient-based methods applied to games in the context of strongly monotone operators (which is implied by strong convex-concavity) (Mokhtari et al., 2020a) and several recent works (Yang et al., 2020; Zhang et al., 2021b; Cohen et al., 2020; Wang & Li, 2020; Xie et al., 2021) have bridged the gap with the lower bound provided for *unbalanced* strongly-convex-strongly-concave objective. There has been a series of papers along this direction (Mokhtari et al., 2020a; Cohen et al., 2020; Lin et al., 2020a; Wang & Li, 2020; Xie et al., 2021), and only very recently have optimal results that reach the lower bound been presented (Kovalev et al., 2021; Thekumparampil et al., 2022; Jin et al., 2022). This work presented improved methods leveraging convex duality. Among these works, only Jin et al. (2022) considers non-bilinear coupling terms, and only Thekumparampil et al. (2022) considers single gradient calls. Note that Jin et al. (2022) consider a finite-sum case, which differs from our setting of a general expectation. Kovalev et al. (2021); Thekumparampil et al. (2022) focus solely on the deterministic setting, and Metelev et al. (2022) present a stochastic version of APDG algorithm (Kovalev et al., 2021) and its extension to a decentralized setting, which is comparable and concurrent with the work of Du et al. (2022).

**Stochastic Case.** There exists a rich literature on stochastic variational inequalities with application to solving stochastic minimax problems (Juditsky et al., 2011; Hsieh et al., 2019; Chavdarova et al., 2019; Alacaoglu & Malitsky, 2022; Zhao, 2022; Beznosikov et al., 2022). However, only a few works have proposed fine-grained bounds suited to the (bi-)SC-SC setting. To the best of our knowledge, most fine-grained bounds have been proposed in the finite-sum setting (Palaniappan & Bach, 2016; Jin et al., 2022) or in the proximal-friendly case (Zhang et al., 2021c). Two closely related works are Li et al. (2022), who provide a convergence rate for extragradient in the purely bilinear setting and Du et al. (2022), who study an accelerated version of extragradient, dubbed as AcceleratedGradient-ExtraGradient (AG-EG) in the bi-SC-SC setting. Our work is in the same vein as Du et al. (2022) but instead employs the optimistic gradient instead of extragradient to handle the bilinear coupling component. Optimistic-gradient-based methods have been considered extensively in the literature due to their need for fewer gradient oracle calls per iteration than standard extragradient and their applicability to the online learning setting (Golowich et al., 2020). Note that, in general, EG and OG methods share some similarities in their analyses, but there are also significant differences (Golowich et al., 2020, §3.1), (Gorbunov et al., 2022, §2). Specifically in our case, using a *single-call algorithm* that reuses previously calculated gradients alters a key recursion (Eq. (D.6)). Al-

though the main part of the proof follows the standard path of estimating Nesterov’s acceleration terms first, an additional squared error norm involving the previous iterates is present, intrinsically implying an additional iterative rule (Eq. (D.7)) in place of the original iterative rule that is essential for proving boundedness of the iterates. In addition, due to the accumulated error across iterates, the maximum stepsize allowed in single-call algorithms is forced to be smaller. We believe that this is not an artifact of our analysis but is a general feature of OG methods.<sup>2</sup>

**Organization.** The rest of this work is organized as follows. §2 introduces the basic settings and assumptions necessary for our algorithm and theoretical analysis. Our proposed AcceleratedGradient-OptimisticGradient (AG-OG) Descent Ascent algorithm is formally introduced in §3.3 and further generalized to Stochastic AcceleratedGradient-OptimisticGradient (S-AG-OG) Descent Ascent in §4. We present our conclusions in §5. Due to space limitations, we defer all proof details along with results of numerical experiments to the supplementary materials.

**Notation.** For two sequences of positive scalars  $\{a_n\}$  and  $\{b_n\}$ , we denote  $a_n = \Omega(b_n)$  (resp.  $a_n = \mathcal{O}(b_n)$ ) if  $a_n \geq Cb_n$  (resp.  $a_n \leq Cb_n$ ) for all  $n$ , and also  $a_n = \Theta(b_n)$  if both  $\Omega(b_n)$  and  $a_n = \mathcal{O}(b_n)$  hold, for some absolute constant  $C > 0$ , and  $\tilde{\mathcal{O}}$  or  $\tilde{\Omega}$  is adopted in turn when  $C$  contains a polylogarithmic factor in problem-dependent parameters. Let  $\lambda_{\max}(\mathbf{A})$  and  $\lambda_{\min}(\mathbf{A})$  denote the maximal and minimal eigenvalues of a real symmetric matrix  $\mathbf{A}$ , and  $\|\mathbf{A}\|$  the operator norm  $\sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}$ . Let vector  $\mathbf{z} = [\mathbf{x}; \mathbf{y}] \in \mathbb{R}^{n+m}$  denote the concatenation of  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{y} \in \mathbb{R}^m$ . For natural number  $K$  let  $[K]$  denote the set  $\{1, \dots, K\}$ . Throughout the paper we also use the standard notation  $\|\cdot\|$  to denote the  $\ell_2$ -norm. We will explain other notation at its first appearance.

## 2 Preliminaries

In minimax optimization the goal is to find the Nash equilibrium (or minimax point) of problem (1.1), defined as a pair  $[\mathbf{x}^*; \mathbf{y}^*] \in \mathcal{X} \times \mathcal{Y}$  satisfying:

$$\mathcal{L}(\mathbf{x}^*, \mathbf{y}) \leq \mathcal{L}(\mathbf{x}^*, \mathbf{y}^*) \leq \mathcal{L}(\mathbf{x}, \mathbf{y}^*).$$

In order to analyze first-order gradient methods for this problem, we assume access to the gradients of the objective  $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{y})$  and  $\nabla_{\mathbf{y}} \mathcal{L}(\mathbf{x}, \mathbf{y})$ . Finding the minimax point of the original convex-concave optimization problem (1.1) and (1.2) reduces to finding the point where the gradients vanish. Accordingly, we use  $W$  to denote the gradient vector field and  $\mathbf{z} = [\mathbf{x}; \mathbf{y}] \in \mathbb{R}^{n+m}$ , and analogously for  $\mathbf{z}^*$ :

$$W(\mathbf{z}) := \begin{pmatrix} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} \mathcal{L}(\mathbf{x}, \mathbf{y}) \end{pmatrix} = \begin{pmatrix} \nabla f(\mathbf{x}) + \nabla_{\mathbf{x}} I(\mathbf{x}, \mathbf{y}) \\ -\nabla_{\mathbf{y}} I(\mathbf{x}, \mathbf{y}) + \nabla g(\mathbf{y}) \end{pmatrix}. \quad (2.1)$$

<sup>2</sup>Limited by space, we refer readers to §D.1 and §D.4 for technical details.

Based on this formulation, our goal is to find the stationary point of the vector field corresponding to the monotone operator  $W(\mathbf{z})$ , namely a  $\mathbf{z}^* = [\mathbf{x}^*; \mathbf{y}^*] \in \mathbb{R}^{n+m}$  satisfying (in the unconstrained case)  $W(\mathbf{z}^*) = 0$ , which is referred to as the *variational inequality (VI) formulation* of minimax optimization (Gidel et al., 2018). The compact representation of the convex-concave minimax problem as a VI allows us to simplify the notation.

In the vector field (2.1), there are individual components that point along the direction optimizing  $f, g$  individually, and a coupling component which corresponds to the gradient vector field of a separable minimax problem. For the individual component, we let  $F(\mathbf{z}) := f(\mathbf{x}) + g(\mathbf{y})$  and correspondingly  $\nabla F(\mathbf{z}) = [\nabla f(\mathbf{x}); \nabla g(\mathbf{y})]$ . For the coupling component, we define the operator  $H(\mathbf{z}) = [\nabla_{\mathbf{x}} I(\mathbf{x}, \mathbf{y}); -\nabla_{\mathbf{y}} I(\mathbf{x}, \mathbf{y})]$ . Note that the representation allows us to write  $W(\mathbf{z})$  as the summation of the two vector fields:  $W(\mathbf{z}) = \nabla F(\mathbf{z}) + H(\mathbf{z})$ .

We introduce our main assumptions as follows:

**Assumption 2.1** (Convexity and Smoothness). We assume that  $f(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}$  is  $\mu_f$ -strongly convex and  $L_f$ -smooth,  $g(\cdot) : \mathbb{R}^m \rightarrow \mathbb{R}$  is  $\mu_g$ -strongly convex and  $L_g$ -smooth, and  $I(\mathbf{x}, \mathbf{y})$  is convex-concave with blockwise smoothness parameters  $L_H = I_{xx} \vee I_{yy} + I_{xy}$ .

This implies that  $F(\mathbf{z})$  is  $(L_f \vee L_g)$ -smooth and  $(\mu_f \wedge \mu_g)$ -strongly convex. In addition  $H(\cdot)$  is monotone, yielding the property that for all  $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^{n+m}$ :

$$\langle H(\mathbf{z}) - H(\mathbf{z}'), \mathbf{z} - \mathbf{z}' \rangle \geq 0. \quad (2.2)$$

The above assumption adds convexity and smoothness constraints to the individual components  $f(\mathbf{x})$  and  $g(\mathbf{y})$ . In addition, for the coupling component  $\mathbf{x}^\top \mathbf{B} \mathbf{y}$  in the separable minimax problem (1.2), without loss of generality, we assume that  $\mathbf{B} \in \mathbb{R}^{n \times m}$ ,  $n \geq m > 0$  is a tall matrix. Note that as  $\mathbf{x}$  and  $\mathbf{y}$  are exchangeable, tall matrices cover all circumstances.

In the stochastic setting, we assume access to an unbiased stochastic oracle  $\tilde{H}(\mathbf{z}; \zeta)$  of  $H(\mathbf{z})$  and an unbiased stochastic oracle  $\tilde{\nabla F}(\mathbf{z}; \xi)$  of  $\nabla F(\mathbf{z})$ . Furthermore, we consider the case where the variances of such stochastic oracles are bounded:

**Assumption 2.2** (Bounded Variance). We assume that the stochastic gradients admit bounded second moments  $\sigma_H^2, \sigma_F^2 \geq 0$ :

$$\begin{aligned} \mathbb{E}_{\xi} \left[ \|\tilde{H}(\mathbf{z}; \zeta) - H(\mathbf{z})\|^2 \right] &\leq \sigma_H^2, \\ \mathbb{E}_{\zeta} \left[ \|\tilde{\nabla F}(\mathbf{z}; \xi) - \nabla F(\mathbf{z})\|^2 \right] &\leq \sigma_F^2. \end{aligned}$$

For ease of exposition, we introduce the overall variance  $\sigma^2 = 3\sqrt{2}\sigma_H^2 + 2\sigma_F^2$ . Note that the bounded noise assumption is common in the stochastic optimization litera-



ture.<sup>3</sup> Under the above assumptions, our goal is to find an  $\epsilon$ -optimal minimax point, a notion defined as follows.

**Definition 2.3** ( $\epsilon$ -Optimal Minimax Point).  $[x; y] \in \mathcal{X} \times \mathcal{Y}$  is called an  $\epsilon$ -optimal minimax point of a convex-concave function  $\mathcal{L}(x, y)$  if  $\|x - x^*\|^2 + \|y - y^*\|^2 \leq \epsilon^2$ .

It is obvious that when the accuracy  $\epsilon = 0$ ,  $[x; y]$  is an (exact) optimal minimax point of  $\mathcal{L}(x, y)$ .

### 3 Accelerated Gradient Optimistic Gradient Descent Ascent

In this section, we discuss key elements of our algorithm design—consisting of *Optimistic Gradient Descent-Ascent* (OGDA) and *Nesterov’s acceleration method*—that together solve the separable minimax problem. Such an approach allows us to demonstrate the main properties of our approach that will eventually guide our analysis in the discrete-time case. In §3.1 and §3.2 we review OGDA and Nesterov’s acceleration. In §3.3 we present our approach to accelerating OGDA for bilinear minimax problems, yielding the Accelerated Gradient-Optimistic Gradient (AG-OG) algorithm, and we prove its convergence. Finally in §3.4 we show that proper restarting on top of the AG-OG algorithm achieves a sharp convergence rate that matches the lower bound of Zhang et al. (2021a).

#### 3.1 Optimistic Gradient Descent Ascent

The Optimistic Gradient Descent Ascent (OGDA) algorithm has received considerable attention in the recent literature, especially for the problem of training Generative Adversarial Networks (GANs) (Goodfellow et al., 2020). In the general variational inequality setting, the iteration of OGDA takes the following form (Popov, 1980):

$$z_{k+\frac{1}{2}} = z_k - \eta W(z_{k-\frac{1}{2}}), \quad z_{k+1} = z_k - \eta W(z_{k+\frac{1}{2}}). \quad (3.1)$$

Note that at step  $k$ , the scheme performs a gradient descent-ascent step at the *extrapolated point*  $z_{k+\frac{1}{2}}$ . Equivalently, with simple algebraic modification (3.1) can be written in a standard form (Gidel et al., 2018):

$$z_{k+\frac{1}{2}} = z_{k-\frac{1}{2}} - 2\eta W(z_{k-\frac{1}{2}}) + \eta W(z_{k-\frac{3}{2}}). \quad (3.2)$$

Treating  $W(z_{k-\frac{1}{2}}) - W(z_{k-\frac{3}{2}})$  as a prediction of the future,  $W(z_{k+\frac{1}{2}}) - W(z_{k-\frac{1}{2}})$ , this update rule can be viewed as an approximation of the implicit *proximal point (PP) method*:

$$z_{k+\frac{1}{2}} = z_{k-\frac{1}{2}} - \eta W(z_{k+\frac{1}{2}}).$$

Another popular tractable approximation of the PP method is the extragradient (EG) method (Korpelevich, 1976): Although similar conceptually to OGDA (3.1), EG requires

<sup>3</sup>We leave the generalization to models of unbounded noise to future work.

two gradient computations per iteration, which doubles the number of gradient computations of OGDA. Both OGDA and EG dynamics (3.1) alleviate cyclic behavior by extrapolation from the past and exhibit a complexity of  $(L/\mu) \log(1/\epsilon)$  (Gidel et al., 2018; Mokhtari et al., 2020a) in general setting (1.1) with  $L$ -smooth,  $\mu$ -strongly-convex- $\mu$ -strongly-concave objectives.<sup>4</sup>

#### 3.2 Nesterov’s Acceleration Scheme

Turning to the minimization problem, while vanilla gradient descent enjoys an iteration complexity of  $\kappa \log(1/\epsilon)$  on  $L$ -smooth,  $\mu$ -strongly convex problems, with  $\kappa = L/\mu$  being the condition number, Nesterov’s method (Nesterov, 1983), when equipped with proper restarting, achieves an improved iteration complexity of  $\sqrt{\kappa} \log(1/\epsilon)$ . We adopt the following version of the Nesterov acceleration, known as the “second scheme” (Tseng, 2008; Lin et al., 2020b):

$$\begin{cases} z_k^{\text{md}} &= \frac{k}{k+2} z_k^{\text{ag}} + \frac{2}{k+2} z_k, \end{cases} \quad (3.3a)$$

$$\begin{cases} z_{k+1} &= z_k - \eta_k \nabla F(z_k^{\text{md}}), \end{cases} \quad (3.3b)$$

$$\begin{cases} z_{k+1}^{\text{ag}} &= \frac{k}{k+2} z_k^{\text{ag}} + \frac{2}{k+2} z_{k+1}. \end{cases} \quad (3.3c)$$

Subtracting (3.3a) from (3.3c) and combining the resulting equation with (3.3b), we conclude

$$\begin{aligned} z_{k+1}^{\text{ag}} - z_k^{\text{md}} &= \frac{2}{k+2} (z_{k+1} - z_k) = -\eta_k \nabla F(z_k^{\text{md}}) \\ \Rightarrow z_{k+1}^{\text{ag}} &= z_k^{\text{md}} - \eta_k \nabla F(z_k^{\text{md}}). \end{aligned} \quad (3.4)$$

Moreover, shifting the index forward by one in (3.3a) and combining it with (3.3c) to cancel the  $z_{k+1}$  term, we obtain

$$\frac{k+2}{k+3} z_{k+1}^{\text{ag}} - z_{k+1}^{\text{md}} = \frac{k}{k+3} z_k^{\text{ag}} - \frac{k+1}{k+3} z_{k+1}^{\text{ag}} \quad (3.5)$$

$$\Rightarrow z_{k+1}^{\text{md}} = z_{k+1}^{\text{ag}} + \frac{k}{k+3} (z_{k+1}^{\text{ag}} - z_k^{\text{ag}}). \quad (3.6)$$

Thus, by a simple notational transformation, (3.4) plus (3.6) (and hence the original update rule (3.3)) is exactly equivalent to the original updates of Nesterov’s acceleration scheme (Nesterov, 1983). Here,  $z_k^{\text{ag}}$  denotes a  $\frac{2}{k}$ -weighted-averaged iteration. In other words, compared with vanilla gradient descent,  $z_{k+1} = z_k - \eta_k \nabla F(z_k)$ , Nesterov’s acceleration conducts a step at the negated gradient direction evaluated at a *predictive iterate* of the weighted-averaged iterate of the sequence. This enables a larger choice of stepsize, reflecting the enhanced stability. An analogous interpretation has been discussed in work on a heavy-ball-based acceleration method (Sebbouh et al., 2021, §1.3).

#### 3.3 Accelerating OGDA on Separable Minimax Problems

In this subsection and §3.4, we show that an organic combination of the two algorithms in §3.1 and §3.2 achieves

<sup>4</sup>In fact an analogous result holds true for general smooth, strongly monotone variational inequalities (Mokhtari et al., 2020a).

**Algorithm 1** AcceleratedGradient-OptimisticGradient (AG-OG)( $z_0^{\text{ag}}, z_0, z_{-1/2}, K$ )

---

```

1: for  $k = 0, 1, \dots, K - 1$  do
2:    $z_k^{\text{md}} = (1 - \alpha_k)z_k^{\text{ag}} + \alpha_k z_k$ 
3:    $z_{k+\frac{1}{2}} = z_k - \eta_k \left( H(z_{k-\frac{1}{2}}) + \nabla F(z_k^{\text{md}}) \right)$ 
4:    $z_{k+1}^{\text{ag}} = (1 - \alpha_k)z_k^{\text{ag}} + \alpha_k z_{k+\frac{1}{2}}$ 
5:    $z_{k+1} = z_k - \eta_k \left( H(z_{k+\frac{1}{2}}) + \nabla F(z_k^{\text{md}}) \right)$ 
6: end for
7: Output:  $z_K^{\text{ag}}$ 
    
```

---

improved convergence rates and when equipped with scheduled restarting, obtains a sharp iteration complexity that matches Jin et al. (2022) while only requiring a single gradient call per iterate. Our algorithm is shown in Algorithm 1. In Line 2 and 4 the update rules of the evaluated point and the extrapolated point of  $f$  follow that in (3.3), while in Line 3 and 5 the updates follow the OGD dynamics (3.1) with each step modified by (3.3b). Algorithm 1 can be seen as a synthesis of OGD and Nesterov’s acceleration, as it reduces to OGD when  $F = 0$  and to Nesterov’s accelerated gradient when  $H = 0$ .

For theoretical analysis, we first state an elementary lemma that establishes the non-expansive property of  $z_k$ . The proof is presented in §G.3.

**Lemma 3.1** (Bounded Iterates). *Under Assumptions 2.1, we set the parameters as  $L = L_f \vee L_g$ ,  $L_H = I_{xx} \vee I_{yy} + I_{xy}$ ,  $\eta_k = \frac{k+2}{2L+\sqrt{3+\sqrt{3}L_H(k+2)}}$  and  $\alpha_k = \frac{2}{k+2}$  in Algorithm 1, at any iterate  $k < K$ ,  $z_k$  stays within the region defined by the initialization  $z_0$ :*

$$\|z_k - z^*\| \leq \|z_0 - z^*\|,$$

where we recall that  $z^*$  denotes the unique solution to Problem (1.2).

**Remark 3.2.** The result in Lemma 3.1 is significant because it represents an improvement over the state-of-the-art (SOTA) result achieved by Lemma 2(b) in Mokhtari et al. (2020b) (considering the case where  $f = 0, g = 0$ ), which only achieves a coefficient of  $\sqrt{2}$  in the Euclidean metric (i.e.,  $\|z_k - z^*\| \leq \sqrt{2}\|z_0 - z^*\|$ ). From a past-gradient perspective, previous analysis has focused on the extra-step ( $z_{k+\frac{1}{2}}$  in our formulation), while our analysis is able to perform a non-expansiveness analysis on the integer steps ( $z_k$ ). This ensures that all iterates do not expand relative to the initialization.

Lemma 3.1 establishes the following last-iterate boundedness: the  $z_k$  iteration is bounded within the ball centered at  $z^*$  and radius  $\|z_0 - z^*\|$  and is essential in proving convergence of iteration  $z_k^{\text{ag}}$ , where handling additional recursions brought by a gradient evaluated from a previous step is the main technical difficulty. With the parameter choice in Lemma 3.1, Line 4 can also be seen as an average step

**Algorithm 2** AcceleratedGradient-OptimisticGradient with restarting (AGOG with restarting)

---

**Require:** Initialization  $z_0^0$ , total number of epochs  $N \geq 1$ , per-epoch iterates ( $K_n : n = 0, \dots, N - 1$ )

```

1: for  $n = 0, 1, \dots, N - 1$  do
2:    $z_{\text{out}} = \text{AG-OG}(z_0^n, z_0^n, z_0^n, K_n)$ 
3:   Set  $z_0^{n+1} \leftarrow z_{\text{out}}$ 
   //Warm-starting from the previous output
4: end for
5: Output:  $z_0^N$ 
    
```

---

that makes last iterates shrink toward the center of convergence. Equipped with Lemma 3.1, we are ready to state the following convergence theorem for discrete-time AG-OG:

**Theorem 3.3.** *Under Assumption 2.1 and setting the parameters as in Lemma 3.1, the output of Algorithm 1 satisfies:*

$$\|z_K^{\text{ag}} - z^*\|^2 \leq \left( \frac{4L}{\mu(K+1)^2} + \frac{2\sqrt{3+\sqrt{3}L_H}}{\mu(K+1)} \right) \|z_0 - z^*\|^2. \quad (3.7)$$

The proof of Theorem 3.3 is provided in §D.1. The selection of  $\alpha_k = \frac{2}{k+2}$  and  $\eta_k = \frac{k+2}{2L+\sqrt{3+\sqrt{3}L_H(k+2)}}$  is vital for Nesterov’s accelerated gradient descent to achieve desirable convergence behavior (Nesterov, 1983). This step-size choice is stronger than the ones used in previous techniques (Chen et al., 2017; Du et al., 2022), which is brought by a fine-tuned analysis of (D.6) in the proof of Theorem 3.3. The convergence rate in (3.7) for strongly convex problems is slow and is not even linear convergence. However, in the next subsection we show how a simple restarting technique not only achieves the linear convergence rate, but also matches the lower bound in Zhang et al. (2021a).

