Nesterov Meets Optimism: Rate-Optimal Separable Minimax Optimization

Chris Junchi Li [*1] Huizhuo Yuan [*2] Gauthier Gidel [3] Quanquan Gu [2] Michael I. Jordan [1 4]

Abstract

We propose a new first-order optimization algorithm — AcceleratedGradient-OptimisticGradient (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:

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

When $\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) : \mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ is a smooth function that is convex in $\boldsymbol{x}$ and concave in $\boldsymbol{y}$, 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

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

We refer to $f(\boldsymbol{x}) - g(\boldsymbol{y})$ as the *individual component*, and $I(\boldsymbol{x},\boldsymbol{y})$ as the *coupling component* of Problem (1.2). Let f be μ_f-strongly convex and L_f-smooth and g be μ_g-strongly convex and L_g-smooth. Let $I(\boldsymbol{x},\boldsymbol{y})$ be convex-concave with blockwise smoothness parameters I_{xx}, I_{xy}, I_{yy} where $||\nabla^2_{xx} I||_{\text{op}} \le I_{xx}$, $||\nabla^2_{xy} I||_{\text{op}} \le I_{xy}$, and $||\nabla^2_{yy} I||_{\text{op}} \le I_{yy}$. Let $I(\boldsymbol{x},\boldsymbol{y})$ be L_H-smooth, and it is straightforward to observe that L_H can be picked as small as $I_{xx} \vee I_{yy} +$

[*] Equal contribution [1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley [2] Department of Computer Sciences, University of California, Los Angeles [3] DIRO, Université de Montréal and Mila [4] Department of Statistics, University of California, Berkeley. Correspondence to: Chris Junchi Li <junchili@berkeley.edu>, Gauthier Gidel <gauthier.gidel@umontreal.ca>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

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:

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

Here we take $I(\boldsymbol{x},\boldsymbol{y})$ as the bilinear coupling function $\boldsymbol{x}^\top \mathbf{B}\boldsymbol{y}$ and is L_H-smooth where L_H can be picked as small as the operator norm $\|\mathbf{B}\|_{\text{op}}$ 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 $\frac{\bar{L}}{\bar{\mu}} \log\left(\frac{1}{\epsilon}\right)$ for finding an ϵ-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 $\boldsymbol{x}$ and $\boldsymbol{y}$, fine-grained convergence rates that depend on the individual strong convexity μ_f, μ_g and smoothness parameters L_f, L_g and also I_{xx}, I_{xy}, I_{yy} are more desirable. In fact, Zhang et al. (2021a) have proved the following iteration complexity lower bound for solving (1.2) via any first-order algorithms under the linear span assumption:

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

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).[1] 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 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. Using a scheduled restarting, we derive an *AcceleratedGradient OptimisticGradient with restarting* (AG-OG with restarting) algorithm that achieves a sharp convergence rate in a variety of settings. We provide theoretical analysis of our algorithm for general separable SC-SC problem (1.2) and compare the results with existing literature under special cases in the form of (1.3) (bi-SC-SC, bi-C-SC and Bilinear).

(ii) Using a scheduled restarting, we derive an *AcceleratedGradient-OptimisticGradient with restarting* (AG-OG with restarting) algorithm that achieves a sharp convergence rate in a variety of settings. For general separable SC-SC setting in (1.2), our algorithm achieves a complexity of $\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{I_{xx}}{\mu_f} \vee \frac{I_{xy}}{\sqrt{\mu_f \mu_g}} \vee \frac{I_{yy}}{\mu_g}\right) \log\left(\frac{1}{\epsilon}\right)$, matching the best known upper bound in Jin et al. (2022). For the setting of bilinearly coupled SC-SC in (1.3), our algorithm achieves a complexity of $\mathcal{O}\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \sqrt{\frac{\|\mathbf{B}\|_{\text{op}}}{\mu_f \mu_g}}\right) \log\left(\frac{1}{\epsilon}\right)$ [Corollary 3.4], which matches the lower bound established by Zhang et al. (2021a). For bi-C-SC, we prove a

[1] 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 $\nabla I(\boldsymbol{x},\boldsymbol{y})$ at least twice per iteration.