### 3.4 Improving Convergence Rates via Restarting and Scaling Reduction

Our algorithm design (Algorithm 2) utilizes the restarting technique, which is a well-established method to accelerate first-order methods in optimization literature (O’Donoghue & Candes, 2015; Roulet & d’Aspremont, 2017; Renegar & Grimmer, 2022). Our variant of restarting accelerates convergence through a novel approach inspired by contemporary variance-reduction strategies, similar to those presented in Li et al. (2022); Du et al. (2022). Our approach is distinct from previous ones (Kovalev et al., 2021; Thekumparampil et al., 2022; Jin et al., 2022) that incorporate the last iterate EG/OGDA with Nesterov’s acceleration. By incorporating the extrapolated step of Nesterov’s method as the average step of OGD and utilizing restarting, we use a two-timescale analysis and scaling reduction technique to achieve optimal results under all regimes. Although our algorithm is a multi-loop algorithm, the simplicity of restarting does not harm the practical aspect of our approach.

Normally, as  $f$  and  $g$  have different strong convexity parameters ( $\mu_f$  and  $\mu_g$ ), it is preferable in practice to have different

stepsizes for the descent step on  $f(x)$  and the ascent step on  $g(y)$  (Du et al., 2017; Lin et al., 2020a; Du et al., 2022). Accordingly, for our analysis we use a scaling reduction technique (Du et al., 2022) that allows us to consider applying a single scaling for all parameters without loss of generality. Setting  $\hat{\mathbf{y}} = \sqrt{\frac{\mu_g}{\mu_f}} \mathbf{y}$ , we have  $\nabla_{\hat{\mathbf{y}}} H(\mathbf{z}) = \sqrt{\frac{\mu_f}{\mu_g}} \nabla_{\mathbf{y}} H(\mathbf{z})$  and  $\nabla_{\hat{\mathbf{y}}} g(\mathbf{y}) = \sqrt{\frac{\mu_f}{\mu_g}} \nabla g(\mathbf{y})$ . Other scaling changes are listed as follows:

$$\begin{aligned} L &= L_f \vee \frac{\mu_f}{\mu_g} L_g, & L_H &= I_{xx} \vee I_{xy} \sqrt{\frac{\mu_f}{\mu_g}} \vee I_{yy} \frac{\mu_f}{\mu_g}, \\ \eta_{k,\mathbf{y}} &= \frac{\eta_k \mu_f}{\mu_g}, & \mu &= \mu_f, \end{aligned} \quad (3.8)$$

where by  $\eta_{k,\mathbf{y}}$  we mean that when updating  $\mathbf{z} = [\mathbf{x}; \mathbf{y}] \in \mathbb{R}^{n+m}$ , we adopt stepsize  $\eta_k$  on the first  $n$  coordinates and  $\eta_{k,\mathbf{y}}$  on the last  $m$  coordinates. With the new scaling and restarting, we obtain Algorithm 2, which we refer to as ‘‘AGOG with restarting.’’ The iteration complexity of AGOG with restarting is stated in the following Corollary 3.4.

**Corollary 3.4.** *Algorithm 2 outputs an  $\epsilon$ -optimal minimax point within a number  $\mathcal{O}(N)$  of iterates, for  $N$  satisfying:*

$$\begin{aligned} N &= \left( \sqrt{\frac{L}{\mu}} + \frac{L_H}{\mu} \right) \log \left( \frac{1}{\epsilon} \right) \\ &= \left( \sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g} + \frac{I_{xx}}{\mu_f} \vee \frac{I_{xy}}{\sqrt{\mu_f \mu_g}} \vee \frac{I_{yy}}{\mu_g}} \right) \log \left( \frac{1}{\epsilon} \right). \end{aligned} \quad (3.9)$$

We defer the proof of the corollary to §D.2. Our rate in Eq. (4.1) of Corollary 3.4 exactly matches the lower bound result in Zhang et al. (2021a) and therefore is optimal.

The analysis in Du et al. (2022), which also adopts a restarted scheme is most similar with ours. However, although OGDA can be written in past-EG form, the algorithm and theoretical analysis are fundamentally different (Golowich et al., 2020). For example, in contrast to EG, the non-expansiveness argument for OGDA does not achieve a unity prefactor (Mokhtari et al., 2020b). Our work proves a strict non-expansive property with prefactor 1, and our technique is new compared with existing EG-based analysis and existing the OGDA-based analysis.

### 3.5 Application to Separable Convex-Strongly-Concave (C-SC) Problem

To extend our strongly-convex-strongly-concave (SC-SC) AGOG algorithm complexity to the convex-strongly-concave (C-SC) setting, we define a *regularization reduction* method that modifies the objective via the addition of a regularization term, which gives the objective function  $\mathcal{L}_\epsilon(\mathbf{x}, \mathbf{y}) = \mathcal{L}(\mathbf{x}, \mathbf{y}) + \epsilon \|\mathbf{x}\|^2$ , where  $\epsilon$  is the desired accuracy of the solution. The following Theorem 3.5 provides the complexity analysis; see §D.3 for the proof.

**Theorem 3.5.** *The output of Algorithm 1 under the same assumptions and stepsize choices of Theorem 3.3 on the*

*objective function  $\mathcal{L}_\epsilon$  achieves a complexity of*

$$\mathcal{O} \left( \left( \sqrt{\frac{L_f}{\epsilon} \vee \frac{L_g}{\mu_g} + \frac{I_{xx}}{\epsilon} \vee \frac{I_{xy}}{\sqrt{\epsilon \mu_g}} \vee \frac{I_{yy}}{\mu_g}} \right) \log \left( \frac{1}{\epsilon} \right) \right)$$

*for the original C-SC problem.*

The work of Thekumparampil et al. (2022) also provides a C-SC case result that is obtained by utilizing the smoothing technique (Nesterov, 2005). Additionally, they present a direct C-SC algorithm without smoothing. On the other hand, Kovalev et al. (2021) focuses on a different perspective on the C-SC problem where  $\mathbf{x}$  and  $\mathbf{y}$  have strong interactions and obtains superlinear complexity of  $\log(\frac{1}{\epsilon})$ . However, both of these papers are limited to bilinear coupling terms. Our result, in contrast, targets a more general separable objective. Our complexity in Theorem 3.5 matches the complexity for regularized reduction in Thekumparampil et al. (2022). Furthermore, Theorem 3.5 is optimal in the bilinearly coupled regime. The reason is that  $\sqrt{\frac{L_f}{\epsilon}}$  is optimal for a pure minimization of convex  $f$  (Nesterov et al., 2018),  $\sqrt{\frac{L_g}{\mu_g}}$  is optimal for a pure maximization of strongly-concave  $g$  (Nesterov et al., 2018), and  $\frac{\|\mathbf{B}\|}{\sqrt{\epsilon \mu_g}}$  matches the lower bound of bilinearly coupled concave-convex minimax optimization (Ouyang & Xu, 2021) when  $f = 0$ .

### 3.6 Application to Bilinear Games

In this subsection, we examine the theoretical performance of our AGOG algorithm (Algorithm 1) when applied to the special case of bilinear games. By using proper averaging and scheduled restarting techniques, our algorithm is able to find near-optimal solutions and achieve an optimal sample complexity up to a constant prefactor. Additionally, we demonstrate that our algorithm can be reduced to a combination of the averaged iterates of the OGDA algorithm and a scheduled restarting procedure, which gives rise to a novel single-call algorithm that achieves an accelerated convergence rate on the bilinear minimax problem itself. Finally, we address the situation where there is stochasticity present in the problem. Throughout this section, we consider Problem (1.2) with  $\nabla f(\mathbf{x}), \nabla g(\mathbf{y})$  being zero almost surely. Moreover, we assume the following bilinear form:

$$I(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{B} \mathbf{y} + \mathbf{x}^\top \mathbf{u}_x + \mathbf{u}_y^\top \mathbf{y}, \quad (3.10)$$

where  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{y} \in \mathbb{R}^m$  with  $n = m$ ,  $\mathbf{B} \in \mathbb{R}^{n \times n}$  is square and full-rank, and  $\mathbf{u}_x, \mathbf{u}_y \in \mathbb{R}^n$  are two parameter vectors. Algorithm 1 reduces to an equivalent form of the OGDA algorithm (in the past extragradient form) with initial condition  $\mathbf{z}_0^{\text{ag}} = \mathbf{z}_{-\frac{1}{2}} = \mathbf{z}_0$ , which gives for all  $k \geq 1$ :

$$\begin{cases} \mathbf{z}_{k+\frac{1}{2}} &= \mathbf{z}_k - \eta_k H(\mathbf{z}_{k-\frac{1}{2}}), \end{cases} \quad (3.11a)$$

$$\begin{cases} \mathbf{z}_{k+1}^{\text{ag}} &= (1 - \alpha_k) \mathbf{z}_k^{\text{ag}} + \alpha_k \mathbf{z}_{k+\frac{1}{2}} \end{cases} \quad (3.11b)$$

$$\begin{cases} \mathbf{z}_{k+1} &= \mathbf{z}_k - \eta_k H(\mathbf{z}_{k+\frac{1}{2}}). \end{cases} \quad (3.11c)$$

By selecting the parameters  $\alpha_k = \frac{2}{k+2}$  and  $\eta_k = \frac{k+2}{2L+c_1 L_H(k+2)}$  with  $L = 0$  and  $c_1 = 2$  in (3.11), we can prove a boundedness lemma (Lemma F.1, presented in F), which is the bilinear game analogue of Lemma 3.1 with an improved scheme of stepsize and demonstrates the non-expansiveness of the last iterate of the OGDA algorithm. The proof is deferred to §G.9.

Non-expansiveness of the iterates further yields the following theorem whose proof is in §F.1.

**Theorem 3.6.** *When specified to the bilinear game case, setting the parameters as  $\alpha_k = \frac{2}{k+2}$  and  $\eta_k = \frac{1}{2L_H}$ , the output of update rules (3.11) satisfies*

$$\|z_K^{ag} - z^*\|^2 \leq \frac{64\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^2} \|z_0 - z^*\|^2. \quad (3.7)$$

Moreover, using the scheduled restarting technique, we obtain a complexity result that matches the lower bound of Ibrahim et al. (2020):

$$O\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B} \mathbf{B}^\top)}} \log\left(\frac{1}{\epsilon}\right)\right).$$

A similar result obtains in a stochastic setting; see §F.2.

## 4 Stochastic Accelerated Gradient Optimistic Gradient Descent Ascent

While the complexity result in Corollary 3.4 has also been obtained in Thekumparampil et al. (2022); Kovalev et al. (2021) and Jin et al. (2022), in addition to conceptual simplicity, our algorithm has the significant advantage that it yields a stochastic version and a convergence rate for the stochastic case. Indeed, the stochastic version of Algorithm 1 and 2 maintains an optimal convergence behavior. The stochastic AG-OG algorithm simply replaces each deterministic gradient with its stochastic counterpart, with noise represented by  $\zeta_t, \xi_t$ . The full stochastic AG-OG algorithm is shown in Algorithm 3 in §C.2.

Based on a boundedness lemma (Lemma D.7, presented in §D.4), which is the stochastic analogue of Lemma 3.1, we can proceed to our stochastic result. See §D.4 for the proof.

**Theorem 4.1.** *Under Assumptions 2.1 and 2.2, set the parameters as in Lemma D.7 and take  $\eta_k = \frac{k+2}{4L+D+4\sqrt{2+\sqrt{2}L_H(k+2)}}$ . Then the output of Algorithm 3 satisfies:*

$$\begin{aligned} & \mathbb{E}\|z_K^{ag} - z^*\|^2 \\ & \leq \left[ \frac{8L}{\mu(K+1)^2} + \frac{7.4(1+C^2)L_H}{\mu(K+1)} \right] \mathbb{E}\|z_0 - z^*\|^2 \\ & \quad + \frac{2(C + \frac{1}{C})\sigma}{\mu\sqrt{K+1}} \sqrt{\mathbb{E}\|z_0 - z^*\|^2}. \end{aligned}$$

**Remark 4.2.** Without knowledge of  $\mathbb{E}\|z_0 - z^*\|^2$ , we need a different scheme to set up the stepsize  $\eta_k$ . We assume

an upper bound on  $\|z_0 - z^*\|^2$  defined as  $\Gamma_0$  and let  $C = \frac{\Gamma_0}{\sqrt{\mathbb{E}\|z_0 - z^*\|^2}}$ . Then, we have that  $D = \frac{\sigma A(K)}{\Gamma_0}$  which is known. Thus,

$$\begin{aligned} & \mathbb{E}\|z_K^{ag} - z^*\|^2 \\ & \leq \left[ \frac{8L}{\mu(K+1)^2} + \frac{14.8L_H}{\mu(K+1)} \right] \Gamma_0^2 + \frac{4\sigma}{\mu\sqrt{K+1}} \Gamma_0. \end{aligned}$$

Analogous to the method in §3.4, we restart the S-AG-OG algorithm properly and achieve the following complexity:

**Corollary 4.3.** *With scheduled restarting imposed on top of Algorithm 1, Algorithm 2 outputs an  $\epsilon$ -optimal minimax point within  $O(N)$  iterations, for  $N$  satisfying:*

$$\begin{aligned} N &= \left( \sqrt{\frac{L}{\mu}} + \frac{L_H}{\mu} \right) \log\left(\frac{1}{\epsilon}\right) + \frac{\sigma^2}{\mu_f^2 \epsilon^2} \\ &= \left( \sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g} + \frac{I_{xx}}{\mu_f} \vee \frac{I_{xy}}{\sqrt{\mu_f \mu_g}} \vee \frac{I_{yy}}{\mu_g}} \right) \log\left(\frac{1}{\epsilon}\right) + \frac{\sigma^2}{\mu_f^2 \epsilon^2}. \end{aligned} \quad (4.1)$$

In the special case of bilinearly coupled SC-SC, the above result reduces to

$$\left( \sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{I_{xy}}{\sqrt{\mu_f \mu_g}} \right) \log\left(\frac{1}{\epsilon}\right) + \frac{\sigma^2}{\mu_f^2 \epsilon^2},$$

which matches that of Du et al. (2022) and is optimal. The reason is that the first term (i.e., bias error) matches the lower bound of bilinearly coupled SC-SC in Zhang et al. (2021a), and the second term (i.e., variance error) matches the statistical minimax rate.

## 5 Discussion

In this paper, we propose novel algorithms for both the deterministic setting (AG-OG) and a stochastic setting (S-AG-OG). These algorithms are characterized by structural interpretability and simplicity. To illustrate these properties, we have provided a continuous-time dynamical approximation that blends optimism with Nesterov's acceleration and an intuitive proof of convergence by leveraging a novel Lyapunov analysis. When discretizing the dynamics using AG-OG, we obtain desirable polynomial convergence behavior in discrete time. Furthermore, by properly restarting the algorithm, we obtain the corresponding AGOG and stochastic AGOG algorithms (with restarting) and prove theoretically that our restarted algorithms enjoys rate-optimal sample complexity for finding an  $\epsilon$ -accurate solution. Future directions include closing the gap between the upper and lower bounds for general separable minimax optimization, and generalizations to nonconvex settings.



## References

- Alacaoglu, A. and Malitsky, Y. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pp. 778–816. PMLR, 2022.
- Azizian, W., Scieur, D., Mitliagkas, I., Lacoste-Julien, S., and Gidel, G. Accelerating smooth games by manipulating spectral shapes. In *International Conference on Artificial Intelligence and Statistics*, pp. 1705–1715. PMLR, 2020.
- Bai, Y. and Jin, C. Provable self-play algorithms for competitive reinforcement learning. *arXiv preprint arXiv:2002.04017*, 2020.
- Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. *Robust optimization*, volume 28. Princeton university press, 2009.
- Beznosikov, A., Polyak, B., Gorbunov, E., Kovalev, D., and Gasnikov, A. Smooth monotone stochastic variational inequalities and saddle point problems—survey. *arXiv preprint arXiv:2208.13592*, 2022.
- Chavdarova, T., Gidel, G., Fleuret, F., and Lacoste-Julien, S. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32, 2019.
- Chen, Y., Lan, G., and Ouyang, Y. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.
- Cohen, M. B., Sidford, A., and Tian, K. Relative lipschitzness in extragradient methods and a direct recipe for acceleration. *arXiv preprint arXiv:2011.06572*, 2020.
- Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., and Song, L. Sbed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pp. 1125–1134. PMLR, 2018.
- Du, S. S. and Hu, W. Linear convergence of the primal-dual gradient method for convex-concave saddle point problems without strong convexity. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 196–205. PMLR, 2019.
- Du, S. S., Chen, J., Li, L., Xiao, L., and Zhou, D. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pp. 1049–1058. PMLR, 2017.
- Du, S. S., Gidel, G., Jordan, M. I., and Li, C. J. Optimal extragradient-based bilinearly-coupled saddle-point optimization. *arXiv preprint arXiv:2206.08573*, 2022.
- Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.
- Gidel, G., Hemmat, R. A., Pezeshki, M., Le Priol, R., Huang, G., Lacoste-Julien, S., and Mitliagkas, I. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1802–1811. PMLR, 2019.
- Golowich, N., Pattathil, S., and Daskalakis, C. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in neural information processing systems*, 2020.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- Gorbunov, E., Taylor, A., and Gidel, G. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. *Advances in neural information processing systems*, 2022.
- Hsieh, Y.-G., Iutzeler, F., Malick, J., and Mertikopoulos, P. On the convergence of single-call stochastic extragradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.
- Ibrahim, A., Azizian, W., Gidel, G., and Mitliagkas, I. Linear lower bounds and conditioning of differentiable games. In *International conference on machine learning*, pp. 4583–4593. PMLR, 2020.
- Jin, Y., Sidford, A., and Tian, K. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. *arXiv preprint arXiv:2202.04640*, 2022.
- Juditsky, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.
- Korpelevich, G. M. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976.
- Kovalev, D., Gasnikov, A., and Richtárik, P. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. *arXiv preprint arXiv:2112.15199*, 2021.
- Li, C. J., Yu, Y., Loizou, N., Gidel, G., Ma, Y., Le Roux, N., and Jordan, M. On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging. In *International Conference on Artificial Intelligence and Statistics*, pp. 9793–9826. PMLR, 2022.

- Liang, T. and Stokes, J. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 907–915. PMLR, 2019.
- Lin, T., Jin, C., and Jordan, M. I. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pp. 2738–2779. PMLR, 2020a.
- Lin, Z., Li, H., and Fang, C. Accelerated optimization for machine learning. *Nature Singapore: Springer*, 2020b.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- Metevlev, D., Rogozin, A., Gasnikov, A., and Kovalev, D. Decentralized saddle-point problems with different constants of strong convexity and strong concavity. *arXiv preprint arXiv:2206.00090*, 2022.
- Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pp. 1497–1507. PMLR, 2020a.
- Mokhtari, A., Ozdaglar, A. E., and Pattathil, S. Convergence rate of  $\mathcal{O}(1/k)$  for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 30(4):3230–3251, 2020b.
- Nemirovski, A. Prox-method with rate of convergence  $\mathcal{O}(1/t)$  for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1): 229–251, 2004.
- Nemirovskij, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.
- Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.
- Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.
- Nesterov, Y. E. A method for solving the convex programming problem with convergence rate  $\mathcal{O}(1/k^2)$ . In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.
- Ouyang, Y. and Xu, Y. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1):1–35, 2021.
- O’donoghue, B. and Candes, E. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.
- Palaniappan, B. and Bach, F. Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems*, 29, 2016.
- Popov, L. D. 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.
- Renegar, J. and Grimmer, B. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of Computational Mathematics*, 22(1):211–256, 2022.
- Roulet, V. and d’Aspremont, A. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017.
- Ryu, E. K., Yuan, K., and Yin, W. Ode analysis of stochastic gradient methods with optimism and anchoring for minimax problems. *arXiv preprint arXiv:1905.10899*, 2019.
- Sebbouh, O., Gower, R. M., and Defazio, A. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pp. 3935–3971. PMLR, 2021.
- Shapley, L. S. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.
- Su, W., Boyd, S., and Candes, E. A differential equation for modeling nesterov’s accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27, 2014.
- Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.
- Thekumparampil, K. K., He, N., and Oh, S. Lifted primal-dual method for bilinearly coupled smooth minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 4281–4308. PMLR, 2022.
- Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.
- Wang, Y. and Li, J. Improved algorithms for convex-concave minimax optimization. *Advances in Neural Information Processing Systems*, 33:4800–4810, 2020.
- Xie, G., Han, Y., and Zhang, Z. Dipa: An improved method for bilinear saddle point problems. *arXiv preprint arXiv:2103.08270*, 2021.

- Yang, J., Zhang, S., Kiyavash, N., and He, N. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33:5667–5678, 2020.
- Zhang, J., Hong, M., and Zhang, S. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, pp. 1–35, 2021a.
- Zhang, S., Yang, J., Guzmán, C., Kiyavash, N., and He, N. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pp. 482–492. PMLR, 2021b.
- Zhang, X., Aybat, N. S., and Gürbüzbalaban, M. Robust accelerated primal-dual methods for computing saddle points. *arXiv preprint arXiv:2111.12743*, 2021c.
- Zhao, R. Accelerated stochastic algorithms for convex-concave saddle-point problems. *Mathematics of Operations Research*, 47(2):1443–1473, 2022.

## A Examples of Separable Minimax Optimization

In this section, we use two examples to showcase the applications of formulation (1.2). We refer the readers for other examples in prior works such as Thekumparampil et al. (2022); Kovalev et al. (2021); Du et al. (2022). In the first example, we demonstrate how the parameters of a linear state-value function can be estimated by solving (1.2). In the second example, we illustrate that turning the constraint in a robust learning problem into penalty and we get an objective in the form of (1.2).

**Policy Evaluation in Reinforcement Learning.** The policy evaluation problem in RL can be formulated as a convex-concave bilinearly coupled minimax problem. Provided a sequence of  $\{s_t, a_t, r_t, s_{t+1}\}_{t=1}^n$  where

- (i)  $s_t, s_{t+1}$  are the current state (at time  $t$ ) and future state (at time  $t + 1$ ), respectively;
- (ii)  $a_t$  is the action at time  $t$  generated by policy  $\pi$ , that is,  $a_t = \pi(s_t)$ ;
- (iii)  $r_t = r(s_t, a_t)$  is the reward obtained after taken action  $a_t$  at state  $s_t$ .

Our goal is to estimate the value function of a fixed policy  $\pi$  in the discounted, infinite-horizon setting with discount factor  $\gamma \in (0, 1)$ , where for each state  $s$

$$V^\pi(s) \equiv \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_t = \pi(s_t), \forall t \geq 0 \right].$$

If a linear function approximation is adopted, i.e.  $V^\pi(s) = \phi(s)^\top \mathbf{x}$  where  $\phi(\cdot)$  is a feature mapping from the state space to feature space, we estimate the model parameter  $\mathbf{x}$  via minimizing the empirical *mean-squared projected Bellman error* (MSPBE):

$$\min_{\mathbf{x}} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_{\mathbf{C}^{-1}}^2. \quad (\text{A.1})$$

where  $\|\mathbf{x}\|_{\mathbf{M}} \equiv \sqrt{\mathbf{x}^\top \mathbf{M} \mathbf{x}}$  denotes the  $\mathbf{M}$ -norm, for positive semi-definite matrix  $\mathbf{M}$ , of an arbitrary vector  $\mathbf{x}$ , and

$$\begin{aligned} \mathbf{A} &= \frac{1}{n} \sum_{t=1}^n \phi(s_t) (\phi(s_t) - \gamma \phi(s_{t+1}))^\top, \\ \mathbf{b} &= \frac{1}{n} \sum_{t=1}^n r_t \phi(s_t), \quad \mathbf{C} = \frac{1}{n} \sum_{t=1}^n \phi(s_t) \phi(s_t)^\top. \end{aligned}$$

Applying first-order optimization directly to (A.1) would necessitate computing the inversion (and storing) of the matrix  $\mathbf{C}$ , or alternatively computing the matrix-vector product  $\mathbf{C}^{-1} \mathbf{v}$  for a vector  $\mathbf{v}$  at each step, and would be computationally costly or even prohibited. To circumvent inverting matrix  $\mathbf{C}$  a reformulation via *conjugate function* can be resorted to; that is, solving (A.1) is equivalent to solving the following minimax (or minimax) problem (Du et al., 2017; Du & Hu, 2019):

$$\min_{\mathbf{x}} \max_{\mathbf{y}} -\mathbf{y}^\top \mathbf{A} \mathbf{x} - \frac{1}{2} \|\mathbf{y}\|_{\mathbf{C}}^2 + \mathbf{b}^\top \mathbf{y}.$$

Such an instance falls under the category of minimax problem (1.2) where the individual component is convex-concave, and is further enhanced to be strongly-convex-strongly-concave when a regularizer term is added on top and  $\mathbf{C}$  is strictly positive definite.

**Robust Learning.** A robust learning or robust optimization problem targets to minimize an objective function (here the least-square) formulated as a minimax (minimax) optimization problem (Ben-Tal et al., 2009; Du & Hu, 2019; Thekumparampil et al., 2022)

$$\min_{\mathbf{x}} \max_{\mathbf{y}: \|\mathbf{y} - \mathbf{y}_0\| \leq \mathcal{R}} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2, \quad (\text{A.2})$$

where  $\mathbf{A}$  is a coefficient matrix and  $\mathbf{y}$  is a noisy observation vector, which is perturbed by a vector of  $\mathcal{R}$ -bounded norm. Transforming (A.2) to a penalized objective gives a formulation of  $\min_{\mathbf{x}} \max_{\mathbf{y}} \frac{1}{2} \|\mathbf{A}\mathbf{x} - \mathbf{y}\|^2 - \rho \|\mathbf{y} - \mathbf{y}_0\|^2$ . When  $\rho$  is selected to be strictly greater than  $\frac{1}{2}$ , we get a strongly-convex-strongly-concave bilinearly coupled minimax optimization problem.



## B Motivations of AG-OG from Continuous-Time Perspectives

We now consider the problem of accelerating the OGDA dynamics. We adopt a continuous-time perspective to motivate our algorithm design and guide our convergence analysis. The design hinges on adopting separate dynamics for the individual and coupling components of the minimax problem.<sup>5</sup>

**Dynamics of OGDA.** Our derivation is based on writing ordinary differential equation (ODE) representations for OGDA and Nesterov’s acceleration, and sequencing these ODEs in a particular way such that the OGDA component only applies to the coupling term  $\mathbf{x}^\top \mathbf{B}\mathbf{y}$ . We begin by rearranging the terms in the OGDA dynamics (3.2), yielding the following updates:

$$\mathbf{z}_{k+\frac{1}{2}} = \mathbf{z}_{k-\frac{1}{2}} - \eta H(\mathbf{z}_{k-\frac{1}{2}}) - \underbrace{\eta \left( H(\mathbf{z}_{k-\frac{1}{2}}) - H(\mathbf{z}_{k-\frac{3}{2}}) \right)}_{\epsilon_{k-1}}.$$

In comparison with some of the alternatives, our OGDA algorithm is featured by the *single-call property* that reuses gradients from past iterates and the sample and iteration complexities match up. Note in our paper  $H$  is a linear operator due to the bilinearity of  $\mathbf{x}^\top \mathbf{B}\mathbf{y}$ , but can be nonlinear in general. For deriving the continuous limit of the OGDA iteration, we let  $\widehat{Z}(t) = \mathbf{z}_{k+\frac{1}{2}}$  for  $t = (k + \frac{1}{2})\Delta t$  and set  $\eta = \Delta t \rightarrow 0$  as  $\Delta t \rightarrow 0$ . Moreover, by the Lipschitz property of  $H$  and using the fact that  $\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}} \leq \mathcal{O}(\eta) \rightarrow 0$ , we have that  $\epsilon_{k-1} \leq o(\eta) = o(\Delta t)$ . Dividing both sides by  $\Delta t$  and letting  $\Delta t \rightarrow 0$ , we obtain a first-order ODE for (3.2) that is equivalent to gradient descent-ascent:

$$\dot{\widehat{Z}} + H(\widehat{Z}) = 0. \quad (\text{B.1})$$

We note here that ODE (B.1) is a coarse-grained dynamics and its discretization can be either EG or GDA as well. However, as can be seen later in Proposition B.1 and Theorem B.2, the intuition of blending (B.1) with another dynamics motivates a strongly interpretable algorithm together with its theoretical analysis.

**Dynamics of Nesterov’s Acceleration.** Su et al. (2014) derived the following ODE for the standard form of Nesterov’s scheme:

$$\ddot{Z} + \frac{3}{t} \dot{Z} + \nabla F(Z) = 0. \quad (\text{B.2})$$

Note that Eq. (3.3) exhibits the same ODE (B.2) with  $Z(k\Delta t) = \mathbf{z}_k^{\text{ag}}$  and  $\Delta t \rightarrow 0$ . Moreover, a compact way of writing (B.2) is

$$\dot{\widehat{Z}} + \frac{t}{2} \nabla F(Z) = 0, \quad (\text{B.3})$$

where  $\widehat{Z} = Z + \frac{t}{2} \dot{Z}$ . Intuitively, while  $Z$  represents the continuous dynamics of  $\mathbf{z}_k^{\text{ag}}$  (averaging point),  $\widehat{Z}$  represents the dynamics of  $\mathbf{z}_k$ . The equality  $\widehat{Z} = Z + \frac{t}{2} \dot{Z}$  exactly matches the relationship between  $\mathbf{z}_k^{\text{ag}}$  and  $\mathbf{z}_k$  shown in the update rule (3.3c). More specifically, starting from the lens of discrete time updates (3.3c) and (3.3b), where  $\mathbf{z}_{k+1} = \frac{k}{2}(\mathbf{z}_{k+1}^{\text{ag}} - \mathbf{z}_k^{\text{ag}}) + \mathbf{z}_{k+1}^{\text{ag}}$ ,  $\mathbf{z}_{k+1} = \mathbf{z}_k - \eta_k \nabla F(\mathbf{z}_k^{\text{md}})$ , by choosing  $L' > L$  where  $L$  is the Lipschitz constant of  $f$ ,  $\eta_k = \frac{k}{2L'}$ ,  $t = k\Delta t$ ,  $\Delta t = \frac{1}{\sqrt{L'}}$  and finally let  $L' \rightarrow \infty$ , we conclude that (also noting that  $\mathbf{z}_k^{\text{md}} - \mathbf{z}_{k+1}^{\text{ag}}$  is  $o(\Delta t)$ ):

$$\widehat{Z} = Z + \frac{t}{2} \dot{Z}, \quad \dot{\widehat{Z}} + \frac{t}{2} \nabla F(Z) = 0,$$

which again leads to the ODE (B.2). On a related note, the dynamics of  $Z$  is the averaged iteration and is highly related to the idea of anchoring (Ryu et al., 2019).

**Blending the Two Dynamics.** To solve the separable minimax problem (1.2), we blend the aforementioned continuous-time dynamics (B.1) corresponding to OGDA on the (linear) coupling component  $H$ , and also (B.3) that conducts Nesterov’s acceleration on the individual component  $f$ . Note that Eqs. (B.3) and (B.1) admit different timescales so the resulting continuous-time dynamics of AG-OG cannot be a direct addition, and a modified timescale analysis is essential to obtain the desired result.

We derive the following proposition, where the derivation of the ODE approximation is deferred to §E:

<sup>5</sup>In fact, a direct acceleration via momentum leads to sub-optimality for the refined case (1.2) where the condition number of the three parts are entangled (Gidel et al., 2019).

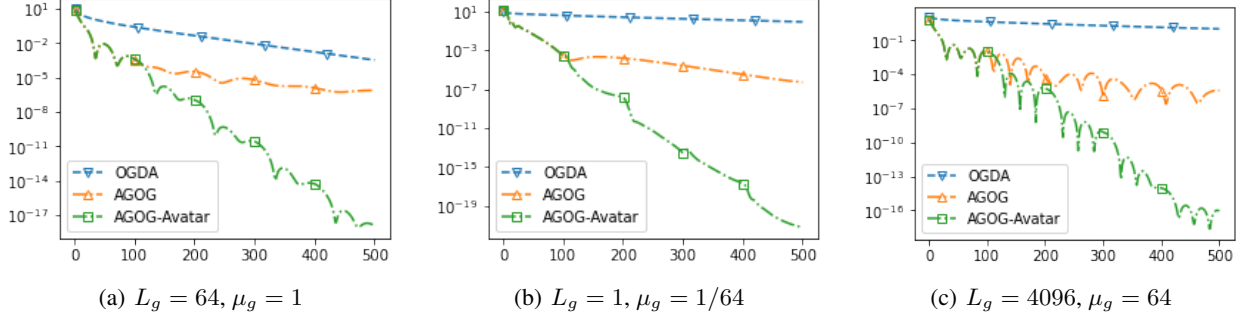


Figure 1. Comparison with OGDA on different problem sets (Deterministic)

**Proposition B.1.** We let  $\eta_k = \frac{k+1}{2L' + \sqrt{3+\sqrt{3}}L'_H(k+1)}$ , where we assume arbitrarily large  $L' > L$  and  $L'_H > L_H$ . Furthermore by letting  $\delta = \frac{1}{\sqrt{L'}\sqrt{L'_H}}$ , as  $\delta \rightarrow 0^+$  and  $\frac{L'_H}{\sqrt{L'}} \rightarrow c \in [0, \infty)$  the ODE for the AG-OG Algorithm 1 under this scaling condition is

$$\dot{\hat{Z}} + c_t \left( H(\hat{Z}) + \nabla F(Z) \right) = 0, \quad \hat{Z} = Z + \frac{t}{2} \dot{Z}, \quad (\text{B.4})$$

where  $c_t \equiv \frac{t(1 \vee c)^2}{2 + \sqrt{3+\sqrt{3}}c(1 \vee c)t}$ . Eq. (B.4) is also equivalent to a single-line higher-order ODE

$$\frac{t}{2} \ddot{Z} + \frac{3}{2} \dot{Z} + c_t \left( \nabla F(Z) + H(Z) + \frac{t}{2} \nabla H(Z) \dot{Z} \right) = 0. \quad (\text{B.5})$$

We finalize this subsection with a convergence rate result for the combined dynamics in ODE (B.4). Throughout the paper we use  $z^*$  to denote a solution to Problem (1.2). We conduct a standard continuous-time analysis based on the following Lyapunov energy function:  $\mathcal{E} = t^2(F(Z) - F(z^*)) + \frac{t}{c_t} \|\hat{Z} - z^*\|^2$ , and compute the time-derivative  $\dot{\mathcal{E}}$ . In contrast with existing analysis for the optimization setting (Su et al., 2014), our analysis repose on the use of a continuous-time Gronwall-inequality-based analysis. The overall result is the following theorem, whose proof is deferred to §E.1:

**Theorem B.2.** We define the gap function as  $V(Z) = F(Z) - F(z^*) + \langle Z - z^*, H(z^*) \rangle$  and have

$$V(Z(T)) \leq \frac{1}{T} \left( \frac{2}{(1 \vee c)^2 T} + \frac{\sqrt{3+\sqrt{3}}c}{1 \vee c} \right) \|z_0 - z^*\|^2.$$

Note in the special case where  $c = 0$ , the result recovers the standard continuous-time convergence rate as in Su et al. (2014). Later in §3.3 we will prove a similar result (Theorem 3.3) in the discrete-time context. We note in the meantime that the continuous-time analysis is much simpler and more informative than the (comparatively more complicated) discrete-time analysis.

## C Experiments

In this section, we study the performance of our AGOG with restartingAlgorithm empirically. We study both deterministic [§C.1] and stochastic settings [§C.2]. In each of these settings we compare our algorithm with the state-of-the-art experimental result.

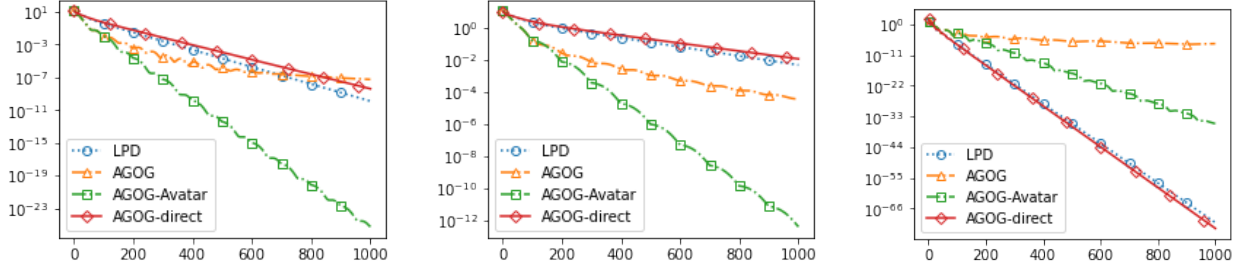
### C.1 Deterministic Setting

We present results on synthetic quadratic game datasets:

$$\mathbf{x}^\top A_1 \mathbf{x} + \mathbf{y}^\top A_2 \mathbf{x} - \mathbf{y}^\top A_3 \mathbf{y}, \quad (\text{C.1})$$

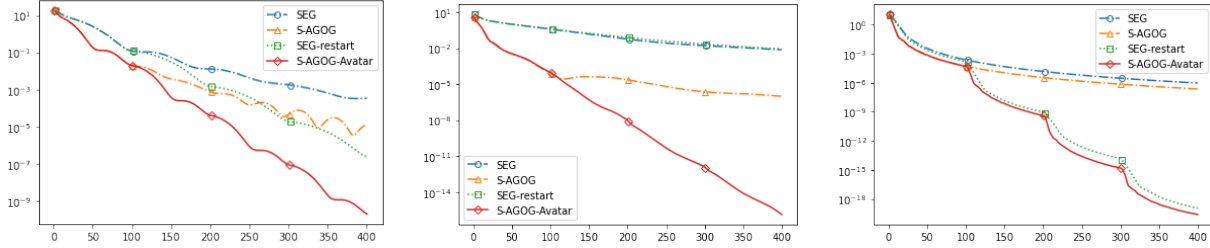
with various settings of the eigenvalues of  $A_1, A_2, A_3$ .

**Comparison with OGDA.** We use the single-call OGDA algorithm (Gidel et al., 2018; Hsieh et al., 2019) as the baseline. In Figure 1 we plot the AGOG algorithm and the AGOG with restarting algorithm under three different instances. We use stepsize  $\eta_k = \frac{k+2}{2L + \sqrt{3+\sqrt{3}}L_H(k+2)}$  in both the AGOG and the AGOG with restarting algorithms and restart AGOG with



(a)  $L_f = L_g = \mu_f = \mu_g = 1, L_H = 356, \mu_H = 101$  (b)  $L_f = L_g = \mu_f = \mu_g = 1, L_H = 725, \mu_H = 101$  (c)  $L_f = L_g = 100, \mu_f = \mu_g = 1, L_H = \mu_H = 1$

Figure 2. Comparison with LPD on different problem sets (Deterministic)



(a)  $L_f = L_g = \mu_f = \mu_g = 1, L_H = 356, \mu_H = 101, \sigma = 0.1$  (b)  $L_f = L_g = 10, \mu_f = \mu_g = \mu_H = 1, L_H = 11, \sigma = 0.1$  (c)  $L_f = L_g = 1, \mu_f = \mu_g = 1/8, L_H = \mu_H = 1, \sigma = 0.1$

Figure 3. Comparison of algorithms on different problem sets (Stochastic)

restarting once every 100 iterates. For the OGDA algorithm, we take stepsize  $\eta = \frac{1}{2(L \vee L_H)}$ . For the parameters of the problem (C.1), we fix  $L_H = 1, L_f = 64, \mu_f = 1$  and change different choices of  $L_g, \mu_g$ . In Figure 1(a) we take  $L_g = 64, \mu_g = 1$ . In Figure 1(b) we take  $L_g = 1, \mu_g = 1/64$  and in Figure 1(c) we take  $L_g = 4096, \mu_g = 64$ . We see from 1(a), 1(b) and 1(c) when the problem has different  $L_f, \mu_f$  and  $L_g, \mu_g$ , changing  $L_g, \mu_g$  has larger impact on OGDA than on AGOG, matching our theoretical observations.

**Comparison with LPD.** Next, we focus on comparison to the Lifted Primal-Dual (LPD) algorithm (Thekumparampil et al., 2022). We implement the AGOG algorithm and its restarted version, the AGOG with restarting. Additionally, inspired by the technique of a single-loop direct-approach in Du et al. (2022), we consider a single-loop algorithm named AGOG-Direct that takes advantage of the strongly-convex strongly-concave property of the problem. We refer readers to Du et al. (2022) for the "direct" method. The parameters of LPD are chosen as described in Thekumparampil et al. (2022). For our AGOG and AGOG with restarting algorithms, we take  $\eta_k = \frac{k+2}{2L + \sqrt{3 + \sqrt{3}L_H}(k+2)}$  and the scaling parameters are taken as in Eq. (3.8). For the AGOG-direct algorithm, we take  $\eta = \frac{1}{(1 + \sqrt{L/\mu_f + (\sqrt{3 + \sqrt{3}L_H})^2/\mu_f^2})\mu_f}$  with the same set of scaling

parameters. We restart AGOG with restarting once every 100 iterates.

In Figure 2(a), the bilinear coupling component  $\mathbf{y}^\top A_2 \mathbf{x}$  is the dominant part. In Figure 2(b), we set the eigenvalues of  $A_2$  even larger than in Figure 2(a). In Figure 2(c),  $\mathbf{x}^\top A_1 \mathbf{x}$  and  $\mathbf{y}^\top A_3 \mathbf{y}$  are the dominant terms. More details on the specific designs of the matrices are shown in the caption of the corresponding figures.

We see from Figures 2(a) and 2(b) that AGOG with restarting (green line) outperforms LPD and MP in regimes where the bilinear term dominates, and when the eigenvalues of the coupling matrix increase, the performance of AGOG with restarting relative to other algorithms is enhanced. This is in accordance with our theoretical analysis. In addition, AGOG with restarting outperforms its non-restarted version (orange line) which has a gentle slope at the end. On the other hand, when the individual component dominates, our AGOG-direct (red line) slightly outperforms LPD. Moreover, AGOG-direct and LPD almost overlap in 2(a) and 2(b).

## C.2 Stochastic Setting

We compared stochastic AGOG and its restarted version (S-AGOG) with Stochastic extragradient (SEG) SEG with restarting, respectively (cf. Li et al., 2022). The complete algorithm is shown in 3. We note that we refer to the averaged iterates version

---

**Algorithm 3** Stochastic AcceleratedGradient-OptimisticGradient (S-AG-OG)( $z_0^{\text{ag}}, z_0, z_{-1/2}, K$ )
 

---

```

1: for  $k = 0, 1, \dots, K - 1$  do
2:    $z_k^{\text{md}} = (1 - \alpha_k)z_k^{\text{ag}} + \alpha_k z_k$ 
3:    $z_{k+\frac{1}{2}} = z_k - \eta_k \left( \tilde{H}(z_{k-\frac{1}{2}}; \zeta_{k-\frac{1}{2}}) + \nabla \tilde{F}(z_k^{\text{md}}; \xi_k) \right)$ 
4:    $z_{k+1}^{\text{ag}} = (1 - \alpha_k)z_k^{\text{ag}} + \alpha_k z_{k+\frac{1}{2}}$ 
5:    $z_{k+1} = z_k - \eta_k \left( \tilde{H}(z_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}) + \nabla \tilde{F}(z_k^{\text{md}}; \xi_k) \right)$ 
6: end for
7: Output:  $z_K^{\text{ag}}$ 
    
```

---

of SEG everywhere when using SEG. For SEG and SEG-restart, we use stepsize  $\eta_k = \frac{1}{2(L \vee L_H)}$ . For AGOG and AGOG with restarting, we use stepsize  $\eta_k = \frac{k+2}{2L + \sqrt{3+3}L_H(k+2)}$ . We restart every 100 gradient calculations for both SEG-restart and AGOG-restart.

We use the same quadratic game setting as in (C.1) except that we assume access only to noisy estimates of  $A_1, A_2, A_3$ . We add Gaussian noise to  $A_1, A_2, A_3$  with  $\sigma = 0.1$  throughout this experiment. We plot the squared norm error with respect to the number of gradient computations in Figure 3. In 3(a) we consider larger eigenvalues for  $A_2$  than  $A_1, A_3$ . In 3(b), we let  $A_1, A_2, A_3$  to be approximately of the same scale. In 3(c), as the scale of the eigenvalues shrinks, the noise is relatively larger than in 3(a) and 3(b). The specific choice of parameters are shown in the caption of the corresponding figures. We see from 3(a), 3(c) and 3(c) that stochastic AGOG with restarting achieves a more desirable convergence speed than SEG-restart. Also, the restarting technique significantly accelerates the convergence, validating our theory.

## D Proof of Main Convergence Results

This section collects the proofs of our main results, Theorem 3.3 [§D.1], Corollary 3.4 [§D.2], and Theorem 4.1 [§D.4].

### D.1 Proof of Theorem 3.3

*Proof.*[Proof of Theorem 3.3] We define the point-wise primal-dual gap function as:

$$V(z, z') := F(z) - F(z') + \langle H(z'), z - z' \rangle. \quad (\text{D.1})$$

Our proof proceeds in the following steps:

**Step 1: Estimating weighted temporal difference in squared norms.** We first prove a result on bounding the temporal difference of the point-wise primal-dual gap between  $z_k^{\text{ag}}$  and  $z^*$ , whose proof is delayed to §G.5.