$\mathcal{O}\left(\sqrt{\frac{L_f}{\epsilon} \vee \frac{L_g}{\mu_g}} + \frac{\|\mathbf{B}\|_{\mathrm{op}}}{\sqrt{\epsilon\mu_g}}\right)\log\left(\frac{1}{\epsilon}\right)$ 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 AG-OG 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., $\frac{\sigma^2}{\mu_f^2\epsilon^2}$ [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 $\Omega\left(\frac{\|\mathbf{B}\|_{\mathrm{op}}}{\sqrt{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}}\right)\log\left(\frac{1}{\epsilon}\right)$ [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.

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 stochastic extragradient method 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. (C.7)). Although 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. (C.8)) 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.[2]

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 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

[2] Limited by space, we refer readers to §C.1 and §C.4 for technical details.

Setting Method	SC-SC	bi-SC-SC	Bilinear	bi-C-SC	Stochastic Rate	Single Call
OGDA (Mokhtari et al., 2020b)	$\widetilde{\mathcal{O}}\left(\frac{L'_f \vee L'_g \vee I_{xy}}{\mu_f \wedge \mu_g}\right)$	$\widetilde{\mathcal{O}}\left(\frac{L_f \vee L_g \vee \\|\mathbf{B}\\|_{\text{op}}}{\mu_f \wedge \mu_g}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\\|\mathbf{B}\\|_{\text{op}}^2}{\lambda_{\min}}\right)$	$\mathcal{O}\left(\frac{L_f \vee L_g \vee \\|\mathbf{B}\\|_{\text{op}}}{\epsilon}\right)$	✓	✓
Proximal Best Response (Wang & Li, 2020)	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L'_f}{\mu_f} \vee \frac{L'_g}{\mu_g}} + \sqrt{\frac{I_{xy}(L'_f \vee L'_g \vee I_{xy})}{\mu_f \mu_g}}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \sqrt{\frac{\\|\mathbf{B}\\|_{\text{op}}(L_f \vee L_g \vee \\|\mathbf{B}\\|_{\text{op}})}{\mu_f \mu_g}}\right)$	—	—	✘	✘
DIPPA (Xie et al., 2021)	—	$\widetilde{\mathcal{O}}\left(\left(\frac{L_f L_g}{\mu_f \mu_g}\left(\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}\right)\right)^{\frac{1}{4}} + \frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\mu_f \mu_g}}\right)$	—	—	✘	✘
LPD (Thekumparampil et al., 2022)	—	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\mu_f \mu_g}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\\|\mathbf{B}\\|_{\text{op}}^2}{\lambda_{\min}}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\epsilon} \vee \frac{L_g}{\mu_g}} + \frac{\\|\boldsymbol{B}\\|_{\text{op}}}{\sqrt{\epsilon \mu_g}}\right)$	✘	✓
APDG (Kovalev et al., 2021) (Metelev et al., 2022)	—	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\mu_f \mu_g}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\\|\mathbf{B}\\|_{\text{op}}^2}{\lambda_{\min}}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f L_g}{\lambda_{\min}}} \vee \frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\lambda_{\min}}}\sqrt{\frac{L_g}{\mu_g}} \vee \frac{\\|\mathbf{B}\\|_{\text{op}}^2}{\lambda_{\min}}\right)$	✓	✘
PD-EG (Jin et al., 2022)	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{I_{xx}}{\mu_f} \vee \frac{I_{xy}}{\sqrt{\mu_f \mu_g}} \vee \frac{I_{yy}}{\mu_g}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\mu_f \mu_g}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\\|\mathbf{B}\\|_{\text{op}}^2}{\lambda_{\min}}\right)$	—	✘	✘
EG+Momentum (Azizian et al., 2020)	—	—	$\widetilde{\mathcal{O}}\left(\frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\lambda_{\min}}}\right)$	—	✘	✘
SEG with Restarting (Li et al., 2022)	—	—	$\widetilde{\mathcal{O}}\left(\frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\lambda_{\min}}}\right)$	—	✓	✘
AG-EG with Restarting (Du et al., 2022)	—	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\mu_f \mu_g}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\lambda_{\min}}}\right)$	—	✓	✘
AG-OG with Restarting (this work)	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{I_{xx}}{\mu_f} \vee \frac{I_{xy}}{\sqrt{\mu_f \mu_g}} \vee \frac{I_{yy}}{\mu_g}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\mu_f \mu_g}}\right)$	$\widetilde{\mathcal{O}}\left(\frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\lambda_{\min}}}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\epsilon} \vee \frac{L_g}{\mu_g}} + \frac{\\|\boldsymbol{B}\\|_{\text{op}}}{\sqrt{\epsilon \mu_g}}\right)$	✓	✓
Lower Bound (Zhang et al., 2021a) (Ibrahim et al., 2020)	$\widetilde{\Omega}\left(\sqrt{\frac{L'_f}{\mu_f} \vee \frac{L'_g}{\mu_g}} + \frac{I_{xy}}{\sqrt{\mu_f \mu_g}}\right)$	$\widetilde{\Omega}\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\mu_f \mu_g}}\right)$	$\widetilde{\Omega}\left(\frac{\\|\mathbf{B}\\|_{\text{op}}}{\sqrt{\lambda_{\min}}}\right)$	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{L_f}{\epsilon} \vee \frac{L_g}{\mu_g}} + \frac{\\|\boldsymbol{B}\\|_{\text{op}}}{\sqrt{\epsilon \mu_g}}\right)$	—	—