**Lemma D.1.** *For arbitrary  $\alpha_k \in (0, 1]$  the iterates of Algorithm 1 satisfy for  $k = 1, \dots, K$  almost surely*

$$V(z_{k+1}^{\text{ag}}, \omega_z) - (1 - \alpha_k)V(z_k^{\text{ag}}, \omega_z) \leq \underbrace{\alpha_k \left\langle \nabla F(z_k^{\text{md}}) + H(z_{k+\frac{1}{2}}), z_{k+\frac{1}{2}} - \omega_z \right\rangle}_I + \underbrace{\frac{L\alpha_k^2}{2} \|z_{k+\frac{1}{2}} - z_k\|^2}_{II}. \quad (\text{D.2})$$

Note that in Lemma D.1, the term I is an inner product with the gradient term and reduces to  $\langle \nabla f(z_k), z_k - \omega_z \rangle$  of the vanilla gradient algorithm if acceleration and optimistic-gradient are removed. The squared term II is brought by gradient evaluated at  $z_k^{\text{md}}$ .

Additionally, throughout the proof of Lemma D.1, we only leverage the convexity and  $L$ -smoothness of  $f$  and the property of  $H$  (2.2), as well as the update rules in Line (2) and Line (4). The proof involves no update rules regarding the gradient updates and hence Lemma D.1 holds for the stochastic case as well.

Next, to further bound the inner product term I, we introduce a general proposition that holds for two updates starting from the same point. Proposition D.2 is a slight modification from the proof of Proposition 4.2 in Chen et al. (2017) and analogous to Lemma 7.1 in Du et al. (2022). We omit the proof here as the argument comes from simple algebraic tricks. Readers can refer to Du et al. (2022) for more details.

**Proposition D.2** (Proposition 4.2 in Chen et al. 2017 and Lemma 7.1 in Du et al. (2022)). *Given an initial point  $\theta \in \mathbb{R}^d$ , two update directions  $\delta_1, \delta_2 \in \mathbb{R}^d$ , and the corresponding results  $\varphi_1, \varphi_2 \in \mathbb{R}^d$  satisfying:*

$$\varphi_1 = \theta - \delta_1, \quad \varphi_2 = \theta - \delta_2. \quad (\text{D.3})$$

*For any point  $z \in \mathbb{R}^d$  we have*

$$\langle \delta_2, \varphi_1 - z \rangle \leq \frac{1}{2} \|\delta_2 - \delta_1\|^2 + \frac{1}{2} [\|\theta - z\|^2 - \|\varphi_2 - z\|^2 - \|\theta - \varphi_1\|^2]. \quad (\text{D.4})$$



Noting that the gradient term  $\nabla F(z_k^{\text{md}}) + H(z_{k+\frac{1}{2}})$  in  $I$  of inequality (D.2) of Lemma D.1 has been used in updating  $z_k$  to  $z_{k+1}$  in Line (5) in Algorithm 1. Comparing Line (5) with Line (3) and by letting  $\theta = z_k$ ,  $\varphi_1 = z_{k+\frac{1}{2}}$ ,  $\varphi_2 = z_{k+1}$  in Proposition D.2, we obtain an upper bound for the inner product term  $I$ :

$$\begin{aligned} \eta_k \cdot I &\leq \frac{1}{2} \eta_k^2 \|H(z_{k+\frac{1}{2}}) - H(z_{k-\frac{1}{2}})\|^2 + \frac{1}{2} \left[ \|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2 - \|z_{k+\frac{1}{2}} - z_k\|^2 \right] \\ &\leq \frac{\eta_k^2 L_H^2}{2} \|z_{k+\frac{1}{2}} - z_{k-\frac{1}{2}}\|^2 + \frac{1}{2} \left[ \|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2 - \|z_{k+\frac{1}{2}} - z_k\|^2 \right], \end{aligned} \quad (\text{D.5})$$

where the last inequality is by properties of  $H$  and the definition of  $L_H$ . Combining Eq. (D.5) with Eq. (D.2), we obtain

$$\begin{aligned} V(z_{k+1}^{\text{ag}}, \omega_z) - (1 - \alpha_k) V(z_k^{\text{ag}}, \omega_z) &\leq \frac{\eta_k \alpha_k L_H^2}{2} \|z_{k+\frac{1}{2}} - z_{k-\frac{1}{2}}\|^2 \\ &\quad + \frac{\alpha_k}{2\eta_k} \left[ \|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2 - \|z_{k+\frac{1}{2}} - z_k\|^2 \right] + \frac{L\alpha_k^2}{2} \|z_{k+\frac{1}{2}} - z_k\|^2. \end{aligned} \quad (\text{D.6})$$

**Step 2: Building and solving the recursion.** To build a iterative rule that connects the previous iterate with the current iterate, we first apply the following lemma to connect  $\|z_{k+\frac{1}{2}} - z_{k-\frac{1}{2}}\|^2$  and  $\|z_{k+\frac{1}{2}} - z_k\|^2$  and reduce Eq. (D.6) to compositions of  $\{\|z_k - \omega_z\|^2\}_{k=0, \dots, K-1}$  and  $\{\|z_{k+\frac{1}{2}} - z_k\|^2\}_{k=0, \dots, K-1}$  terms. The proof of Lemma D.3 is deferred to §G.6.

**Lemma D.3.** *For any stepsize sequence  $\{\eta_k\}_{k=0, \dots, K-1}$  satisfying for some positive constant  $c > 0$  and the Lipchitz parameter  $L_H$  that  $\eta_k L_H \leq \sqrt{\frac{c}{2}}$  holds for all  $k$ . In Algorithm 1, the following holds for any  $k \in [K-1]$ :*

$$\left\| z_{k+\frac{1}{2}} - z_{k-\frac{1}{2}} \right\|^2 \leq 2c^k \sum_{\ell=0}^k c^{-\ell} \left\| z_{\ell+\frac{1}{2}} - z_{\ell} \right\|^2. \quad (\text{D.7})$$

Combining Eq. (D.7) with (D.6), bringing in the choice of  $\alpha_k = \frac{2}{k+2}$  and by rearranging the terms, we obtain the following relation:

$$\begin{aligned} V(z_{k+1}^{\text{ag}}, \omega_z) - \frac{k}{k+2} V(z_k^{\text{ag}}, \omega_z) &\leq \frac{1}{\eta_k(k+2)} \left[ \|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2 \right] \\ &\quad - \left( \frac{1}{\eta_k(k+2)} - \frac{2L}{(k+2)^2} \right) \|z_{k+\frac{1}{2}} - z_k\|^2 + \frac{2\eta_k L_H^2}{k+2} \sum_{\ell=0}^k c^{k-\ell} \|z_{\ell+\frac{1}{2}} - z_{\ell}\|^2. \end{aligned}$$

Multiplying both sides by  $(k+2)^2$ , we obtain

$$\begin{aligned} (k+2)^2 V(z_{k+1}^{\text{ag}}, \omega_z) - [(k+1)^2 - 1] V(z_k^{\text{ag}}, \omega_z) &\leq \frac{k+2}{\eta_k} \left[ \|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2 \right] \\ &\quad - \left( \frac{k+2}{\eta_k} - 2L \right) \|z_{k+\frac{1}{2}} - z_k\|^2 + 2(k+2)\eta_k L_H^2 \sum_{\ell=0}^k c^{k-\ell} \|z_{\ell+\frac{1}{2}} - z_{\ell}\|^2. \end{aligned}$$

Taking  $\eta_k = \frac{k+2}{2L + \sqrt{\frac{2}{c}} L_H (k+2)}$ , we have

$$\frac{k+2}{\eta_k} - 2L \geq \sqrt{\frac{2}{c}} (k+2) L_H,$$

and the previous inequality reduces to

$$\begin{aligned} (k+2)^2 V(z_{k+1}^{\text{ag}}, \omega_z) - [(k+1)^2 - 1] V(z_k^{\text{ag}}, \omega_z) &\leq \left( 2L + \sqrt{\frac{2}{c}} L_H (k+2) \right) \left[ \|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2 \right] \\ &\quad - \sqrt{\frac{2}{c}} (k+2) L_H \|z_{k+\frac{1}{2}} - z_k\|^2 + (k+2) \sqrt{2c} L_H \sum_{\ell=0}^k c^{k-\ell} \|z_{\ell+\frac{1}{2}} - z_{\ell}\|^2 \end{aligned}$$

Rearranging and summing over  $k$  from 0 to  $K - 1$ , we have

$$\begin{aligned}
 & [(K+1)^2 - 1] V(\mathbf{z}_K^{\text{ag}}, \omega_{\mathbf{z}}) + \left( 2L + \sqrt{\frac{2}{c}} L_H (K+1) \right) \|\mathbf{z}_K - \omega_{\mathbf{z}}\|^2 \\
 & \leq \left( 2L + \sqrt{\frac{2}{c}} L_H \right) \|\mathbf{z}_0 - \omega_{\mathbf{z}}\|^2 + \sqrt{\frac{2}{c}} L_H \sum_{k=0}^{K-1} \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \sum_{k=0}^{K-1} V(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}) \\
 & \quad - \underbrace{\sqrt{\frac{2}{c}} L_H \sum_{k=0}^{K-1} (k+2) \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2}_{\text{III}_1} + \underbrace{\sqrt{2c} L_H \sum_{k=0}^{K-1} (k+2) \sum_{\ell=0}^k c^{k-\ell} \|\mathbf{z}_{\ell+\frac{1}{2}} - \mathbf{z}_\ell\|^2}_{\text{III}_2}.
 \end{aligned}$$

Simple algebra yields

$$\text{III}_2 = \sum_{\ell=0}^{K-1} \|\mathbf{z}_{\ell+\frac{1}{2}} - \mathbf{z}_\ell\|^2 \sum_{k=\ell}^{K-1} (k+2) c^{k-\ell} \leq \sum_{\ell=0}^{K-1} \left[ \frac{\ell+2}{1-c} + \frac{c}{(1-c)^2} \right] \|\mathbf{z}_{\ell+\frac{1}{2}} - \mathbf{z}_\ell\|^2.$$

Straightforward derivations give that if we choose  $c = \frac{2}{3+\sqrt{3}}$ , the inequality

$$\sqrt{\frac{2}{c}} (k+2) \geq \sqrt{2c} \left[ \frac{k+2}{1-c} + \frac{c}{(1-c)^2} \right]$$

holds for all  $k \geq 0$ . Thus, summing  $\text{III}_1$  and  $\text{III}_2$  terms we have

$$-\sqrt{\frac{2}{c}} L_H \text{III}_1 + 2\sqrt{\frac{c}{2}} L_H \text{III}_2 \leq 0.$$

Finally, we solve the recursion and conclude

$$\begin{aligned}
 & [(K+1)^2 - 1] V(\mathbf{z}_K^{\text{ag}}, \omega_{\mathbf{z}}) + \left( 2L + \sqrt{\frac{2}{c}} L_H (K+1) \right) \|\mathbf{z}_K - \omega_{\mathbf{z}}\|^2 \\
 & \leq \left( 2L + \sqrt{\frac{2}{c}} L_H \right) \|\mathbf{z}_0 - \omega_{\mathbf{z}}\|^2 + \sqrt{\frac{2}{c}} L_H \sum_{k=0}^{K-1} \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \sum_{k=0}^{K-1} V(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}). \tag{D.8}
 \end{aligned}$$

**Step 3: Proving  $\mathbf{z}_k$  stays bounded with respect to  $\mathbf{z}^*$ .** In Lemma 3.1, we show that  $\mathbf{z}_k$  is always bounded within the ball centered at  $\mathbf{z}^*$  with radius  $\|\mathbf{z}_0 - \mathbf{z}^*\|$  and the proof of this lemma is presented in §G.3.

**Lemma D.4** (Lemma 3.1). *Under Assumption 2.1. Set the parameters as  $L = L_f \vee L_g$ ,  $L_H = \sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}$ ,  $\eta_k = \frac{k+2}{2L + \sqrt{3+\sqrt{3}} L_H (k+2)}$  and  $\alpha_k = \frac{2}{k+2}$  in Algorithm 1, at any iterate  $k < K$ ,  $\mathbf{z}_k$  stays within the region defined by the initialization  $\mathbf{z}_0$ :*

$$\|\mathbf{z}_k - \mathbf{z}^*\|^2 \leq \|\mathbf{z}_0 - \mathbf{z}^*\|^2,$$

where we use  $\mathbf{z}^*$  to denote the unique solution of Problem (1.2).

**Step 4: Combining everything together.** Bringing the bounded iterates results in Lemma 3.1 into the recursion (D.8), setting  $\omega_{\mathbf{z}} = \mathbf{z}^*$  and rearranging, we obtain the following:

$$\begin{aligned}
 (K+1)^2 V(\mathbf{z}_K^{\text{ag}}, \mathbf{z}^*) & \leq (K+1)^2 V(\mathbf{z}_K^{\text{ag}}, \mathbf{z}^*) + \left( 2L + \sqrt{\frac{2}{c}} L_H (K+1) \right) \|\mathbf{z}_K - \mathbf{z}^*\|^2 \\
 & \leq \left( 2L + \sqrt{\frac{2}{c}} L_H (K+1) \right) \|\mathbf{z}_0 - \mathbf{z}^*\|^2.
 \end{aligned}$$

Dividing both sides by  $(K+1)^2$  and noting that  $V(\mathbf{z}_K^{\text{ag}}, \mathbf{z}^*) \geq \frac{\mu}{2} \mathbb{E} \|\mathbf{z}_K^{\text{ag}} - \mathbf{z}^*\|^2$ , bringing in the choice of  $c = \frac{2}{3+\sqrt{3}}$  and we conclude our proof of Theorem 3.3.  $\square$

We finally remark that a limitation of this convergence rate bound is that the coefficient for  $L_H$  in our stepsize choosing scheme is  $\sqrt{3 + \sqrt{3}} \approx 2.175$  while an improved stepsize in this special case is  $\frac{1}{2L_H}$ , yielding a sharper coefficient 2. Although the slight difference in constant factors does not harm the practical performance drastically, we anticipate that this constant might be further improved and leave it to future work.

## D.2 Proof of Corollary 3.4

*Proof.*[Proof of Corollary 3.4] The proof of restarting argument is direct. By Eq. (3.7), if we want  $\|z_K^{\text{ag}} - z^*\|^2 \leq \frac{1}{e} \|z_0 - z^*\|^2$  to hold, we can choose  $K$  such that

$$\frac{4L}{\mu(K+1)^2} \leq \frac{1}{2e}, \quad \frac{2\sqrt{3 + \sqrt{3}}L_H}{\mu(K+1)} \leq \frac{1}{2e}.$$

This is equivalent to

$$K+1 \geq \sqrt{\frac{8eL}{\mu}}, \quad K+1 \geq \frac{4e\sqrt{3 + \sqrt{3}}L_H}{\mu}.$$

For a given threshold  $\epsilon > 0$ , with the output of every epoch satisfying  $\|z_K^{\text{ag}} - z^*\|^2 \leq \frac{1}{e} \|z_0 - z^*\|^2$ , the total epochs required to obtain an  $\epsilon$ -optimal minimax point would be  $\log\left(\frac{\|z_0 - z^*\|^2}{\epsilon}\right)$ . Thus, the total number of iterates required to get within the  $\epsilon$  threshold would be:

$$\mathcal{O}\left(\sqrt{\frac{L}{\mu}} + \frac{L_H}{\mu}\right) \cdot \log\left(\frac{1}{\epsilon}\right).$$

Bringing the choice of scaling parameters in (3.8) and we conclude our proof of Corollary 3.4.  $\square$

## D.3 Proof of Theorem 3.5

*Proof.*[Proof of Theorem 3.5] For minimax problem, we recall that we define the primal-dual gap function as:

$$V(z, z') = F(z) - F(z') + \langle H(z'), z - z' \rangle.$$

We have for any pair of parameters  $(\hat{x}, \hat{y})$  and  $(x, y)$ :

$$\begin{aligned} \mathcal{L}(\hat{x}, y) - \mathcal{L}(x, \hat{y}) &= f(\hat{x}) - f(x) + g(\hat{y}) - g(y) + I(\hat{x}, y) - I(x, \hat{y}) \\ &= f(\hat{x}) - f(x) + g(\hat{y}) - g(y) + I(\hat{x}, y) - I(x, y) + I(x, y) - I(x, \hat{y}) \\ &\leq f(\hat{x}) - f(x) + g(\hat{y}) - g(y) + \langle H(z), \hat{z} - z \rangle \leq V(\hat{z}, z). \end{aligned}$$

Similarly for the regularized problem we define

$$V_\epsilon(z, z') = V(z, z') + \frac{\epsilon}{2} \|x\|^2 - \frac{\epsilon}{2} \|x'\|^2,$$

and by definition of  $\mathcal{L}_\epsilon$  we have

$$\mathcal{L}_\epsilon(\hat{x}, y) - \mathcal{L}_\epsilon(x, \hat{y}) \leq V_\epsilon(\hat{z}, z),$$

and moreover,

$$\max_{y \in \mathcal{Y}} \mathcal{L}_\epsilon(\hat{x}, y) - \min_{x \in \mathcal{X}} \mathcal{L}_\epsilon(x, \hat{y}) \leq \max_{z \in \mathcal{X} \times \mathcal{Y}} V_\epsilon(\hat{z}, z).$$

By the AGOG algorithm 1 on the regularized objective  $\mathcal{L}_\epsilon$ , if we can find a pair  $z = (\hat{x}, \hat{y})$  such that

$$\max_{y \in \mathcal{Y}} \mathcal{L}_\epsilon(\hat{x}, y) - \min_{x \in \mathcal{X}} \mathcal{L}_\epsilon(x, \hat{y}) \leq \max_{z \in \mathcal{X} \times \mathcal{Y}} V_\epsilon(\hat{z}, z) \leq \epsilon.$$

The result would imply for the original C-SC problem that

$$\begin{aligned} \max_{\mathbf{y} \in \mathcal{Y}} \mathcal{L}(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\mathbf{x}, \hat{\mathbf{y}}) &= \max_{\mathbf{y} \in \mathcal{Y}} \mathcal{L}(\hat{\mathbf{x}}, \mathbf{y}) + \frac{\epsilon}{2} \|\hat{\mathbf{x}}\|^2 - \left( \min_{\mathbf{x} \in \mathcal{X}} \mathcal{L}(\mathbf{x}, \hat{\mathbf{y}}) + \frac{\epsilon}{2} \|\hat{\mathbf{x}}\|^2 \right) \\ &\leq \max_{\mathbf{y} \in \mathcal{Y}} \left( \mathcal{L}(\hat{\mathbf{x}}, \mathbf{y}) + \frac{\epsilon}{2} \|\hat{\mathbf{x}}\|^2 \right) - \min_{\mathbf{x} \in \mathcal{X}} \left( \mathcal{L}(\mathbf{x}, \hat{\mathbf{y}}) + \frac{\epsilon}{2} \|\mathbf{x}\|^2 \right) \\ &\leq \max_{\mathbf{y} \in \mathcal{Y}} \mathcal{L}_\epsilon(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} \mathcal{L}_\epsilon(\mathbf{x}, \hat{\mathbf{y}}) \leq \epsilon. \end{aligned}$$

The left of this subsection is devoted to finding the gradient complexity of finding a  $\hat{\mathbf{z}} \in \mathcal{Z}$  such that  $\max_{\mathbf{z} \in \mathcal{Z}} V_\epsilon(\hat{\mathbf{z}}, \mathbf{z})$ . By utilizing the results in **Step 2** in the proof of Theorem 3.3 to the objective  $V_\epsilon$ , we have

$$\begin{aligned} V_\epsilon(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}) - \frac{k}{k+2} V_\epsilon(\mathbf{z}_k^{\text{ag}}, \omega_{\mathbf{z}}) &\leq \frac{1}{\eta_k(k+2)} [\|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2] \\ &\quad - \left( \frac{1}{\eta_k(k+2)} - \frac{2L}{(k+2)^2} \right) \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + \frac{2\eta_k L_H^2}{k+2} \sum_{\ell=0}^k c^{k-\ell} \|\mathbf{z}_{\ell+\frac{1}{2}} - \mathbf{z}_\ell\|^2. \end{aligned}$$

Multiplying both sides by  $(k+2)(k+1)$ , we obtain

$$\begin{aligned} (k+2)(k+1) V_\epsilon(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}) - (k+1)k V_\epsilon(\mathbf{z}_k^{\text{ag}}, \omega_{\mathbf{z}}) &\leq \frac{k+1}{\eta_k} [\|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2] \\ &\quad - \left( \frac{k+1}{\eta_k} - 2L \right) \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + 2(k+1)\eta_k L_H^2 \sum_{\ell=0}^k c^{k-\ell} \|\mathbf{z}_{\ell+\frac{1}{2}} - \mathbf{z}_\ell\|^2. \end{aligned}$$

Taking  $\eta_k = \frac{k+1}{2L + \sqrt{\frac{2}{c} L_H (k+1)}}$ ,  $c = \frac{2}{3 + \sqrt{5}}$  and adopting similar techniques as in the proof of Theorem 3.3, we have

$$(K+2)(K+1) V_\epsilon(\mathbf{z}_K^{\text{ag}}, \omega_{\mathbf{z}}) + \left( 2L + \sqrt{\frac{2}{c} L_H K} \right) \|\mathbf{z}_K - \omega_{\mathbf{z}}\|^2 \leq 2L \|\mathbf{z}_0 - \omega_{\mathbf{z}}\|^2 + \sqrt{\frac{2}{c} L_H} \sum_{k=0}^{K-1} \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2. \quad (\text{D.9})$$

Taking  $\omega_{\mathbf{z}} = \mathbf{z}_\epsilon^*$  where  $\mathbf{z}_\epsilon^*$  is the solution of the objective. Similarly as in Lemma 3.1 in the proof of Theorem 3.3, we can apply the same bootstrapping argument and derive for  $\mu_\epsilon$  being the strongly convexity parameter of  $V_\epsilon$ ,  $L_\epsilon$  being the smoothness parameter of the regularized  $F$ , the following inequality

$$\|\mathbf{z}_K^{\text{ag}} - \mathbf{z}_\epsilon^*\|^2 \leq \left( \frac{4L}{\mu_\epsilon(K+1)K} + \frac{2\sqrt{3 + \sqrt{5} L_H}}{\mu_\epsilon(K+1)} \right) \|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\|^2.$$

Applying the same restarting as in Corollary 3.4, the total number of iterates required to get within the  $\epsilon$  threshold (in terms of  $\|\mathbf{z}_K^{\text{ag}} - \mathbf{z}_\epsilon^*\|^2$ ) should be

$$\mathcal{O} \left( \sqrt{\frac{L_\epsilon}{\mu_\epsilon}} + \frac{L_H}{\mu_\epsilon} \right) \cdot \log \left( \frac{1}{\epsilon} \right).$$

We note that in previous iterates  $n = 0, \dots, N-2$  in Algorithm 2, we have obtained a  $\mathbf{z}_0$  such that  $\|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\|^2 \leq \epsilon$ . We then analyze at iteration  $n = N-1$ . Again from Equation (D.9), letting  $\omega_{\mathbf{z}} := \mathbf{z}_K^* = \arg \max_{\mathbf{z} \in \mathcal{Z}} V_\epsilon(\mathbf{z}_K^{\text{ag}}, \mathbf{z})$ , we have

$$(K+2)(K+1) V_\epsilon(\mathbf{z}_K^{\text{ag}}, \mathbf{z}_K^*) + \left( 2L + \sqrt{\frac{2}{c} L_H K} \right) \|\mathbf{z}_K - \mathbf{z}_K^*\|^2 \leq 2L \|\mathbf{z}_0 - \mathbf{z}_K^*\|^2 + \sqrt{\frac{2}{c} L_H} \sum_{k=0}^{K-1} \|\mathbf{z}_k - \mathbf{z}_K^*\|^2.$$

As  $V_\epsilon(\mathbf{z}_K^{\text{ag}}, \mathbf{z}_K^*) \geq 0$ , we can apply the same bootstrapping argument and derive

$$\frac{\mu_\epsilon}{2} \|\mathbf{z}_K^{\text{ag}} - \mathbf{z}_K^*\|^2 \leq V_\epsilon(\mathbf{z}_K^{\text{ag}}, \mathbf{z}_K^*) \leq \frac{2L_\epsilon + \sqrt{3 + \sqrt{5} L_H K}}{K(K+1)} \|\mathbf{z}_0 - \mathbf{z}_K^*\|^2. \quad (\text{D.10})$$



On the other hand, we also have that

$$\frac{\mu_\epsilon}{2} \|\mathbf{z}_K^{\text{ag}} - \mathbf{z}_\epsilon^*\|^2 \leq V_\epsilon(\mathbf{z}_K^{\text{ag}}, \mathbf{z}_\epsilon^*) \leq \frac{2L + \sqrt{3 + \sqrt{5}L_H K}}{K(K+1)} \|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\|^2. \quad (\text{D.11})$$

By analyzing the two inequalities (D.10) and (D.11), we obtain that for sufficiently large  $K$  (in an order of  $\mathcal{O}\left(\sqrt{\frac{L_\epsilon}{\mu_\epsilon}} + \frac{L_H}{\mu_\epsilon}\right)$ ) such that  $\frac{4L+2\sqrt{3+\sqrt{5}L_H K}}{\mu_\epsilon K(K+1)} \leq \frac{1}{c_2}$ ,

$$\begin{aligned} \|\mathbf{z}_K^* - \mathbf{z}_\epsilon^*\| &\leq \|\mathbf{z}_K^{\text{ag}} - \mathbf{z}_K^*\| + \|\mathbf{z}_K^{\text{ag}} - \mathbf{z}_\epsilon^*\| \leq \frac{1}{c_2} [\|\mathbf{z}_0 - \mathbf{z}_K^*\| + \|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\|] \\ &\leq \frac{1}{c_2} [\|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\| + \|\mathbf{z}_\epsilon^* - \mathbf{z}_K^*\| + \|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\|] \leq \frac{2}{c_2 - 1} \|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\|. \end{aligned}$$

Furthermore, we apply (D.10) again and derive

$$\max_{\mathbf{z} \in \mathcal{Z}} V_\epsilon(\mathbf{z}_K^{\text{ag}}, \mathbf{z}) = V_\epsilon(\mathbf{z}_K^{\text{ag}}, \mathbf{z}_K^*) \leq \frac{1}{c_2} \|\mathbf{z}_0 - \mathbf{z}_K^*\|^2 \leq \frac{\|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\|}{c_2} \leq \frac{\|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\| + \|\mathbf{z}_\epsilon^* - \mathbf{z}_K^*\|}{c_2} \leq \frac{c_2 + 1}{c_2(c_2 - 1)} \|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\|.$$

Taking  $c_2 = 3$ , and noting that we have obtained  $\|\mathbf{z}_0 - \mathbf{z}_\epsilon^*\|^2 \leq \epsilon$  in previous restarted iterates and combining the technique at the beginning of the proof of Theorem 3.5, the total number of iterates in order to get  $\max_{\mathbf{y} \in \mathcal{Y}} \mathcal{L}_\epsilon(\hat{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} \mathcal{L}_\epsilon(\mathbf{x}, \hat{\mathbf{y}}) \leq \max_{\mathbf{z} \in \mathcal{Z}} V_\epsilon(\hat{\mathbf{z}}, \mathbf{z}) \leq \epsilon$  is  $\mathcal{O}\left(\sqrt{\frac{L}{\epsilon \wedge \mu_g}} + \frac{L_H}{\epsilon \wedge \mu_g}\right) \cdot \log\left(\frac{1}{\epsilon}\right)$ . Applying a scaling reduction argument as in (3.8) (the stepsize is dependent on  $\epsilon$  after scaling reduction) gives a final complexity of

$$\mathcal{O}\left(\left(\sqrt{\frac{L_f}{\epsilon}} \vee \frac{L_g}{\mu_g} + \frac{I_{xx}}{\epsilon} \vee \frac{I_{xy}}{\sqrt{\epsilon} \mu_g} \vee \frac{I_{yy}}{\mu_g}\right) \log\left(\frac{1}{\epsilon}\right)\right).$$

□

#### D.4 Proof of Theorem 4.1

*Proof.*[Proof of Theorem 4.1] For the stochastic case, we use the primal-dual gap function (D.1) and proceeds in the following

**Step 1: Estimating weighted temporal difference in squared norms.** We mentioned in the proof of Theorem 3.3 that Lemma D.1 holds for the stochastic case as well. Thus, we have

$$V(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}) - (1 - \alpha_k)V(\mathbf{z}_k^{\text{ag}}, \omega_{\mathbf{z}}) \leq \underbrace{\alpha_k \langle \nabla F(\mathbf{z}_k^{\text{md}}) + H(\mathbf{z}_{k+\frac{1}{2}}), \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \rangle}_{\text{I}} + \underbrace{\frac{L\alpha_k^2}{2} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2}_{\text{II}}. \quad (\text{D.2})$$

By applying Proposition D.2 to the iterates of Algorithm 3. Taking  $\mathbf{x} = \mathbf{z}_k$ ,  $\phi_1 = \mathbf{z}_{k+\frac{1}{2}}$ ,  $\phi_2 = \mathbf{z}_{k+1}$  in Proposition D.2 and recalling the update rules in Algorithm 3, we obtain the following stochastic version of inequality (D.5):

$$\begin{aligned} &\eta_k \cdot \langle \nabla \tilde{F}(\mathbf{z}_k^{\text{md}}; \xi_k) + \nabla \tilde{H}(\mathbf{z}_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}), \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \rangle \\ &\leq \frac{1}{2} \eta_k^2 \underbrace{\|\tilde{H}(\mathbf{z}_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}) - \tilde{H}(\mathbf{z}_{k-\frac{1}{2}}; \zeta_{k-\frac{1}{2}})\|^2}_{(a)} + \frac{1}{2} \left[ \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 \right]. \end{aligned}$$

**Step 2: Building and solving the recursion.** Note that in the stochastic case, unlike Step 2 in the proof of Theorem 3.3, before connecting  $\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2$  with  $\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2$  to get an iterative rule, we need to bound the expectation of (a) with additional noise first.

Throughout the rest of the proof of Theorem 4.1, we denote

$$\Delta_h^{k+\frac{1}{2}} = \tilde{H}(\mathbf{z}_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}) - H(\mathbf{z}_{k+\frac{1}{2}}), \quad \Delta_f^k = \nabla \tilde{F}(\mathbf{z}_k^{\text{md}}; \xi_k) - \nabla F(\mathbf{z}_k^{\text{md}}).$$

Taking expectation over term (a) in above, we use the following lemma to depict the upper bound of the quantity. The proof is delayed to §G.7.

**Lemma D.5.** For any  $\beta > 0$ , under Assumption 2.2, we have

$$\mathbb{E} \|\tilde{H}(\mathbf{z}_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}) - \tilde{H}(\mathbf{z}_{k-\frac{1}{2}}; \zeta_{k-\frac{1}{2}})\|^2 \leq (1 + \beta) L_H^2 \mathbb{E} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 + \left(2 + \frac{1}{\beta}\right) \sigma_H^2. \quad (\text{D.12})$$

Taking  $\beta = 1$  in Lemma D.5 and bringing the result into the expectation of (D.2), we obtain that

$$\begin{aligned} & \mathbb{E} V(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}) - (1 - \alpha_k) \mathbb{E} V(\mathbf{z}_k^{\text{ag}}, \omega_{\mathbf{z}}) \\ & \leq \frac{\alpha_k \eta_k}{2} \left[ 2L_H^2 \mathbb{E} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 + 3\sigma_H^2 \right] + \alpha_k \mathbb{E} \left\langle \Delta_f^k + \Delta_h^{k+\frac{1}{2}}, \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \right\rangle \\ & \quad + \frac{L\alpha_k^2}{2} \mathbb{E} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + \frac{\alpha_k}{2\eta_k} \mathbb{E} \left[ \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 \right]. \end{aligned} \quad (\text{D.13})$$

Following the above inequality and following similar techniques as in Step 2 of the proof of Theorem 3.3, we can derive the following Lemma D.6, whose proof is delayed to §G.8.

**Lemma D.6.** For the choice of stepsize such that  $\eta_k L_H \leq \frac{\sqrt{c}}{2}$  holds for all  $k$  and any constant  $r > 0$ , we have

$$\begin{aligned} & \mathbb{E} V(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}) - (1 - \alpha_k) \mathbb{E} V(\mathbf{z}_k^{\text{ag}}, \omega_{\mathbf{z}}) \leq \frac{\alpha_k}{2\eta_k} \mathbb{E} [\|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2] + \frac{3\alpha_k \eta_k}{2(1-c)} \sigma_H^2 \\ & \quad + 2\alpha_k \eta_k L_H^2 \sum_{\ell=0}^k c^{k-\ell} \mathbb{E} \|\mathbf{z}_{\ell+\frac{1}{2}} - \mathbf{z}_\ell\|^2 - \left( \frac{r\alpha_k}{2\eta_k} - \frac{L\alpha_k^2}{2} \right) \mathbb{E} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + \frac{\alpha_k \eta_k}{2(1-r)} \sigma_F^2. \end{aligned}$$

Recalling that  $\alpha_k = \frac{2}{k+2}$ , we have

$$\begin{aligned} & \mathbb{E} V(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}) - \frac{k}{k+2} \mathbb{E} V(\mathbf{z}_k^{\text{ag}}, \omega_{\mathbf{z}}) \leq \frac{1}{\eta_k(k+2)} \mathbb{E} [\|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2] \\ & \quad + \frac{4\eta_k L_H^2}{k+2} \sum_{\ell=0}^k c^{k-\ell} \mathbb{E} \|\mathbf{z}_{\ell+\frac{1}{2}} - \mathbf{z}_\ell\|^2 - \left( \frac{r}{\eta_k(k+2)} - \frac{2L}{(k+2)^2} \right) \mathbb{E} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 \\ & \quad + \frac{3\eta_k}{(1-c)(k+2)} \sigma_H^2 + \frac{\eta_k}{(1-r)(k+2)} \sigma_F^2. \end{aligned}$$

Multiplying both sides by  $(k+2)^2$  and taking  $r = \frac{1}{2}$ , we obtain

$$\begin{aligned} & (k+2)^2 \mathbb{E} V(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}) - [(k+1)^2 - 1] \mathbb{E} V(\mathbf{z}_k^{\text{ag}}, \omega_{\mathbf{z}}) \\ & \leq \frac{k+2}{\eta_k} \mathbb{E} [\|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2] + 4\eta_k L_H^2 (k+2) \sum_{\ell=0}^k c^{k-\ell} \mathbb{E} \|\mathbf{z}_{\ell+\frac{1}{2}} - \mathbf{z}_\ell\|^2 \\ & \quad - \left( \frac{r(k+2)}{\eta_k} - 2L \right) \mathbb{E} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + \frac{3\eta_k(k+2)}{1-c} \sigma_H^2 + \frac{\eta_k(k+2)}{1-r} \sigma_F^2 \\ & \leq \frac{k+2}{\eta_k} \mathbb{E} [\|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2] + 4\eta_k L_H^2 (k+2) \sum_{\ell=0}^k c^{k-\ell} \mathbb{E} \|\mathbf{z}_{\ell+\frac{1}{2}} - \mathbf{z}_\ell\|^2 \\ & \quad - \left( \frac{k+2}{2\eta_k} - 2L \right) \mathbb{E} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + \frac{3\eta_k(k+2)}{1-c} \sigma_H^2 + 2\eta_k(k+2) \sigma_F^2. \end{aligned}$$

Telescoping over  $k = 0, 1, \dots, K-1$  and using the same techniques as in the proof of Theorem 3.3, we have for  $\frac{k+2}{2\eta_k} \geq 2L + \frac{1}{\sqrt{c}} L_H (k+2)$  and  $c = \frac{1}{2+\sqrt{2}}$  ( $c/(1-c) = \sqrt{2} - 1$ , and recall  $\sigma^2 = 3\sqrt{2}\sigma_H^2 + 2\sigma_F^2$ ) so that

$$\begin{aligned} & [(K+1)^2 - 1] \mathbb{E} V(\mathbf{z}_K^{\text{ag}}, \mathbf{z}^*) + \frac{K+1}{\eta_{K-1}} \mathbb{E} \|\mathbf{z}_K - \mathbf{z}^*\|^2 \\ & \leq \frac{2}{\eta_0} \mathbb{E} \|\mathbf{z}_0 - \mathbf{z}^*\|^2 + \frac{2}{\sqrt{c}} L_H \sum_{k=1}^{K-1} \mathbb{E} \|\mathbf{z}_k - \mathbf{z}^*\|^2 + \sum_{k=0}^{K-1} (k+2) \eta_k \sigma^2 - \sum_{k=0}^{K-1} \mathbb{E} V(\mathbf{z}_{k+1}^{\text{ag}}, \mathbf{z}^*). \end{aligned} \quad (\text{D.14})$$

**Step 3: Proving  $z_k$  stays within a neighbourhood of  $z^*$ .** We introduce the following Lemma D.7, whose proof is in §G.4

**Lemma D.7.** *Given the maximum epoch number  $K > 0$  and stepsize sequence  $\{\eta_k\}_{k \in [K]}$  satisfying*

(a)  $\frac{k+2}{\eta_k} - \frac{k+1}{\eta_{k-1}} = \frac{2}{\sqrt{c}} L_H$  for any  $k < K$ , we have for  $\forall k \in [K-1]$ :

$$\|z_k - z^*\|^2 \leq \|z_0 - z^*\|^2 + \frac{\eta_0}{2} \sum_{k=0}^{K-1} (k+2) \eta_k \sigma^2.$$

(b) In addition if  $\eta_k \leq \frac{k+2}{D}$  for  $\forall k \in [K-1]$  where  $D$  will be specified in (c) and taking  $A(K) := \sqrt{(K+1)(K+2)(2K+3)/6}$ , we have

$$\|z_k - z^*\|^2 \leq \|z_0 - z^*\|^2 + \frac{A(K)^2 \sigma^2}{D^2}. \quad (\text{D.15})$$

(c) Taking  $D = \frac{\sigma}{C} \frac{A(K)}{\sqrt{\mathbb{E}\|z_0 - z^*\|^2}}$  for some absolute constant  $C > 0$ , bound (D.15) reduces to

$$\|z_k - z^*\|^2 \leq (1 + C^2) \|z_0 - z^*\|^2. \quad (\text{D.16})$$

**Step 4: Combining everything together.** Combining the choice of stepsize  $\eta_k$  in (a), (b) in Lemma D.7 and  $\frac{k+2}{2\eta_k} \geq 2L + \frac{1}{\sqrt{c}} L_H (k+2)$ , and bound (D.14) with Eq. (D.16), by rearranging the terms again, we conclude that for  $\eta_k = \frac{k+2}{4L+D+4\sqrt{2+\sqrt{2}}L_H(k+2)}$ ,

$$\begin{aligned} (K+1)^2 \mathbb{E}V(z_K^{\text{ag}}, z^*) &\leq \left( 4L + 2\sqrt{2+\sqrt{2}}(K+1)(1+C^2)L_H \right) \mathbb{E}\|z_0 - z^*\|^2 \\ &\quad + \left( C + \frac{1}{C} \right) \sigma A(K) \sqrt{\mathbb{E}\|z_0 - z^*\|^2}. \end{aligned}$$

Dividing both sides by  $(K+1)^2$  and noting that  $V(z_K^{\text{ag}}, z^*) \geq \frac{\mu}{2} \mathbb{E}\|z_K^{\text{ag}} - z^*\|^2$ , we have

$$\mathbb{E}\|z_K^{\text{ag}} - z^*\|^2 \leq \left[ \frac{8L}{\mu(K+1)^2} + \frac{7.4(1+C^2)L_H}{\mu(K+1)} \right] \mathbb{E}\|z_0 - z^*\|^2 + \frac{2(C + \frac{1}{C})\sigma}{\mu\sqrt{K+1}} \sqrt{\mathbb{E}\|z_0 - z^*\|^2},$$

hence concluding the entire proof of Theorem 4.1. □

## E Understanding the Continuous-Time Dynamics of AG-OG

Due to the complexity of discrete analysis, in §B we describe the continuous dynamics of AG-OG as a hybrid dynamics of OGDA and Nesterov's acceleration. After introducing the discrete dynamics of AG-OG in §3.3, we are now ready to present a formal proposition that connects Algorithm 1 with its ODE. We adopt similar stepsize scheme as in Theorem 3.3 by letting

$\eta_k = \frac{k+1}{2L' + \sqrt{3+\sqrt{3}}L'_H(k+1)}$  where  $L'$  and  $L'_H$  can be arbitrarily large. Furthermore, we set  $\Delta t = \frac{1}{\sqrt{L' \vee L'_H}}$ , let  $\frac{L'_H}{L'} \rightarrow c$  and

use the notation  $c_t = \frac{t(1 \vee c)^2}{2 + \sqrt{3+\sqrt{3}}c(1 \vee c)t}$ . We derive the following proposition:

**Proposition E.1** (Proposition B.1). *The ODE for the AG-OG Algorithm 1 is*

$$\frac{t}{2} \ddot{Z} + \frac{3}{2} \dot{Z} + c_t \left( \nabla F(Z) + HZ + \frac{t}{2} H \dot{Z} \right) = 0. \quad (\text{E.1})$$

*Proof.* [Proof of Proposition B.1] We recall that in Algorithm 1, Line 3 and 5 yield:

$$\begin{aligned} z_{k+\frac{1}{2}} &= z_k - \eta_k \left( H(z_{k-\frac{1}{2}}) + \nabla F(z_k^{\text{md}}) \right), \\ z_{k+1} &= z_k - \eta_k \left( H(z_{k+\frac{1}{2}}) + \nabla F(z_k^{\text{md}}) \right). \end{aligned}$$

Subtracting the first line from the second line, we have

$$\mathbf{z}_{k+1} = \mathbf{z}_{k+\frac{1}{2}} - \eta_k \left( H(\mathbf{z}_{k+\frac{1}{2}}) - H(\mathbf{z}_{k-\frac{1}{2}}) \right). \quad (\text{E.2})$$

Furthermore, by shifting indices, we have

$$\mathbf{z}_{k+\frac{3}{2}} = \mathbf{z}_{k+1} - \eta_{k+1} \left( H(\mathbf{z}_{k+\frac{1}{2}}) + \nabla F(\mathbf{z}_{k+1}^{\text{md}}) \right). \quad (\text{E.3})$$

Combining (E.2) and (E.3), we derive the iterative update rule on  $\mathbf{z}_{k+\frac{1}{2}}$  that

$$\begin{aligned} \mathbf{z}_{k+\frac{3}{2}} &= \mathbf{z}_{k+\frac{1}{2}} - (\eta_k + \eta_{k+1}) H(\mathbf{z}_{k+\frac{1}{2}}) + \eta_k H(\mathbf{z}_{k-\frac{1}{2}}) - \eta_{k+1} \nabla F(\mathbf{z}_{k+1}^{\text{md}}) \\ &= \mathbf{z}_{k+\frac{1}{2}} - \eta_{k+1} H(\mathbf{z}_{k+\frac{1}{2}}) - \eta_{k+1} \nabla F(\mathbf{z}_{k+1}^{\text{md}}) - \eta_k \left( H(\mathbf{z}_{k+\frac{1}{2}}) - H(\mathbf{z}_{k-\frac{1}{2}}) \right). \end{aligned}$$

Moreover, Line 2 and 4 in Algorithm 1 imply that

$$\mathbf{z}_{k+1}^{\text{md}} = \mathbf{z}_{k+1}^{\text{ag}} + \frac{k}{k+3} (\mathbf{z}_{k+1}^{\text{ag}} - \mathbf{z}_k^{\text{ag}}).$$

Thus, we obtain

$$\begin{aligned} \mathbf{z}_{k+\frac{3}{2}} - \mathbf{z}_{k+\frac{1}{2}} &= -\eta_{k+1} \left[ H(\mathbf{z}_{k+\frac{1}{2}}) + \nabla F(\mathbf{z}_{k+1}^{\text{ag}}) \right] - \eta_k \left[ H(\mathbf{z}_{k+\frac{1}{2}}) - H(\mathbf{z}_{k-\frac{1}{2}}) \right] \\ &\quad - \eta_{k+1} \left[ \nabla F(\mathbf{z}_{k+1}^{\text{md}}) - \nabla F(\mathbf{z}_{k+1}^{\text{ag}}) \right]. \end{aligned} \quad (\text{E.4})$$

Recalling that in Theorem 3.3 we take the stepsize  $\eta_k = \frac{k+2}{2L+\sqrt{3+\sqrt{3}L_H}(k+2)}$  and noting that the stepsize can actually be chosen for any  $L' > L, L'_H > L_H$  as  $\eta_k = \frac{k+2}{2L'+\sqrt{3+\sqrt{3}L'_H}(k+2)}$ . We let  $t = k\Delta t$  where  $\Delta t = \frac{1}{\sqrt{L'}\sqrt{L'_H}}$ . Thus, we obtain

$$\eta_k = \frac{k+2}{2L' + \sqrt{3+\sqrt{3}L'_H}(k+2)} = \frac{t\Delta t + 2(\Delta t)^2}{2L'(\Delta t)^2 + \sqrt{3+\sqrt{3}L'_H}(t\Delta t + 2(\Delta t)^2)}.$$

We use  $c = \frac{L'_H}{\sqrt{L'}}$  to depict the relationship between  $L'_H$  and  $L'$ , simple algebra yields:

$$L'_H = \frac{c}{(1 \vee c)\Delta t}, \quad L' = \frac{1}{(1 \vee c)^2(\Delta t)^2}.$$