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.

$a_n \geq C b_n$ (resp. $a_n \leq C b_n$) 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 $\widetilde{\mathcal{O}}$ or $\widetilde{\Omega}$ 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 $\|\mathbf{A}\|_{\text{op}}$ the operator norm $\sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}$. Let vector $\boldsymbol{z} = [\boldsymbol{x}; \boldsymbol{y}] \in \mathbb{R}^{n+m}$ denote the concatenation of $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{y} \in \mathbb{R}^m$. We use $\wedge$ (resp. $\vee$) to denote the bivariate $\min$ (resp. $\max$) throughout this paper. For natural number K let $[K]$ denote the set $\{1, \ldots, K\}$. Throughout the paper we also use the standard notation $\|\cdot\|$ to denote the ℓ_2-norm and $\|\cdot\|_{\text{op}}$ to denote the operator norm of a matrix. We will explain other notations at their first appearances.

2. Preliminaries

In minimax optimization the goal is to find an (approximate) Nash equilibrium (or minimax point) of problem (1.1) (or (1.2)), defined as a pair $[\boldsymbol{x}^*; \boldsymbol{y}^*] \in \mathcal{X} \times \mathcal{Y}$ satisfying:

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

In order to analyze first-order gradient methods for this problem, we assume access to the gradients of the objective $\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ and $\nabla_{\boldsymbol{y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$. 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 $\boldsymbol{z} = [\boldsymbol{x}; \boldsymbol{y}] \in \mathbb{R}^{n+m}$:

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

Based on this formulation, our goal is to find the stationary point of the vector field correponding to the monotone operator $W(\boldsymbol{z})$, namely a point $\boldsymbol{z}^* = [\boldsymbol{x}^*; \boldsymbol{y}^*] \in \mathbb{R}^{n+m}$ satisfying (in the unconstrained case) $W(\boldsymbol{z}^*) = 0$, 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 $F(\boldsymbol{z}) := f(\boldsymbol{x}) + g(\boldsymbol{y})$ and correspondingly $\nabla F(\boldsymbol{z}) = [\nabla f(\boldsymbol{x}); \nabla g(\boldsymbol{y})]$. For the coupling component, we define the operator $H(\boldsymbol{z}) = [\nabla_{\boldsymbol{x}} I(\boldsymbol{x}, \boldsymbol{y}); -\nabla_{\boldsymbol{y}} I(\boldsymbol{x}, \boldsymbol{y})]$. Note that the representation allows us to write $W(\boldsymbol{z})$ as the summation of the two vector fields: $W(\boldsymbol{z}) = \nabla F(\boldsymbol{z}) + H(\boldsymbol{z})$.

We introduce our main assumptions as follows:

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

This implies that $F(\boldsymbol{z})$ 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 $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^{n+m}$:

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

The above assumption adds convexity and smoothness constraints to the individual components $f(\boldsymbol{x})$ and $g(\boldsymbol{y})$. In addition, for the coupling component $\boldsymbol{x}^\top \mathbf{B} \boldsymbol{y}$ 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 $\boldsymbol{x}$ and $\boldsymbol{y}$ are exchangeable, tall matrices cover all circumstances.