Combining this with the value of  $\eta_k$  and we obtain that

$$\begin{aligned} \eta_k &= \frac{t\Delta t + 2(\Delta t)^2}{\frac{2}{(1 \vee c)^2} + \sqrt{3+\sqrt{3}\frac{c}{1 \vee c}}(t+2\Delta t)} = \frac{(1 \vee c)^2(t+2\Delta t)}{2 + \sqrt{3+\sqrt{3}c(1 \vee c)}(t+2\Delta t)} \Delta t \\ &= \frac{(1 \vee c)^2 t}{2 + \sqrt{3+\sqrt{3}c(1 \vee c)}t} \Delta t + (\Delta t)^2, \end{aligned} \quad (\text{E.5})$$

which goes to 0 as  $L', L'_H \rightarrow \infty$  and  $\Delta t \rightarrow 0$ . We let  $c_t = \frac{(1 \vee c)^2 t}{2 + \sqrt{3+\sqrt{3}c(1 \vee c)}t}$  and (E.5) can be shortened as:

$$\eta_k = c_t \Delta t + o(\Delta t).$$

With this choice of  $\eta_k$ , by Taylor expansion we have:

$$\begin{aligned} H(\mathbf{z}_{k+\frac{1}{2}}) - H(\mathbf{z}_{k-\frac{1}{2}}) &\leq \mathcal{O}(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}) \leq \mathcal{O}(\eta_k) = o(1), \\ \nabla F(\mathbf{z}_{k+1}^{\text{md}}) - \nabla F(\mathbf{z}_{k+1}^{\text{ag}}) &\leq \mathcal{O}(\mathbf{z}_{k+1}^{\text{md}} - \mathbf{z}_{k+1}^{\text{ag}}) \\ &\stackrel{(a)}{\leq} \mathcal{O}(\mathbf{z}_{k+1}^{\text{ag}} - \mathbf{z}_k^{\text{ag}}) \leq \mathcal{O}\left(\frac{2}{k+2}(\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k^{\text{ag}})\right) = o(1). \end{aligned}$$



we rewrite (E.4) in continuous dynamics by letting  $\widehat{Z}(t) = z_{k-\frac{1}{2}}$  and  $Z(t) = z_{k+1}^{\text{ag}}$ :

$$\begin{aligned}\widehat{Z}(t+1) - \widehat{Z}(t) &= -\eta_{k+1} \left[ H(\widehat{Z}(t)) + \nabla F(Z(t)) \right] + o(\eta_k) + o(\eta_{k+1}) \\ &= -c_t \Delta t \left[ H(\widehat{Z}(t)) + \nabla F(Z(t)) \right] + o(\Delta t).\end{aligned}$$

Dividing both sides by  $\Delta t$  and with  $\Delta t \rightarrow 0$ , we obtain

$$\dot{\widehat{Z}}(t) + c_t \left[ H(\widehat{Z}(t)) + \nabla F(Z(t)) \right] = 0. \quad (\text{E.6})$$

In the final step we calculate the relationship between  $\widehat{Z}(t)$  and  $Z(t)$ . By Line 4, we have

$$z_{k+1} = \frac{k}{2} \left( z_{k+\frac{1}{2}}^{\text{ag}} - z_k^{\text{ag}} \right) + z_{k+1}^{\text{ag}},$$

which is equivalent to

$$\widehat{Z}(t) = \frac{t}{2} \frac{Z(t) - Z(t-1)}{\Delta t} + Z(t).$$

Letting  $\Delta t \rightarrow 0$ , we have

$$\widehat{Z}(t) = Z(t) + \frac{t}{2} \dot{Z}(t). \quad (\text{E.7})$$

Combining (E.6) with (E.7) we conclude our proof.  $\square$

### E.1 Proof of Theorem B.2

*Proof.*[Proof of Theorem B.2] We first provide an estimate of the time derivative  $\dot{\mathcal{E}}$  of the Lyapunov function corresponding to (B.4), and the result is shown in Lemma E.2:

**Lemma E.2.** *We set the Lyapunov function as defined in the following (E.8):*

$$\mathcal{E} = t^2 V(Z) + \frac{t}{c_t} \|\widehat{Z} - z^*\|^2, \quad (\text{E.8})$$

where  $c_t = \frac{(1 \vee c)^2 t}{2 + \sqrt{3 + \sqrt{3}c}(1 \vee c)t}$ . Given the dynamics in (B.4) starting from  $Z(0) = z_0$ , we have

$$\dot{\mathcal{E}} = \frac{d}{dt} \left[ t^2 V(Z) + \frac{t}{c_t} \|\widehat{Z} - z^*\|^2 \right] \leq \sqrt{3 + \sqrt{3}} \frac{c}{1 \vee c} \|\widehat{Z} - z^*\|^2. \quad (\text{E.9})$$

We postpone the proof of Lemma E.2 to §G.1. Note that both sides of (E.9) in Lemma E.2 presents the quantity  $\frac{t}{c_t} \|\widehat{Z} - z^*\|^2$  in its original and gradient forms, respectively. By integrating on both sides and applying a Gronwall-type technique, we obtain the following Lemma E.3 which shows that the continuous-time dynamics of AG-EG-ODE are non-expansive with respect to the minimax point  $z^*$ .

**Lemma E.3.** *We have*

$$\|\widehat{Z} - z^*\| \leq \|z_0 - z^*\|. \quad (\text{E.10})$$

We postpone the rigorous proof of Lemma E.3 to §G.2. Now bringing (E.10) in Lemma E.3 into (E.9), we conclude that

$$\dot{\mathcal{E}} = \frac{d}{dt} \left[ t^2 V(Z) + \frac{t}{c_t} \|\widehat{Z} - z^*\|^2 \right] \leq \sqrt{3 + \sqrt{3}} \frac{c}{1 \vee c} \|z_0 - z^*\|^2.$$

Integrating both sides gives

$$T^2 V(Z) + \frac{2 + \sqrt{3 + \sqrt{3}c}(1 \vee c)T}{(1 \vee c)^2} \|\widehat{Z} - z^*\|^2 - \frac{2}{(1 \vee c)^2} \|z_0 - z^*\|^2 \leq \sqrt{3 + \sqrt{3}} \frac{c}{1 \vee c} T \|z_0 - z^*\|^2.$$

Rearranging and dividing both sides by  $T^2$ , we obtain that

$$V(Z) \leq V(Z) + \frac{2 + \sqrt{3 + \sqrt{3}c}(1 \vee c)T}{(1 \vee c)^2 T^2} \|\hat{Z} - z^*\|^2 \leq \frac{2 + \sqrt{3 + \sqrt{3}c}(1 \vee c)T}{(1 \vee c)^2 T^2} \|z_0 - z^*\|^2,$$

which concludes our proof.  $\square$

## F Proof of Bilinear Game Cases

### F.1 Proof of Theorem 3.6

*Proof.* [Proof of Theorem 3.6]

**Step 1: Non-expansiveness of OGD last-iterate.** We start by the non-expansiveness Lemma, whose proof is in §G.9.

**Lemma F.1** (Bounded Iterates). *Following (3.11), at any iterate  $k < K$ ,  $z_k$  stays within the region defined by the initialization  $z_0$ :*

$$\|z_k - z^*\| \leq \|z_0 - z^*\|,$$

where we recall that  $z^* = [x^*; y^*]$  denotes the unique solution of Problem (1.3) with  $\nabla f, \nabla g = 0$  and  $I$  defined in (3.10).

By Lemma F.1, for any  $0 \leq k < K$ , we have

$$\|z_k - z^*\| \leq \|z_0 - z^*\|.$$

**Step 2:** Recalling that we take  $\alpha_k = \frac{2}{k+2}$  in (3.11b) of (3.11). Thus, we obtain the following

$$z_{k+1}^{\text{ag}} = \frac{k}{k+2} z_k^{\text{ag}} + \frac{2}{k+2} z_{k+\frac{1}{2}}.$$

Subtracting both sides by  $z^*$  and multiplying both sides by  $(k+1)(k+2)$ , we have

$$(k+1)(k+2) (z_{k+1}^{\text{ag}} - z^*) = k(k+1) (z_k^{\text{ag}} - z^*) + 2(k+1) (z_{k+\frac{1}{2}} - z^*).$$

Telescoping over  $k = 0, \dots, K-1$  and we conclude that

$$K(K+1) (z_K^{\text{ag}} - z^*) = 2 \sum_{k=0}^{K-1} (k+1) (z_{k+\frac{1}{2}} - z^*). \quad (\text{F.1})$$

Moreover, according to the update rule (3.11c), we have that

$$Kz_K - \sum_{k=0}^{K-1} z_k = \sum_{k=0}^{K-1} (k+1)z_{k+1} - (k+1)z_k = \sum_{k=0}^{K-1} \eta_k (k+1) H(z_{k+\frac{1}{2}}). \quad (\text{F.2})$$

Recalling that  $\eta_k = \frac{1}{2L_H}$ , combining (F.2) with (F.1) and taking the squared norm on both sides, we conclude that

$$\begin{aligned} \|K(K+1) (z_K^{\text{ag}} - z^*)\|^2 &= \left\| 2 \sum_{k=0}^{K-1} (k+1) (z_{k+\frac{1}{2}} - z^*) \right\|^2 \leq \frac{1}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})} \left\| 2 \sum_{k=0}^{K-1} (k+1) H(z_{k+\frac{1}{2}}) \right\|^2 \\ &= \frac{16L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})} \|Kz_K - \sum_{k=0}^{K-1} z_k\|^2 = \frac{16L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})} \left\| \sum_{k=0}^{K-1} [(z_K - z^*) - (z_k - z^*)] \right\|^2 \\ &\leq \frac{16L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})} K \cdot \sum_{k=0}^{K-1} [2\|z_K - z^*\|^2 + 2\|z_k - z^*\|^2]. \end{aligned} \quad (\text{F.3})$$

Applying non-expansiveness in Lemma F.1 in (F.10), bringing  $L_H = \sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}$  and rearranging, we conclude that

$$\|z_K^{\text{ag}} - z^*\|^2 \leq \frac{64\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^2} \|z_0 - z^*\|^2.$$

Restarting every  $\left\lceil 8\sqrt{\frac{e\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})}} \right\rceil$  iterates for a total of  $\log\left(\frac{\|z_0 - z^*\|}{\epsilon}\right)$  times, we obtain the final sample complexity of

$$\mathcal{O}\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})}} \log\left(\frac{1}{\epsilon}\right)\right)$$

for the nonstochastic setting.  $\square$

## F.2 Stochastic Bilinear Game Case

For the stochastic AGOG with restarting for bilinear case, our iteration spells

$$\begin{cases} z_{k+\frac{1}{2}} &= z_k - \eta_k \tilde{H}(z_{k-\frac{1}{2}}; \zeta_{k-\frac{1}{2}}), \end{cases} \quad (\text{F.4a})$$

$$\begin{cases} z_{k+1}^{\text{ag}} &= (1 - \alpha_k) z_k^{\text{ag}} + \alpha_k z_{k+\frac{1}{2}}, \end{cases} \quad (\text{F.4b})$$

$$\begin{cases} z_{k+1} &= z_k - \eta_k \tilde{H}(z_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}). \end{cases} \quad (\text{F.4c})$$

We are able to derive the following theorem.

**Theorem F.2** (Convergence of stochastic AG-OG, bilinear case). *When specified to the stochastic bilinear game case, setting the parameters as  $\alpha_k = \frac{2}{k+2}$  and  $\eta_k = \frac{1}{2L_H}$ , the output of update rules (F.4) satisfies for any  $\gamma > 0$ ,*

$$\mathbb{E}\|z_K^{\text{ag}} - z^*\|^2 \leq (1 + \gamma) \left( \frac{128L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^2} \|z_0 - z^*\|^2 + \frac{48}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)} \sigma_H^2 \right) + \frac{4(1 + \frac{1}{\gamma})\sigma_H^2}{3\lambda_{\min}(\mathbf{B}^\top \mathbf{B})K}.$$

Moreover, by operating scheduled restarting technique, it yields a complexity result of

$$\mathcal{O}\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})}} \log\left(\frac{\sqrt[4]{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})} \lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\sigma_H}\right) + \frac{\sigma_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B}) \epsilon^2}\right)$$

We begin by proving the following lemma, which is the stochastic version of Lemma F.1. It is worth noting that when setting  $\sigma_H = 0$  and  $\beta = 0$  in Lemma F.3 below, it reduces to the non-expansiveness of deterministic iterates (3.11). Moreover, Lemma F.3 holds for any monotonic  $H(\cdot)$ .

**Lemma F.3** (Bounded Iterates (Stochastic)). *Following (F.4), for any  $\beta > 0$ , taking  $\eta_k = \frac{1}{2L_H\sqrt{(1+\beta)}}$  at any iterate  $k < K$ ,  $z_k$  stays within the region defined by the initialization  $z_0$ :*

$$\|z_k - z^*\|^2 \leq \|z_0 - z^*\|^2 + \eta_k^2 \left(2 + \frac{1}{\beta}\right) K \sigma_H^2,$$

where we recall that  $z^*$  denotes the unique solution of Problem (1.2) with  $\nabla f, \nabla g = 0$  and  $I$  defined in (3.10).

*Proof.*[Proof of Lemma F.3] The optimal condition of the problem yields  $H(z^*) = 0$  for  $z^*$  being the solution of the VI. By the monotonicity of  $H(\cdot)$ , let  $z = z_{k+\frac{1}{2}}$  and  $z' = \omega_z$  in (2.2), we have that

$$\langle H(z_{k+\frac{1}{2}}) - H(\omega_z), z_{k+\frac{1}{2}} - \omega_z \rangle \geq 0, \forall \omega_z \in \mathcal{Z}. \quad (\text{F.5})$$

Let  $\varphi_1 = z_{k+\frac{1}{2}}$ ,  $\varphi_2 = z_{k+1}$ ,  $\theta = z_k$ ,  $\delta_1 = \eta_k \tilde{H}(z_{k-\frac{1}{2}}; \zeta_{k-\frac{1}{2}})$ ,  $\delta_2 = \eta_k \tilde{H}(z_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}})$  and  $z = \omega_z$  in Proposition D.2, we have

$$\begin{aligned} & \eta_k \langle \tilde{H}(z_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}), z_{k+\frac{1}{2}} - \omega_z \rangle \\ & \leq \frac{\eta_k^2}{2} \|\tilde{H}(z_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}) - \tilde{H}(z_{k-\frac{1}{2}}; \zeta_{k-\frac{1}{2}})\|^2 + \frac{1}{2} [\|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2 - \|z_k - z_{k+\frac{1}{2}}\|^2]. \end{aligned} \quad (\text{F.6})$$

Recalling the results in Lemma D.5 that

$$\mathbb{E} \|\tilde{H}(\mathbf{z}_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}) - \tilde{H}(\mathbf{z}_{k-\frac{1}{2}}; \zeta_{k-\frac{1}{2}})\|^2 \leq (1+\beta)L_H^2 \mathbb{E} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 + \left(2 + \frac{1}{\beta}\right) \sigma_H^2.$$

Taking expectation over (F.6), combining it with (F.5) and letting  $\omega_{\mathbf{z}} = \mathbf{z}^*$ , we obtain

$$\begin{aligned} 0 &= \eta_k \mathbb{E} \left\langle H(\mathbf{z}^*), \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}^* \right\rangle \\ &\leq \frac{\eta_k^2 (1+\beta) L_H^2}{2} \mathbb{E} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 + \frac{1}{2} \mathbb{E} \left[ \|\mathbf{z}_k - \mathbf{z}^*\|^2 - \|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2 - \|\mathbf{z}_k - \mathbf{z}_{k+\frac{1}{2}}\|^2 \right] + \frac{\eta_k^2 \left(2 + \frac{1}{\beta}\right)}{2} \sigma_H^2. \end{aligned} \quad (\text{F.7})$$

Next, we move on to estimate  $\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2$ . As we know that via Young's and Cauchy-Schwarz's inequalities and the update rules (3.11a) and (3.11c), for all  $k \geq 1$

$$\begin{aligned} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 &\leq 2\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + 2\|\mathbf{z}_k - \mathbf{z}_{k-\frac{1}{2}}\|^2 \\ &\leq 2\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + 2\eta_{k-1}^2 L_H^2 \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2. \end{aligned}$$

Multiplying both sides by 2 and moving one term to the right hand gives for all  $k \geq 1$

$$\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \leq 4\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + 4\eta_{k-1}^2 L_H^2 \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2 - \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2.$$

Bringing this into (G.12) and noting that  $\eta_{k-1} \leq \frac{1}{2L_H}$  as well as  $\eta_k \leq \frac{1}{2L_H}$ , we have

$$\begin{aligned} 0 &\leq \frac{\eta_k^2 (1+\beta) L_H^2}{2} \mathbb{E} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 + \frac{1}{2} \mathbb{E} \left[ \|\mathbf{z}_k - \mathbf{z}^*\|^2 - \|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2 - \|\mathbf{z}_k - \mathbf{z}_{k+\frac{1}{2}}\|^2 \right] + \frac{\eta_k^2 \left(2 + \frac{1}{\beta}\right)}{2} \sigma_H^2 \\ &\leq \frac{1}{2} \mathbb{E} \left[ \|\mathbf{z}_k - \mathbf{z}^*\|^2 - \|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2 \right] + \frac{\eta_k^2 (1+\beta) L_H^2}{2} \mathbb{E} \left[ \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2 - \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \right] \\ &\quad - \left( \frac{1}{2} - 2\eta_k^2 (1+\beta) L_H^2 \right) \mathbb{E} \|\mathbf{z}_k - \mathbf{z}_{k+\frac{1}{2}}\|^2 + \frac{\eta_k^2 \left(2 + \frac{1}{\beta}\right)}{2} \sigma_H^2. \end{aligned}$$

Taking  $\eta_k$  satisfying  $\eta_k \leq \frac{1}{L_H \sqrt{4(1+\beta)}}$  and by rearranging the above inequality, we have

$$\begin{aligned} 0 &\leq \frac{1}{2} \mathbb{E} \left[ \|\mathbf{z}_k - \mathbf{z}^*\|^2 - \|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2 \right] + \frac{1}{8} \mathbb{E} \left[ \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2 - \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \right] \\ &\quad - \left( \frac{1}{2} - \frac{1}{2} \right) \mathbb{E} \|\mathbf{z}_k - \mathbf{z}_{k+\frac{1}{2}}\|^2 + \frac{\eta_k^2 \left(2 + \frac{1}{\beta}\right)}{2} \sigma_H^2 \\ &\leq \frac{1}{2} \mathbb{E} \left[ \|\mathbf{z}_k - \mathbf{z}^*\|^2 + \frac{1}{4} \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2 - \|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2 - \frac{1}{4} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \right] + \frac{\eta_k^2 \left(2 + \frac{1}{\beta}\right)}{2} \sigma_H^2. \end{aligned}$$

Rearranging the above inequality and we conclude that

$$\mathbb{E} \left[ \|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2 + \frac{1}{4} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \right] \leq \mathbb{E} \left[ \|\mathbf{z}_k - \mathbf{z}^*\|^2 + \frac{1}{4} \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2 \right] + \eta_k^2 \left(2 + \frac{1}{\beta}\right) \sigma_H^2.$$

Telescoping over  $k = 0, 1, \dots, K-1$  and noting that  $\mathbf{z}_{-\frac{1}{2}} = \mathbf{z}_{-\frac{3}{2}} = \mathbf{z}_0$  and  $\eta_k$  is taken as a constant for the bilinear case, we have

$$\|\mathbf{z}_K - \mathbf{z}^*\|^2 \leq \|\mathbf{z}_K - \mathbf{z}^*\|^2 + \frac{1}{4} \|\mathbf{z}_{K-\frac{1}{2}} - \mathbf{z}_{K-\frac{3}{2}}\|^2 \leq \|\mathbf{z}_0 - \mathbf{z}^*\|^2 + \eta_k^2 \left(2 + \frac{1}{\beta}\right) K \sigma_H^2,$$

which concludes our proof of Lemma F.3.  $\square$

Recalling that we take  $\alpha_k = \frac{2}{k+2}$  in (F.4b) of (F.4). Thus, we obtain the following

$$z_{k+1}^{\text{ag}} = \frac{k}{k+2} z_k^{\text{ag}} + \frac{2}{k+2} z_{k+\frac{1}{2}}.$$

Subtracting both sides by  $z^*$  and multiplying both sides by  $(k+1)(k+2)$ , we have

$$(k+1)(k+2) (z_{k+1}^{\text{ag}} - z^*) = k(k+1) (z_k^{\text{ag}} - z^*) + 2(k+1) (z_{k+\frac{1}{2}} - z^*).$$

Telescoping over  $k = 0, \dots, K-1$  and we conclude that

$$K(K+1) (z_K^{\text{ag}} - z^*) = 2 \sum_{k=0}^{K-1} (k+1) (z_{k+\frac{1}{2}} - z^*). \quad (\text{F.8})$$

Moreover, according to the update rule (F.4c), we have that

$$\begin{aligned} K z_K - \sum_{k=0}^{K-1} z_k &= \sum_{k=0}^{K-1} (k+1) z_{k+1} - (k+1) z_k = \sum_{k=0}^{K-1} \eta_k (k+1) \tilde{H}(z_{k+\frac{1}{2}}, \zeta_{k+\frac{1}{2}}) \\ &= \sum_{k=0}^{K-1} \eta_k (k+1) \left[ H(z_{k+\frac{1}{2}}) + \Delta_h^{k+\frac{1}{2}} \right]. \end{aligned} \quad (\text{F.9})$$

Recalling that  $\eta_k = \frac{1}{L_H \sqrt{4(1+\beta)}}$ , combining (F.9) with (F.8) and taking the squared norm on both sides, we conclude that

$$\begin{aligned} &\|K(K+1) (z_K^{\text{ag}} - z^*)\|^2 \\ &= \|2 \sum_{k=0}^{K-1} (k+1) (z_{k+\frac{1}{2}} - z^*)\|^2 \leq \frac{1}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})} \|2 \sum_{k=0}^{K-1} (k+1) H(z_{k+\frac{1}{2}})\|^2 \\ &= \frac{16(1+\beta)(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})} \|K z_K - \sum_{k=0}^{K-1} z_k\|^2 + \frac{4(1+\frac{1}{\gamma})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})} \left\| \sum_{k=0}^{K-1} (k+1) \Delta_h^{k+\frac{1}{2}} \right\|^2 \\ &\leq \frac{16(1+\beta)(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})} K \cdot \sum_{k=0}^{K-1} [2\|z_K - z^*\|^2 + 2\|z_k - z^*\|^2] + \frac{4(1+\frac{1}{\gamma})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})} \left\| \sum_{k=0}^{K-1} (k+1) \Delta_h^{k+\frac{1}{2}} \right\|^2. \end{aligned} \quad (\text{F.10})$$