In the stochastic setting, we assume access to an unbiased stochastic oracle $\widetilde{H}(\boldsymbol{z}; \zeta)$ of $H(\boldsymbol{z})$ and an unbiased stochastic oracle $\nabla \widetilde{F}(\boldsymbol{z}; \xi)$ of $\nabla F(\boldsymbol{z})$. 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 $\sigma_H^2, \sigma_F^2 \geq 0$:

$$\mathbb{E}_\xi \left[||\widetilde{H}(\boldsymbol{z}; \zeta) - H(\boldsymbol{z})||^2 \right] \leq \sigma_H^2,$$
$$\mathbb{E}_\zeta \left[||\nabla \widetilde{F}(\boldsymbol{z}; \xi) - \nabla F(\boldsymbol{z})||^2 \right] \leq \sigma_F^2.$$

For ease of exposition, we introduce the overall variance $\sigma^2 = 3\sqrt{2}\sigma_H^2 + 2\sigma_F^2$. Note that the noise variance bound assumption is common in the stochastic optimization literature.[3] Under the above assumptions, our goal is to find an *ϵ-optimal minimax point*, a notion defined as follows.

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

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

3. AcceleratedGradient OptimisticGradient Descent Ascent

In this section, we discuss key elements of our algorithm design—consisting of *OptimisticGradient 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 OptimisticGradient 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):

$$\boldsymbol{z}_{k+\frac{1}{2}} = \boldsymbol{z}_k - \eta W(\boldsymbol{z}_{k-\frac{1}{2}}), \quad \boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta W(\boldsymbol{z}_{k+\frac{1}{2}}). \tag{3.1}$$

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

$$\boldsymbol{z}_{k+\frac{1}{2}} = \boldsymbol{z}_{k-\frac{1}{2}} - 2\eta W(\boldsymbol{z}_{k-\frac{1}{2}}) + \eta W(\boldsymbol{z}_{k-\frac{3}{2}}). \tag{3.2}$$

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

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

Another popular tractable approximation of the PP method is the extragradient (EG) method (Korpelevich, 1976): Although conceptually similar to OGDA (3.1), EG requires two gradient queries per iteration and hence doubles the overall number of gradient computations. Both OGDA and EG dynamics (3.1) alleviate cyclic behavior by extrapolation from the past and exhibit a complexity of $\mathcal{O}(L/\mu \log(1/\epsilon))$ (Gidel et al., 2018; Mokhtari et al., 2020a) in general setting (1.1) with L-smooth, μ-strongly-convex-μ-strongly-concave objectives.[4]

3.2. Nesterov's Acceleration Scheme

Turning to the minimization problem, while vanilla gradient descent enjoys an iteration complexity of $\mathcal{O}(\kappa \log(1/\epsilon))$ on L-smooth, μ-strongly convex problems, with $\kappa = L/\mu$ being the condition number, Nesterov's method (Nesterov,

[3] We leave the generalization to models of unbounded noise to future work.

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

1983), when equipped with proper restarting, achieves an improved iteration complexity of $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$. We adopt the following version of the Nesterov acceleration, known as the "second scheme" (Tseng, 2008; Lin et al., 2020b):

$$\begin{cases} z_k^{\text{md}} &= \frac{k}{k+2} z_k^{\text{ag}} + \frac{2}{k+2} z_k, & \text{(3.3a)} \\ z_{k+1} &= z_k - \eta_k \nabla F(z_k^{\text{md}}), & \text{(3.3b)} \\ z_{k+1}^{\text{ag}} &= \frac{k}{k+2} z_k^{\text{ag}} + \frac{2}{k+2} z_{k+1}. & \text{(3.3c)} \end{cases}$$

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

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

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

$$\frac{k+2}{k+3} z_{k+1}^{\text{ag}} - z_{k+1}^{\text{md}} = \frac{k}{k+3} z_k^{\text{ag}} - \frac{k+1}{k+3} z_{k+1}^{\text{ag}} \tag{3.5}$$
$$\Rightarrow \quad z_{k+1}^{\text{md}} = z_{k+1}^{\text{ag}} + \frac{k}{k+3}\left(z_{k+1}^{\text{ag}} - z_k^{\text{ag}}\right). \tag{3.6}$$

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^{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 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 (3.3a) and (3.3c), while in Lines 3 and 5 the updates follow the OGDA dynamics (3.1) with each step modified by (3.3b). Algorithm 1 can be seen as a synthesis of OGDA and Nesterov's acceleration, as it reduces to OGDA when $\nabla F = 0$ and to Nesterov's accelerated gradient when $H = 0$.