Dividing both sides of (F.10) by  $K^2(K+1)^2$  and taking expectation, we have

$$\begin{aligned} \mathbb{E}\|z_K^{\text{ag}} - z^*\|^2 &\leq \frac{16(1+\beta)(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})K(K+1)^2} \sum_{k=0}^{K-1} \mathbb{E} [2\|z_K - z^*\|^2 + 2\|z_k - z^*\|^2] \\ &\quad + \frac{4(1+\frac{1}{\gamma})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})K^2(K+1)^2} \sum_{k=0}^{K-1} (k+1)^2 \mathbb{E}\|\Delta_h^{k+\frac{1}{2}}\|^2 \\ &\leq \frac{64(1+\beta)(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^2} \left[ \|z_0 - z^*\|^2 + \eta_k^2(2+\frac{1}{\beta})K\sigma_H^2 \right] + \frac{4(1+\frac{1}{\gamma})\sigma_H^2}{3\lambda_{\min}(\mathbf{B}^\top \mathbf{B})K} \\ &= \frac{64(1+\beta)(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^2} \|z_0 - z^*\|^2 + \frac{16(1+\gamma)(2+\frac{1}{\beta})}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)} \sigma_H^2 + \frac{4(1+\frac{1}{\gamma})\sigma_H^2}{3\lambda_{\min}(\mathbf{B}^\top \mathbf{B})K}, \end{aligned}$$

Minimize over  $\beta$ , we have  $\beta = \frac{\sigma_H \sqrt{K+1}}{2L_H \|z_0 - z^*\|}$  and the first two terms become

$$\frac{64(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^2} \|z_0 - z^*\|^2 + \frac{32(1+\gamma)}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)} \sigma_H^2 + \frac{32(1+\gamma)L_H \sigma_H}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^{3/2}} \|z_0 - z^*\|$$



If we take  $\beta = 1$ , the above reduces to

$$\begin{aligned}\mathbb{E}\|\mathbf{z}_K^{\text{ag}} - \mathbf{z}^*\|^2 &\leq \frac{128(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^2}\|\mathbf{z}_0 - \mathbf{z}^*\|^2 + \frac{48(1+\gamma)}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)}\sigma_H^2 + \frac{4(1+\frac{1}{\gamma})\sigma_H^2}{3\lambda_{\min}(\mathbf{B}^\top \mathbf{B})K} \\ &= (1+\gamma) \left( \frac{128L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^2}\|\mathbf{z}_0 - \mathbf{z}^*\|^2 + \frac{48}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)}\sigma_H^2 \right) + \frac{4(1+\frac{1}{\gamma})\sigma_H^2}{3\lambda_{\min}(\mathbf{B}^\top \mathbf{B})K}\end{aligned}$$

Further minimizing over  $\gamma$ , we conclude that so

$$\begin{aligned}\sqrt{\mathbb{E}\|\mathbf{z}_K^{\text{ag}} - \mathbf{z}^*\|^2} &\leq \sqrt{\frac{128L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^2}\|\mathbf{z}_0 - \mathbf{z}^*\|^2 + \frac{48}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)}\sigma_H^2} + \sqrt{\frac{4\sigma_H^2}{3\lambda_{\min}(\mathbf{B}^\top \mathbf{B})K}} \\ &\leq \sqrt{\frac{128L_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)^2}\|\mathbf{z}_0 - \mathbf{z}^*\|^2} + \sqrt{\frac{48\sigma_H^2}{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})(K+1)}} + \sqrt{\frac{4\sigma_H^2}{3\lambda_{\min}(\mathbf{B}^\top \mathbf{B})K}} \\ &\leq \frac{1}{\sqrt{\lambda_{\min}(\mathbf{B}^\top \mathbf{B})}} \left( \frac{8\sqrt{2}L_H\|\mathbf{z}_0 - \mathbf{z}^*\|}{K+1} + \frac{8.083\sigma_H}{\sqrt{K}} \right)\end{aligned}$$

By operating restarting techniques the same way as in the explanation of Corollary 3.2 in [Du et al. \(2022\)](#), we conclude Theorem F.2.

## G Proof of Auxiliary Lemmas

### G.1 Proof of Lemma E.2

*Proof.*[Proof of Lemma E.2] Since  $\mathcal{E}(t)$  is set as

$$\mathcal{E} = t^2 V(Z) + \frac{t}{c_t} \|\widehat{Z} - \mathbf{z}^*\|^2, \quad (\text{E.8})$$

we have its time derivative

$$\frac{d\mathcal{E}}{dt} = \underbrace{2tV(Z) + t^2\langle \nabla V(Z), \dot{Z} \rangle + \frac{2t}{c_t} \langle \widehat{Z} - \mathbf{z}^*, \dot{\widehat{Z}} \rangle}_{\text{bracketed part}} + \sqrt{3 + \sqrt{3}} \frac{c}{1 \vee c} \|\widehat{Z} - \mathbf{z}^*\|^2.$$

We want to show the bracketed part in above is nonpositive, i.e.

$$2t[F(Z) - F(\mathbf{z}^*) + \langle Z - \mathbf{z}^*, H(\mathbf{z}^*) \rangle] + t^2\langle \nabla F(Z) + H(\mathbf{z}^*), \dot{Z} \rangle + \frac{2t}{c_t} \langle \widehat{Z} - \mathbf{z}^*, \dot{\widehat{Z}} \rangle \leq 0.$$

Saddle definition gives  $\nabla F(\mathbf{z}^*) + H(\mathbf{z}^*) = 0$ , and hence  $\mu$ -strong convexity of  $F$  implies

$$\begin{aligned}V(Z) &= F(Z) - F(\mathbf{z}^*) + \langle Z - \mathbf{z}^*, H(\mathbf{z}^*) \rangle \\ &\geq \langle Z - \mathbf{z}^*, \nabla F(\mathbf{z}^*) + H(\mathbf{z}^*) \rangle + \frac{\mu}{2} \|Z - \mathbf{z}^*\|^2 = \frac{\mu}{2} \|Z - \mathbf{z}^*\|^2 \geq 0.\end{aligned}$$

Denote

$$2t[F(Z) - F(\mathbf{z}^*) + \langle Z - \mathbf{z}^*, H(\mathbf{z}^*) \rangle] + t^2\langle \nabla F(Z) + H(\mathbf{z}^*), \dot{Z} \rangle + \frac{2t}{c_t} \langle \widehat{Z} - \mathbf{z}^*, \dot{\widehat{Z}} \rangle \equiv \text{I} + \text{II} + \text{III}.$$

Then using  $\widehat{Z} = Z + \frac{t}{2} \dot{Z}$  we have

$$\begin{aligned}\text{III} &= \frac{2t}{c_t} \langle \widehat{Z} - \mathbf{z}^*, -c_t[\nabla F(Z) + H(\widehat{Z})] \rangle = -2t\langle \widehat{Z} - \mathbf{z}^*, \nabla F(Z) + H(\widehat{Z}) \rangle \\ &\leq -2t\langle \widehat{Z} - \mathbf{z}^*, \nabla F(Z) + H(\mathbf{z}^*) \rangle = -\langle 2t(Z - \mathbf{z}^*) + t^2\dot{Z}, \nabla F(Z) + H(\mathbf{z}^*) \rangle,\end{aligned}$$

where in the third equality above we used the property of monotone operator  $H$  (and also linearity) and conclude  $\langle \widehat{Z} - \mathbf{z}^*, H(\widehat{Z} - \mathbf{z}^*) \rangle \geq 0$  which is actually = 0 for bilinear operator  $H$ . Therefore

$$\text{I} + \text{II} = 2t[F(Z) - F(\mathbf{z}^*) + \langle Z - \mathbf{z}^*, H(\mathbf{z}^*) \rangle] + t^2\langle \dot{Z}, \nabla F(Z) + H(\mathbf{z}^*) \rangle,$$

and

$$\begin{aligned} \text{I} + \text{II} + \text{III} &= 2t [F(Z) - F(\mathbf{z}^*) + \langle Z - \mathbf{z}^*, H(\mathbf{z}^*) \rangle] - \langle 2t(Z - \mathbf{z}^*), \nabla F(Z) + H(\mathbf{z}^*) \rangle \\ &= 2t [F(Z) - F(\mathbf{z}^*) - \langle Z - (\mathbf{z}^*), \nabla F(Z) \rangle] \leq 0, \end{aligned}$$

where the last step uses the convexity of  $f$ . This concludes the desired result of Lemma E.2.  $\square$

## G.2 Proof of Lemma E.3

*Proof.*[Proof of Lemma E.3] To proceed with proof we adopt a Gronwall-type argument. Note

$$\dot{\mathcal{E}}(t) = \frac{d}{dt} \left[ t^2 V(Z(t)) + \frac{t}{c_t} \|\widehat{Z}(t) - \mathbf{z}^*\|^2 \right] \leq \sqrt{3 + \sqrt{3}} \frac{c}{1 \vee c} \|\widehat{Z}(t) - \mathbf{z}^*\|^2.$$

Taking integrals on both sides  $\int_0^T dt$  gives

$$T^2 V(Z(T)) + \frac{T}{c_t} \|\widehat{Z}(T) - \mathbf{z}^*\|^2 - \frac{2}{(1 \vee c)^2} \|\mathbf{z}_0 - \mathbf{z}^*\|^2 \leq \sqrt{3 + \sqrt{3}} \frac{c}{1 \vee c} \int_0^T \|\widehat{Z}(t) - \mathbf{z}^*\|^2 dt.$$

Let  $y(T) \equiv \int_0^T \|\widehat{Z}(t) - \mathbf{z}^*\|^2 dt$  then we have by removing the first term

$$\frac{T}{c_t} y'(T) - \frac{2}{(1 \vee c)^2} y'(0) = \frac{2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) T}{(1 \vee c)^2} y'(T) - \frac{2}{(1 \vee c)^2} y'(0) \leq \sqrt{3 + \sqrt{3}} \frac{c}{1 \vee c} y(T), \quad (\text{G.1})$$

which gives, via quotient rule,

$$\begin{aligned} \frac{d}{dT} \left( \frac{c_t y(T)}{T} \right) &= \frac{d}{dT} \left( \frac{(1 \vee c)^2 y(T)}{2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) T} \right) \\ &= (1 \vee c)^2 \frac{(2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) T) y'(T) - c (1 \vee c) y(T)}{(2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) T)^2} \leq \frac{2(1 \vee c)^2 y'(0)}{(2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) T)^2}, \end{aligned}$$

so

$$\begin{aligned} \frac{(1 \vee c)^2 y(T)}{2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) T} - \frac{(1 \vee c)^2 y(0)}{2} &\leq 2(1 \vee c)^2 y'(0) \int_0^T \frac{1}{(2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) t)^2} dt \\ &= 2 \frac{1 \vee c}{\sqrt{3 + \sqrt{3}} c} y'(0) \left( \frac{1}{2} - \frac{1}{2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) T} \right). \end{aligned}$$

Therefore

$$y(T) \leq 2 \frac{1}{\sqrt{3 + \sqrt{3}} c (1 \vee c)} y'(0) \left( \frac{2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) T}{2} - 1 \right) = y'(0) T.$$

Plugging in back (G.1) gives

$$\frac{T}{c_t} y'(T) - \frac{2}{(1 \vee c)^2} y'(0) \leq \sqrt{3 + \sqrt{3}} \frac{c}{1 \vee c} y(T) \leq \sqrt{3 + \sqrt{3}} \frac{c}{1 \vee c} y'(0) T,$$

so

$$\frac{T}{c_t} y'(T) = \frac{2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) T}{(1 \vee c)^2} y'(T) \leq \frac{2 + \sqrt{3 + \sqrt{3}} c (1 \vee c) T}{(1 \vee c)^2} y'(0),$$

so  $y'(T) \leq y'(0)$  which indicates

$$\|\widehat{Z}(T) - \mathbf{z}^*\|^2 \leq \|\mathbf{z}_0 - \mathbf{z}^*\|^2,$$

completing the proof of Lemma E.3.  $\square$

### G.3 Proof of Lemma 3.1

*Proof.*[Proof of Lemma 3.1] Following (D.8), we let  $\omega_z = z^*$ . Due to the non-negativity of  $V(\cdot, z^*)$ , we can eliminate the  $V$  terms and have:

$$\left(2L + \sqrt{\frac{2}{c}}L_H(K+1)\right) \|z_K - z^*\|^2 \leq \left(2L + \sqrt{\frac{2}{c}}L_H\right) \|z_0 - z^*\|^2 + \sqrt{\frac{2}{c}}L_H \sum_{k=0}^{K-1} \|z_k - z^*\|^2.$$

We adopt a "bootstrapping" argument, similar as the Gronwall-type analysis in the Proof of Theorem B.2. We define  $M_K = \max_{0 \leq k \leq K-1} \|z_k - z^*\|^2$  and taking a maximum on each term on the right hand side of the above inequality, we conclude that

$$\begin{aligned} \left(2L + \sqrt{\frac{2}{c}}L_H(K+1)\right) \|z_K - z^*\|^2 &\leq \left(2L + \sqrt{\frac{2}{c}}L_H\right) M_{K-1} + \sqrt{\frac{2}{c}}L_H \sum_{k=0}^{K-1} M_{K-1} \\ &= \left(2L + \sqrt{\frac{2}{c}}L_H(K+1)\right) M_{K-1}. \end{aligned}$$

Thus, we know that  $\|z_K - z^*\|^2 \leq M_{K-1}$  and hence  $M_K = M_{K-1}$  always holds. That yields  $M_K = M_0$ , and we conclude the proof of Lemma 3.1.  $\square$

### G.4 Proof of Lemma D.7

*Proof.*[Proof of Lemma D.7] Starting from (D.14) that

$$\begin{aligned} &[(K+1)^2 - 1] \mathbb{E}V(z_K^{\text{ag}}, z^*) + \frac{K+1}{\eta_{K-1}} \mathbb{E}\|z_K - z^*\|^2 \\ &\leq \frac{2}{\eta_0} \mathbb{E}\|z_0 - z^*\|^2 + \frac{2}{\sqrt{c}}L_H \sum_{k=1}^{K-1} \mathbb{E}\|z_k - z^*\|^2 + \sum_{k=0}^{K-1} (k+2)\eta_k \sigma^2 - \sum_{k=0}^{K-1} \mathbb{E}V(z_{k+1}^{\text{ag}}, z^*). \end{aligned}$$

We first omit the  $V(\cdot, \cdot)$  terms and have

$$\frac{K+1}{\eta_{K-1}} \mathbb{E}\|z_K - z^*\|^2 \leq \frac{2}{\eta_0} \mathbb{E}\|z_0 - z^*\|^2 + \frac{2}{\sqrt{c}}L_H \sum_{k=1}^{K-1} \mathbb{E}\|z_k - z^*\|^2 + \sum_{k=0}^{K-1} (k+2)\eta_k \sigma^2. \quad (\text{G.2})$$

Rewrite  $\|z_K - z^*\|^2$  as the difference between two summations, we obtain:

$$\frac{K+1}{\eta_{K-1}} \left( \sum_{k=1}^K - \sum_{k=1}^{K-1} \right) \mathbb{E}\|z_k - \omega_z\|^2 \leq \frac{2}{\eta_0} \mathbb{E}\|z_0 - z^*\|^2 + \frac{2}{\sqrt{c}}L_H \sum_{k=1}^{K-1} \mathbb{E}\|z_k - z^*\|^2 + \sum_{k=0}^{K-1} (k+2)\eta_k \sigma^2.$$

Rearranging the terms and by the first condition (a) that  $\frac{k+2}{\eta_k} - \frac{k+1}{\eta_{k-1}} = \frac{2}{\sqrt{c}}L_H$ , we have:

$$\frac{K+1}{\eta_{K-1}} \sum_{k=1}^K \mathbb{E}\|z_k - z^*\|^2 \leq \frac{2}{\eta_0} \mathbb{E}\|z_0 - z^*\|^2 + \frac{K+2}{\eta_K} \sum_{k=1}^{K-1} \mathbb{E}\|z_k - \omega_z\|^2 + \sum_{k=0}^{K-1} (k+2)\eta_k \sigma^2.$$

To construct a valid iterative rule, we divide both sides of the above inequality with  $\frac{(K+1)(K+2)}{\eta_{K-1}\eta_K}$  and obtain the following:

$$\frac{\eta_K}{K+2} \sum_{k=1}^K \mathbb{E}\|z_k - z^*\|^2 \leq \frac{\eta_{K-1}}{K+1} \sum_{k=1}^{K-1} \mathbb{E}\|z_k - \omega_z\|^2 + \frac{\eta_{K-1}\eta_K}{(K+1)(K+2)} \left[ \frac{2}{\eta_0} \mathbb{E}\|z_0 - z^*\|^2 + \sum_{k=0}^{K-1} (k+2)\eta_k \sigma^2 \right].$$

Here we slightly abuse the notations and use  $K$  to denote an arbitrary iteration during the process of the algorithm and use  $\mathcal{K}$  to denote the fixed total number of iterates. Thus,  $\sum_{k=0}^{K-1} (k+2)\eta_k \sigma^2 \leq \sum_{k=0}^{\mathcal{K}-1} (k+2)\eta_k \sigma^2$  is an upper bound that does

not change with the choice of  $K$ . It follows that:

$$\begin{aligned} \frac{\eta_K}{K+2} \sum_{k=1}^K \mathbb{E} \|z_k - z^*\|^2 &\leq \frac{\eta_{K-1}}{K+1} \sum_{k=1}^{K-1} \mathbb{E} \|z_k - \omega_z\|^2 + \frac{\sqrt{c}}{2L_H} \left[ \frac{\eta_{K-1}}{K+1} - \frac{\eta_K}{K+2} \right] \left[ \frac{2}{\eta_0} \mathbb{E} \|z_0 - z^*\|^2 + \sum_{k=0}^{K-1} (k+2) \eta_k \sigma^2 \right] \\ &\leq \frac{\sqrt{c}}{2L_H} \left[ \frac{\eta_0}{2} - \frac{\eta_K}{K+2} \right] \left[ \frac{2}{\eta_0} \mathbb{E} \|z_0 - z^*\|^2 + \sum_{k=0}^{K-1} (k+2) \eta_k \sigma^2 \right]. \end{aligned}$$

Dividing both sides by  $\frac{\eta_K}{K+2}$ , the result follows:

$$\sum_{k=1}^K \mathbb{E} \|z_k - z^*\|^2 \leq \frac{\sqrt{c}}{2L_H} \left[ \frac{\eta_0(K+2)}{2\eta_K} - 1 \right] \left[ \frac{2}{\eta_0} \mathbb{E} \|z_0 - z^*\|^2 + \sum_{k=0}^{K-1} (k+2) \eta_k \sigma^2 \right].$$

Bringing this into Eq. (G.2), we conclude that

$$\frac{K+1}{\eta_{K-1}} \mathbb{E} \|z_K - z^*\|^2 \leq \frac{\eta_0(K+1)}{2\eta_{K-1}} \left[ \frac{2}{\eta_0} \mathbb{E} \|z_0 - z^*\|^2 + \sum_{k=0}^{K-1} (k+2) \eta_k \sigma^2 \right].$$

Dividing both sides by  $\frac{K+1}{\eta_{K-1}}$  and we have:

$$\mathbb{E} \|z_K - z^*\|^2 \leq \frac{\eta_0}{2} \left[ \frac{2}{\eta_0} \mathbb{E} \|z_0 - z^*\|^2 + \sum_{k=0}^{K-1} (k+2) \eta_k \sigma^2 \right].$$

Now we change back using the notation  $K$  to denote the total iterates and  $k$  is the iterates indexes, we have

$$\mathbb{E} \|z_k - z^*\|^2 \leq \mathbb{E} \|z_0 - z^*\|^2 + \frac{\eta_0}{2} \sum_{k=0}^{K-1} (k+2) \eta_k \sigma^2,$$

which concludes the proof of (a) of Lemma D.7. Additionally, if  $\eta_k \leq \frac{k+2}{D}$  for some quantity  $D$ , we have

$$\sum_{k=0}^{K-1} (k+2) \eta_k \leq \sum_{k=0}^{K-1} \frac{(k+2)^2}{D} \leq \frac{(K+1)(K+2)(2K+3)}{6D}.$$

We use  $A(K) = \sqrt{(K+1)(K+2)(2K+3)/6}$  and noting that  $\eta_0 \leq \frac{2}{D}$ , we have

$$\mathbb{E} \|z_k - z^*\|^2 \leq \mathbb{E} \|z_0 - z^*\|^2 + \frac{A(K)^2 \sigma^2}{D^2},$$

which concludes our proof of (b). And (c) follows by straightforward calculations.  $\square$

## G.5 Proof of Lemma D.1

*Proof.*[Proof of Lemma D.1] Recalling that  $F$  is  $L$ -smooth. To upper-bound the difference in pointwise primal-dual gap between iterates, we first estimate the difference in function values of  $f$  via gradients at the extrapolation point. For all  $u \in \mathcal{Z}$ , the convexity and  $L$ -smoothness of  $F(\cdot)$  implies that:

$$\begin{aligned} F(z_{k+1}^{\text{ag}}) - F(u) &= F(z_{k+1}^{\text{ag}}) - F(z_k^{\text{md}}) - (F(u) - F(z_k^{\text{md}})) \\ &\leq \langle \nabla F(z_k^{\text{md}}), z_{k+1}^{\text{ag}} - z_k^{\text{md}} \rangle + \frac{L}{2} \|z_{k+1}^{\text{ag}} - z_k^{\text{md}}\|^2 - \langle \nabla F(z_k^{\text{md}}), u - z_k^{\text{md}} \rangle. \end{aligned}$$

Taking  $u = \omega_z$  and  $u = z_k^{\text{ag}}$  respectively, we conclude that

$$F(z_{k+1}^{\text{ag}}) - F(\omega_z) \leq \langle \nabla F(z_k^{\text{md}}), z_{k+1}^{\text{ag}} - z_k^{\text{md}} \rangle + \frac{L}{2} \|z_{k+1}^{\text{ag}} - z_k^{\text{md}}\|^2 - \langle \nabla F(z_k^{\text{md}}), \omega_z - z_k^{\text{md}} \rangle, \quad (\text{G.3})$$

$$F(z_{k+1}^{\text{ag}}) - F(z_k^{\text{ag}}) \leq \langle \nabla F(z_k^{\text{md}}), z_{k+1}^{\text{ag}} - z_k^{\text{md}} \rangle + \frac{L}{2} \|z_{k+1}^{\text{ag}} - z_k^{\text{md}}\|^2 - \langle \nabla F(z_k^{\text{md}}), z_k^{\text{ag}} - z_k^{\text{md}} \rangle. \quad (\text{G.4})$$

Multiplying (G.3) by  $\alpha_k$  and (G.4) by  $(1 - \alpha_k)$  and adding them up, we have

$$\begin{aligned} F(\mathbf{z}_{k+1}^{\text{ag}}) - \alpha_k F(\omega_{\mathbf{z}}) - (1 - \alpha_k) F(\mathbf{z}_k^{\text{ag}}) &\leq \langle \nabla F(\mathbf{z}_k^{\text{md}}), \mathbf{z}_{k+1}^{\text{ag}} - (1 - \alpha_k) \mathbf{z}_k^{\text{ag}} - \omega_{\mathbf{z}} \rangle + \frac{L}{2} \|\mathbf{z}_{k+1}^{\text{ag}} - \mathbf{z}_k^{\text{md}}\|^2 \\ &= \underbrace{\alpha_k \langle \nabla F(\mathbf{z}_k^{\text{md}}), \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \rangle}_{\text{I(a)}} + \underbrace{\frac{L\alpha_k^2}{2} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2}_{\text{II}}, \end{aligned} \quad (\text{G.5})$$

where by subtracting Line (2) from Line (4) of Algorithm 1 and by Line (4) it self, the last equality of (G.5) follows.