For theoretical analysis, we first state a nonexpansiveness lemma of z_k with respect to z^*, the unique solution to Problem (1.2). The proof is presented in §E.2.

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

```
1: for $k = 0, 1, \ldots, K-1$ 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: $z_K^{\text{ag}}$
```

Lemma 3.1 (Nonexpansiveness)**.** *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 and choose initialization $z_{-\frac{1}{2}} = z_0^{ag} = z_0$, at any iterate $k < K$ we have*

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

Remark 3.2. The result in Lemma 3.1 is significant in that it establishes the *last-iterate nonexpansiveness* ruled by the initialization z_0: the z_k iteration stays in the ball centered at z^* with radius $||z_0 - z^*||$. This is essential in proving convergence results of iteration z_k^{ag} where the main technical difficulty lies upon the additional recursive analysis due to gradient evaluation in a previous iterate. From a past extragradient perspective, earlier analysis was focusing on the half iterates in extragradient step (3) ($z_{k+\frac{1}{2}}$ in our formulation). In contrast, we perform a nonexpansiveness analysis on the integer steps (z_k), serving as a critical improvement over the best previous result achieved by Mokhtari et al. (2020b, Lemma 2(b)) (consider the bilinear coupling case where $f = 0$, $g = 0$), which merely admits a factor of $\sqrt{2}$ in terms of the Euclidean metric (i.e., $||z_k - z^*|| \le \sqrt{2}||z_0 - z^*||$).

With the parameter choice in Lemma 3.1, Line 4 can also be seen as an average step 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 on problem* (1.2) *satisfies:*

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

Here we use μ to denote $\mu_f \wedge \mu_g$.

The proof of Theorem 3.3 is provided in §C.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 stepsize choice is

Algorithm 2 AcceleratedGradient-OptimisticGradient with restarting (AG-OG with restarting)

```
Require: Initialization $z_0^0$, total number of epochs $N \geq 1$,
    per-epoch iterates $(K_n : n = 0, \dots, N-1)$
for $n = 0, 1, \dots, N-1$ do
    $z_{\text{out}} = \text{AG-OG}(z_0^n, z_0^n, z_0^n, K_n)$
    Set $z_0^{n+1} \leftarrow z_{\text{out}}$
    //Warm-starting from the previous output
end for
Output: $z_0^N$
```

larger than the ones used in previous techniques (Chen et al., 2017; Du et al., 2022), which is brought by a fine-tuned analysis of (C.7) in the proof of Theorem 3.3. The convergence rate in (3.7) for strongly convex problems is slow and not even linear. 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) in a broad regime of parameters.

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 OGDA 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 (μ_f and μ_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 $\widehat{\boldsymbol{y}} = \sqrt{\frac{\mu_g}{\mu_f}}\boldsymbol{y}$, we have $\nabla_{\widehat{\boldsymbol{y}}} H(\boldsymbol{z}) = \sqrt{\frac{\mu_f}{\mu_g}}\nabla_{\boldsymbol{y}} H(\boldsymbol{z})$ and $\nabla_{\widehat{\boldsymbol{y}}} g(\boldsymbol{y}) = \sqrt{\frac{\mu_f}{\mu_g}}\nabla g(\boldsymbol{y})$. Other scaling changes are listed as follows:

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

where by $\eta_{k,\boldsymbol{y}}$ we mean that when updating $\boldsymbol{z} = [\boldsymbol{x}; \boldsymbol{y}] \in \mathbb{R}^{n+m}$, we adopt stepsize η_k on the $\boldsymbol{x}$-part (first n coordinates) and $\eta_{k,\boldsymbol{y}}$ on the $\boldsymbol{y}$-part (last m coordinates). Writing out in details, the update rules with adjusted stepsizes on $[\boldsymbol{x}; \boldsymbol{y}]$ are as follows:

$$
\begin{cases}
x_k^{\text{md}} = (1-\alpha_k)x_k^{\text{ag}} + \alpha_k x_k \\
y_k^{\text{md}} = (1-\alpha_k)y_k^{\text{ag}} + \alpha_k y_k \\
x_{k+\frac{1}{2}} = x_k - \eta_k \left(I_x(x_{k-\frac{1}{2}}, y_{k-\frac{1}{2}}) + \nabla f(x_k^{\text{md}})\right) \\
y_{k+\frac{1}{2}} = y_k - \eta_{k,\boldsymbol{y}} \left(-I_y(x_{k-\frac{1}{2}}, y_{k-\frac{1}{2}}) + \nabla g(y_k^{\text{md}})\right) \\
x_{k+1}^{\text{ag}} = (1-\alpha_k)x_k^{\text{ag}} + \alpha_k x_{k+\frac{1}{2}} \\
y_{k+1}^{\text{ag}} = (1-\alpha_k)y_k^{\text{ag}} + \alpha_k y_{k+\frac{1}{2}} \\
x_{k+1} = x_k - \eta_k \left(I_x(x_{k+\frac{1}{2}}, y_{k+\frac{1}{2}}) + \nabla f(x_k^{\text{md}})\right) \\
y_{k+1} = y_k - \eta_{k,\boldsymbol{y}} \left(-I_y(x_{k+\frac{1}{2}}, y_{k+\frac{1}{2}}) + \nabla g(y_k^{\text{md}})\right)
\end{cases}
$$

With the new scaling and restarting, we obtain Algorithm 2, which we refer to as "AG-OG with restarting." The iteration complexity of AG-OG with restarting is stated in the following Corollary 3.4.

Corollary 3.4. *Algorithm 2 on problem* (1.2) *with* $K_n = \left\lceil \sqrt{8e\frac{L}{\mu}} \vee 4e\sqrt{3+\sqrt{3}}\frac{L_H}{\mu} \right\rceil$ *outputs an* ϵ*-optimal minimax point within a number* $\mathcal{O}(N)$ *of iterates, for* N *satisfying:*

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

We defer the proof of the corollary to §C.2. When restricted to the bilinear-coupled problem (1.3), Eq. (4.1) reduces to $\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \frac{\|\mathbf{B}\|_{\text{op}}}{\sqrt{\mu_f \mu_g}}\right)\log\left(\frac{1}{\epsilon}\right)$, which exactly matches the lower bound result in Zhang et al. (2021a) and therefore is optimal under this special instance.

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) AG-OG 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 $\mathcal{L}_\epsilon(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) + \epsilon||\boldsymbol{x}||^2$, where ϵ is the desired accuracy of the solution. The following Theorem 3.5 provides the complexity analysis; see §C.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 the ϵ-optimal minimax point of L within the sample 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 $\boldsymbol{x}$ and $\boldsymbol{y}$ 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 bilinear coupling case (1.3). 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 $\frac{\|\mathbf{B}\|_{\text{op}}}{\sqrt{\epsilon\mu_g}}$ matches the lower bound of bilinearly coupled concave-convex minimax optimization (Ouyang & Xu, 2021) when $f = 0$.

3.6. Application to Bilinear Games

While the complexity result for deterministic case 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. 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 $\nabla f(\boldsymbol{x}), \nabla g(\boldsymbol{y})$ being zero almost surely. Moreover, we assume the following bilinear form:

$$I(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \mathbf{B} \boldsymbol{y} + \boldsymbol{x}^\top \boldsymbol{u}_{\boldsymbol{x}} + \boldsymbol{u}_{\boldsymbol{y}}^\top \boldsymbol{y}, \tag{3.10}$$

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

$$\begin{cases} \boldsymbol{z}_{k+\frac{1}{2}} &= \boldsymbol{z}_k - \eta_k H(\boldsymbol{z}_{k-\frac{1}{2}}), & (3.11a) \\ \boldsymbol{z}_{k+1}^{\text{ag}} &= (1-\alpha_k)\boldsymbol{z}_k^{\text{ag}} + \alpha_k \boldsymbol{z}_{k+\frac{1}{2}} & (3.11b) \\ \boldsymbol{z}_{k+1} &= \boldsymbol{z}_k - \eta_k H(\boldsymbol{z}_{k+\frac{1}{2}}) \,. & (3.11c) \end{cases}$$

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 D.1, presented in D), 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 §E.8.

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

$$||\boldsymbol{z}_K^{ag} - \boldsymbol{z}^*||^2 \leq \frac{64\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})(K+1)^2}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2. \tag{3.7}$$

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

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

An extended result is also obtained in the stochastic setting; we refer interested readers to §D.2.