Recalling that  $\mathbf{z}_k^{\text{ag}}$  corresponds to regular iterates and  $\mathbf{z}_k^{\text{md}}$  corresponds to the extrapolated iterates of Nesterov's acceleration scheme. The squared error term II in (G.5) is brought by gradient calculation at the extrapolated point instead of the regular point. Note that if we do an implicit version of Nesterov such that  $\mathbf{z}_{k-1}^{\text{md}} = \mathbf{z}_k^{\text{ag}}$ , this squared term goes to zero, and the convergence analysis would be the same as in OGDA. This could potentially result in a new implicit algorithm with better convergence guarantee.

On the other hand, for the coupling term of the updates, we have

$$\langle H(\omega_{\mathbf{z}}), \mathbf{z}_{k+1}^{\text{ag}} - \omega_{\mathbf{z}} \rangle - (1 - \alpha_k) \langle H(\omega_{\mathbf{z}}), \mathbf{z}_k^{\text{ag}} - \omega_{\mathbf{z}} \rangle = \alpha_k \langle H(\omega_{\mathbf{z}}), \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \rangle \leq \underbrace{\alpha_k \langle H(\mathbf{z}_{k+\frac{1}{2}}), \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \rangle}_{\text{I(b)}}, \quad (\text{G.6})$$

where the last equality comes from the monotonicity property of  $H(\cdot)$  that:

$$\langle H(\mathbf{z}_{k+\frac{1}{2}}) - H(\omega_{\mathbf{z}}), \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \rangle \geq 0.$$

Summing up Eq. (G.6) and Eq. (G.5), we obtain the following:

$$\begin{aligned} F(\mathbf{z}_{k+1}^{\text{ag}}) - \alpha_k F(\omega_{\mathbf{z}}) - (1 - \alpha_k) F(\mathbf{z}_k^{\text{ag}}) + \langle H(\omega_{\mathbf{z}}), \mathbf{z}_{k+1}^{\text{ag}} - \omega_{\mathbf{z}} \rangle - (1 - \alpha_k) \langle H(\omega_{\mathbf{z}}), \mathbf{z}_k^{\text{ag}} - \omega_{\mathbf{z}} \rangle \\ \leq \underbrace{\alpha_k \langle \nabla F(\mathbf{z}_k^{\text{md}}) + H(\mathbf{z}_{k+\frac{1}{2}}), \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \rangle}_{\text{I}} + \underbrace{\frac{L\alpha_k^2}{2} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2}_{\text{II}}, \end{aligned}$$

where I is the summation of I(a) and I(b). This concludes our proof of Lemma D.1 by bringing in the definitions of  $V(\mathbf{z}_{k+1}^{\text{ag}}, \mathbf{z}^*)$  and  $V(\mathbf{z}_k^{\text{ag}}, \mathbf{z}^*)$ .  $\square$

## G.6 Proof of Lemma D.3

*Proof.*[Proof of Lemma D.3] By Young's inequality and Cauchy-Schwarz inequality, we have that

$$\begin{aligned} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 &\leq 2 \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + 2 \|\mathbf{z}_k - \mathbf{z}_{k-\frac{1}{2}}\|^2 \\ &\stackrel{(a)}{\leq} 2 \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + 2\eta_{k-1}^2 L_H^2 \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2 \stackrel{(b)}{\leq} 2 \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + c \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2, \end{aligned} \quad (\text{G.7})$$

where (a) is by Line (3) and (5) of Algorithm 1 and the definition of  $L_H$ , and (b) is by the condition in Lemma D.3 that  $\eta_k L_H \leq \sqrt{\frac{c}{2}}$ . Recursively applying the above gives (D.7) which is repeated as:

$$\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \leq 2c^k \sum_{\ell=0}^k c^{-\ell} \|\mathbf{z}_{\ell+\frac{1}{2}} - \mathbf{z}_{\ell}\|^2. \quad (\text{D.7})$$

Indeed, from (G.7)

$$c^{-k} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 - c^{-(k-1)} \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2 \leq 2c^{-k} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2,$$

so telescoping over  $k = 1, \dots, K$  gives

$$c^{-K} \|\mathbf{z}_{K+\frac{1}{2}} - \mathbf{z}_{K-\frac{1}{2}}\|^2 - \|\mathbf{z}_{\frac{1}{2}} - \mathbf{z}_{-\frac{1}{2}}\|^2 \leq 2 \sum_{k=1}^K c^{-k} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2,$$

which is just (due to  $\mathbf{z}_0 = \mathbf{z}_{-\frac{1}{2}}$ )

$$\begin{aligned}
 c^{-K} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}} \right\|^2 &\leq 2 \sum_{k=1}^K c^{-k} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\|^2 + \left\| \mathbf{z}_{\frac{1}{2}} - \mathbf{z}_0 \right\|^2 \\
 &\leq 2 \sum_{k=1}^K c^{-k} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\|^2 + \sum_{k=0}^0 c^{-k} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\|^2 \\
 &\leq 2 \sum_{k=2}^K c^{-k} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\|^2 + 2 \sum_{k=0}^0 c^{-k} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\|^2 = 2 \sum_{k=1}^K c^{-k} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\|^2.
 \end{aligned}$$

This gives (D.7) and the entire Lemma D.3.  $\square$

### G.7 Proof of Lemma D.5

*Proof.*[Proof of Lemma D.5] We recall that we denote

$$\Delta_h^{k+\frac{1}{2}} = \tilde{H}(\mathbf{z}_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}) - H(\mathbf{z}_{k+\frac{1}{2}}), \quad \Delta_f^k = \nabla \tilde{F}(\mathbf{z}_k^{\text{md}}; \xi_k) - \nabla F(\mathbf{z}_k^{\text{md}}).$$

Then, we have

$$\mathbb{E} \left\| \tilde{H}(\mathbf{z}_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}) - \tilde{H}(\mathbf{z}_{k-\frac{1}{2}}; \zeta_{k-\frac{1}{2}}) \right\|^2 = \mathbb{E} \left\| H(\mathbf{z}_{k+\frac{1}{2}}) - H(\mathbf{z}_{k-\frac{1}{2}}) + \Delta_h^{k+\frac{1}{2}} - \Delta_h^{k-\frac{1}{2}} \right\|^2.$$

By first taking expectation over  $\zeta_{k+\frac{1}{2}}$  condition on  $\mathbf{z}_{k+\frac{1}{2}}$  given, we have

$$\begin{aligned}
 \text{LHS} &\leq \mathbb{E} \left\| H(\mathbf{z}_{k+\frac{1}{2}}) - H(\mathbf{z}_{k-\frac{1}{2}}) - \Delta_h^{k-\frac{1}{2}} \right\|^2 + \mathbb{E} \left\| \Delta_h^{k+\frac{1}{2}} \right\|^2 \\
 &\leq (1 + \beta) \mathbb{E} \left\| H(\mathbf{z}_{k+\frac{1}{2}}) - H(\mathbf{z}_{k-\frac{1}{2}}) \right\|^2 + (1 + \frac{1}{\beta}) \mathbb{E} \left\| \Delta_h^{k-\frac{1}{2}} \right\|^2 + \mathbb{E} \left\| \Delta_h^{k+\frac{1}{2}} \right\|^2 \\
 &\leq (1 + \beta) L_H^2 \mathbb{E} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}} \right\|^2 + (1 + \frac{1}{\beta}) \mathbb{E} \left\| \Delta_h^{k-\frac{1}{2}} \right\|^2 + \mathbb{E} \left\| \Delta_h^{k+\frac{1}{2}} \right\|^2.
 \end{aligned}$$

Recalling that by Assumption 2.2,  $\mathbb{E} \left\| \Delta_h^{k+\frac{1}{2}} \right\|^2 \leq \sigma_H^2$  and  $\mathbb{E} \left\| \Delta_h^{k-\frac{1}{2}} \right\|^2 \leq \sigma_H^2$ , we conclude our proof of Lemma D.5.  $\square$

### G.8 Proof of Lemma D.6

*Proof.*[Proof of Lemma D.6] By inequality (D.13), we have

$$\begin{aligned}
 &\mathbb{E} V(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}) - (1 - \alpha_k) \mathbb{E} V(\mathbf{z}_k^{\text{ag}}, \omega_{\mathbf{z}}) \\
 &\leq \frac{\alpha_k \eta_k}{2} \left[ 2L_H^2 \mathbb{E} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}} \right\|^2 + 3\sigma_H^2 \right] + \alpha_k \mathbb{E} \left\langle \Delta_f^k + \Delta_h^{k+\frac{1}{2}}, \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \right\rangle \\
 &\quad + \frac{L\alpha_k^2}{2} \mathbb{E} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\|^2 + \frac{\alpha_k}{2\eta_k} \mathbb{E} \left[ \left\| \mathbf{z}_k - \omega_{\mathbf{z}} \right\|^2 - \left\| \mathbf{z}_{k+1} - \omega_{\mathbf{z}} \right\|^2 - \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\|^2 \right]
 \end{aligned}$$

The inner product term can be decomposed into

$$\begin{aligned}
 &\mathbb{E} \left\langle \Delta_f^k + \Delta_h^{k+\frac{1}{2}}, \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \right\rangle \\
 &= \mathbb{E} \left\langle \Delta_h^{k+\frac{1}{2}}, \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \right\rangle + \mathbb{E} \left\langle \Delta_f^k, \mathbf{z}_k - \omega_{\mathbf{z}} \right\rangle + \mathbb{E} \left\langle \Delta_f^k, \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\rangle = \mathbb{E} \left\langle \Delta_f^k, \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\rangle,
 \end{aligned}$$

Where the expectation of the first two terms all equals 0. Thus, we obtain

$$\begin{aligned}
 &\mathbb{E} V(\mathbf{z}_{k+1}^{\text{ag}}, \omega_{\mathbf{z}}) - (1 - \alpha_k) \mathbb{E} V(\mathbf{z}_k^{\text{ag}}, \omega_{\mathbf{z}}) \\
 &\leq \frac{\alpha_k \eta_k}{2} \left[ 2L_H^2 \mathbb{E} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}} \right\|^2 + 3\sigma_H^2 \right] + \alpha_k \mathbb{E} \left\langle \Delta_f^k, \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\rangle \\
 &\quad + \frac{\alpha_k}{2\eta_k} \mathbb{E} \left[ \left\| \mathbf{z}_k - \omega_{\mathbf{z}} \right\|^2 - \left\| \mathbf{z}_{k+1} - \omega_{\mathbf{z}} \right\|^2 \right] - \left( \frac{\alpha_k}{2\eta_k} - \frac{L\alpha_k^2}{2} \right) \mathbb{E} \left\| \mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k \right\|^2.
 \end{aligned}$$



For any  $r > 0$ , we pair up

$$-\frac{(1-r)\alpha_k}{2\eta_k}\mathbb{E}\|z_{k+\frac{1}{2}} - z_k\|^2 + \alpha_k\mathbb{E}\langle \Delta_f^k, z_{k+\frac{1}{2}} - z_k \rangle \leq \frac{\alpha_k\eta_k}{2(1-r)}\mathbb{E}\|\Delta_f^k\|^2,$$

and thus

$$\begin{aligned} & \mathbb{E}V(z_{k+1}^{\text{ag}}, \omega_z) - (1 - \alpha_k)\mathbb{E}V(z_k^{\text{ag}}, \omega_z) \\ & \leq \frac{\alpha_k\eta_k}{2} \left[ 2L_H^2\mathbb{E}\|z_{k+\frac{1}{2}} - z_{k-\frac{1}{2}}\|^2 + 3\sigma_H^2 \right] + \frac{\alpha_k\eta_k}{2(1-r)}\mathbb{E}\|\Delta_f^k\|^2 \\ & \quad + \frac{\alpha_k}{2\eta_k}\mathbb{E}[\|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2] - \left( \frac{r\alpha_k}{2\eta_k} - \frac{L\alpha_k^2}{2} \right) \mathbb{E}\|z_{k+\frac{1}{2}} - z_k\|^2. \end{aligned} \quad (\text{G.8})$$

Next, we connect  $\|z_{k+\frac{1}{2}} - z_{k-\frac{1}{2}}\|^2$  with the squared norms  $\|z_{\ell+\frac{1}{2}} - z_\ell\|^2$ . For  $\eta_k$  satisfying  $\eta_k L_H \leq \frac{\sqrt{c}}{2}$ , we have

$$\begin{aligned} & \mathbb{E}\left\|z_{k+\frac{1}{2}} - z_{k-\frac{1}{2}}\right\|^2 \leq 2\mathbb{E}\|z_{k+\frac{1}{2}} - z_k\|^2 + 2\mathbb{E}\|z_k - z_{k-\frac{1}{2}}\|^2 \\ & = 2\mathbb{E}\|z_{k+\frac{1}{2}} - z_k\|^2 + 2\eta_{k-1}^2\mathbb{E}\|\tilde{H}(z_{k-\frac{1}{2}}) - \tilde{H}(z_{k-\frac{3}{2}})\|^2 \\ & = 2\mathbb{E}\|z_{k+\frac{1}{2}} - z_k\|^2 + 2\eta_{k-1}^2\mathbb{E}\|H(z_{k-\frac{1}{2}}) - H(z_{k-\frac{3}{2}}) + \Delta_h^{k-\frac{3}{2}}\|^2 + 2\eta_{k-1}^2\mathbb{E}\|\Delta_h^{k-\frac{1}{2}}\|^2 \\ & \leq 2\mathbb{E}\|z_{k+\frac{1}{2}} - z_k\|^2 + 4\eta_{k-1}^2L_H^2\mathbb{E}\|z_{k-\frac{1}{2}} - z_{k-\frac{3}{2}}\|^2 + 6\eta_{k-1}^2\sigma_H^2 \\ & = 2\sum_{\ell=0}^k c^{k-\ell}\mathbb{E}\|z_{\ell+\frac{1}{2}} - z_\ell\|^2 + 6\sum_{\ell=0}^k c^{k-\ell}\eta_{\ell-1}^2\sigma_H^2. \end{aligned} \quad (\text{G.9})$$

Bringing Eq. (G.9) into (G.8), we have

$$\begin{aligned} & \mathbb{E}V(z_{k+1}^{\text{ag}}, \omega_z) - (1 - \alpha_k)\mathbb{E}V(z_k^{\text{ag}}, \omega_z) \\ & \leq \frac{\alpha_k\eta_k}{2} \left[ 4L_H^2\sum_{\ell=0}^k c^{k-\ell}\mathbb{E}\|z_{\ell+\frac{1}{2}} - z_\ell\|^2 + 12L_H^2\sum_{\ell=0}^k c^{k-\ell}\eta_{\ell-1}^2\sigma_H^2 + 3\sigma_H^2 \right] + \frac{\alpha_k\eta_k}{2(1-r)}\sigma_F^2 \\ & \quad + \frac{\alpha_k}{2\eta_k}\mathbb{E}[\|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2] - \left( \frac{r\alpha_k}{2\eta_k} - \frac{L\alpha_k^2}{2} \right) \mathbb{E}\|z_{k+\frac{1}{2}} - z_k\|^2 \\ & \leq \frac{\alpha_k\eta_k}{2} \left[ 4L_H^2\sum_{\ell=0}^k c^{k-\ell}\mathbb{E}\|z_{\ell+\frac{1}{2}} - z_\ell\|^2 + 3\frac{c}{1-c}\sigma_H^2 + 3\sigma_H^2 \right] + \frac{\alpha_k\eta_k}{2(1-r)}\sigma_F^2 \\ & \quad + \frac{\alpha_k}{2\eta_k}\mathbb{E}[\|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2] - \left( \frac{r\alpha_k}{2\eta_k} - \frac{L\alpha_k^2}{2} \right) \mathbb{E}\|z_{k+\frac{1}{2}} - z_k\|^2 \\ & \leq 2\alpha_k\eta_kL_H^2\sum_{\ell=0}^k c^{k-\ell}\mathbb{E}\|z_{\ell+\frac{1}{2}} - z_\ell\|^2 + \frac{3\alpha_k\eta_k}{2(1-c)}\sigma_H^2 + \frac{\alpha_k\eta_k}{2(1-r)}\sigma_F^2 \\ & \quad + \frac{\alpha_k}{2\eta_k}\mathbb{E}[\|z_k - \omega_z\|^2 - \|z_{k+1} - \omega_z\|^2] - \left( \frac{r\alpha_k}{2\eta_k} - \frac{L\alpha_k^2}{2} \right) \mathbb{E}\|z_{k+\frac{1}{2}} - z_k\|^2, \end{aligned}$$

which concludes our proof of Lemma D.6.  $\square$

## G.9 Proof of Lemma F.1

*Proof.*[Proof of Lemma F.1] The optimal condition of the problem yields  $H(z^*) = 0$  for  $z^*$  being the solution of the VI. By the monotonicity of  $H(\cdot)$ , let  $z = z_{k+\frac{1}{2}}$  and  $z' = \omega_z$  in (2.2), we have that

$$\langle H(z_{k+\frac{1}{2}}) - H(\omega_z), z_{k+\frac{1}{2}} - \omega_z \rangle \geq 0, \quad \forall \omega_z \in \mathcal{Z}. \quad (\text{G.10})$$

Let  $\varphi_1 = \mathbf{z}_{k+\frac{1}{2}}$ ,  $\varphi_2 = \mathbf{z}_{k+1}$ ,  $\boldsymbol{\theta} = \mathbf{z}_k$ ,  $\boldsymbol{\delta}_1 = \eta_k H(\mathbf{z}_{k-\frac{1}{2}})$ ,  $\boldsymbol{\delta}_2 = \eta_k H(\mathbf{z}_{k+\frac{1}{2}})$  and  $\mathbf{z} = \omega_{\mathbf{z}}$  in Proposition D.2, we have

$$\begin{aligned} \eta_k \left\langle H(\mathbf{z}_{k+\frac{1}{2}}), \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \right\rangle &\leq \frac{\eta_k^2}{2} \|H(\mathbf{z}_{k+\frac{1}{2}}) - H(\mathbf{z}_{k-\frac{1}{2}})\|^2 + \frac{1}{2} \left[ \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_k - \mathbf{z}_{k+\frac{1}{2}}\|^2 \right] \\ &\leq \frac{\eta_k^2 L_H^2}{2} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 + \frac{1}{2} \left[ \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_k - \mathbf{z}_{k+\frac{1}{2}}\|^2 \right], \end{aligned} \quad (\text{G.11})$$

where the last inequality follows by the  $L_H$ -Lipschitzness of the  $H$  operator. Combining (G.11) with (G.10), we obtain

$$0 = \eta_k \left\langle H(\omega_{\mathbf{z}}), \mathbf{z}_{k+\frac{1}{2}} - \omega_{\mathbf{z}} \right\rangle \leq \frac{\eta_k^2 L_H^2}{2} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 + \frac{1}{2} \left[ \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_k - \mathbf{z}_{k+\frac{1}{2}}\|^2 \right]. \quad (\text{G.12})$$

Next, we move on to estimate  $\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2$ . As we know that via Young's and Cauchy-Schwarz's inequalities and the update rules (3.11a) and (3.11c), for all  $k \geq 1$

$$\begin{aligned} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 &\leq 2\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + 2\|\mathbf{z}_k - \mathbf{z}_{k-\frac{1}{2}}\|^2 \\ &\leq 2\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + 2\eta_{k-1}^2 L_H^2 \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2. \end{aligned}$$

Multiplying both sides by 2 and moving one term to the right hand gives for all  $k \geq 1$

$$\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \leq 4\|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_k\|^2 + 4\eta_{k-1}^2 L_H^2 \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2 - \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2.$$

Bringing this into (G.12) and noting that  $\eta_{k-1} \leq \frac{1}{2L_H}$  as well as  $\eta_k \leq \frac{1}{2L_H}$ , we have

$$\begin{aligned} 0 &\leq \frac{\eta_k^2 L_H^2}{2} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 + \frac{1}{2} \left[ \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_k - \mathbf{z}_{k+\frac{1}{2}}\|^2 \right] \\ &\leq \frac{1}{2} \left[ \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2 \right] + \frac{\eta_k^2 L_H^2}{2} \left[ \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2 - \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \right] - \left( \frac{1}{2} - 2\eta_k^2 L_H^2 \right) \|\mathbf{z}_k - \mathbf{z}_{k+\frac{1}{2}}\|^2 \\ &\leq \frac{1}{2} \left[ \|\mathbf{z}_k - \omega_{\mathbf{z}}\|^2 + \frac{1}{4} \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2 - \|\mathbf{z}_{k+1} - \omega_{\mathbf{z}}\|^2 - \frac{1}{4} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \right]. \end{aligned}$$

Rearranging the above inequality and take  $\omega_{\mathbf{z}} = \mathbf{z}^*$  and we conclude that

$$\|\mathbf{z}_{k+1} - \mathbf{z}^*\|^2 + \frac{1}{4} \|\mathbf{z}_{k+\frac{1}{2}} - \mathbf{z}_{k-\frac{1}{2}}\|^2 \leq \|\mathbf{z}_k - \mathbf{z}^*\|^2 + \frac{1}{4} \|\mathbf{z}_{k-\frac{1}{2}} - \mathbf{z}_{k-\frac{3}{2}}\|^2.$$

Telescoping over  $k = 0, 1, \dots, K-1$  and noting that  $\mathbf{z}_{-\frac{1}{2}} = \mathbf{z}_{-\frac{3}{2}} = \mathbf{z}_0$ , we have

$$\|\mathbf{z}_K - \mathbf{z}^*\|^2 \leq \|\mathbf{z}_K - \mathbf{z}^*\|^2 + \frac{1}{4} \|\mathbf{z}_{K-\frac{1}{2}} - \mathbf{z}_{K-\frac{3}{2}}\|^2 \leq \|\mathbf{z}_0 - \mathbf{z}^*\|^2,$$

which concludes our proof of Lemma F.1.  $\square$