4. Stochastic AcceleratedGradient OptimisticGradient Descent Ascent

In this subsection, we generalize the theoretical performance of our AG-OG algorithm (Algorithm 1 and 2) to the stochastic case where the rate-optimal convergence behavior is maintained. The stochastic AG-OG algorithm replaces each batch gradient with its unbiased stochastic counterpart, with noise indices represented by ζ_t, ξ_t. The full stochastic AG-OG algorithm is shown in Algorithm 3 in §B.2.

Based on a generalized nonexpasiveness lemma (Lemma C.7, presented in §C.4) which is the stochastic analogue of Lemma 3.1, we can proceed the analysis and arrive at our stochastic result. See §C.4 for the proof.

Theorem 4.1. *Under Assumptions 2.1 and 2.2, we take* $\eta_k = \frac{k+2}{4L+D+4\sqrt{2+\sqrt{2}}L_H(k+2)}$ *where* $D = \frac{\sigma}{C}\frac{A(K)}{\sqrt{\mathbb{E}||\boldsymbol{z}_0-\boldsymbol{z}^*||^2}}$ *for* $A(K) := \sqrt{(K+1)(K+2)(2K+3)/6}$ *and some absolute constant* $C > 0$. *Then the output of Algorithm 3 on problem* (1.2) *satisfies:*

$$\begin{aligned}&\mathbb{E}||\boldsymbol{z}_K^{ag} - \boldsymbol{z}^*||^2\\ &\le \left[\frac{8L}{\mu(K+1)^2} + \frac{7.4(1+C^2)L_H}{\mu(K+1)}\right]\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2\\ &\quad + \frac{2(C+\frac{1}{C})\sigma}{\mu\sqrt{K+1}}\sqrt{\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2}.\end{aligned}$$

Remark 4.2. Without knowledge of expected initial distance $\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2$ to the true minimax point, we need an alternative selection of stepsize η_k. We assume an upper bound on $||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2$ defined as Γ_0 and let $C = \frac{\Gamma_0}{\sqrt{\mathbb{E}||\boldsymbol{z}_0-\boldsymbol{z}^*||^2}}$. The quantity $D = \frac{\sigma A(K)}{\Gamma_0}$ is hence known. Thus

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

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 3, Algorithm 2 on problem* (1.2) *outputs an* ϵ*-optimal minimax point within* $\mathcal{O}(N)$ *iterations, for* N *satisfying:*

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

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 rate-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 worst-case 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) which organically blends optimism with Nesterov's acceleration, featuring structural interpretability and simplicity. Leveraging novel Lyapunov analysis, these algorithms achieve desirable polynomial convergence behavior. Further by properly restarting the algorithms, AG-OG and its stochastic version theoretically enjoy rate-optimal sample complexity for finding an ϵ-accurate solution. Future directions include closing the gap between the upper and lower bounds for general separable minimax optimization, and generalizations to nonconvex settings.

Acknowledgements

We sincerely thank the anonymous reviewers for their helpful comments, and thank Simon Shaolei Du for highly inspiring discussions. This work is supported in part by Canada CIFAR AI Chair to GG, by the National Science Foundation CAREER Award 1906169 to QG, by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764 and also the Vannevar Bush Faculty Fellowship program under grant number N00014-21-1-2941 and NSF grant IIS-1901252 to MIJ.

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. Sbeed: 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.

Golowich, N., Pattathil, S., and Daskalakis, C. Tight last-iterate convergence rates for no-regret learning in multiplayer 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.

Metelev, 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 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 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 o (1/kˆ 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.

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.

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. Dippa: 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 of robust learning problem, we illustrate how turning the disk constraint into a penalty term allows us to obtain 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. We are provided a sequence of four-tuple $\{(s_t, a_t, r_t, s_{t+1})\}_{t=1}^n$, 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 π, 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 π in the discounted, infinite-horizon setting with discount factor $\gamma \in (0,1)$, where for each state s the discounted reward

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

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

$$\min_{\boldsymbol{x}} \frac{1}{2}\|\mathbf{A}\boldsymbol{x} - \boldsymbol{b}\|^2_{\mathbf{C}^{-1}}. \tag{A.1}$$

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

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

Applying first-order optimization directly to (A.1) would necessitate computing (and storing) the inversion of matrix $\mathbf{C}$, or at least, the matrix-vector product $\mathbf{C}^{-1}\boldsymbol{v}$ for given vector $\boldsymbol{v}$ at each step, which 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 problem (Du et al., 2017; Du & Hu, 2019):

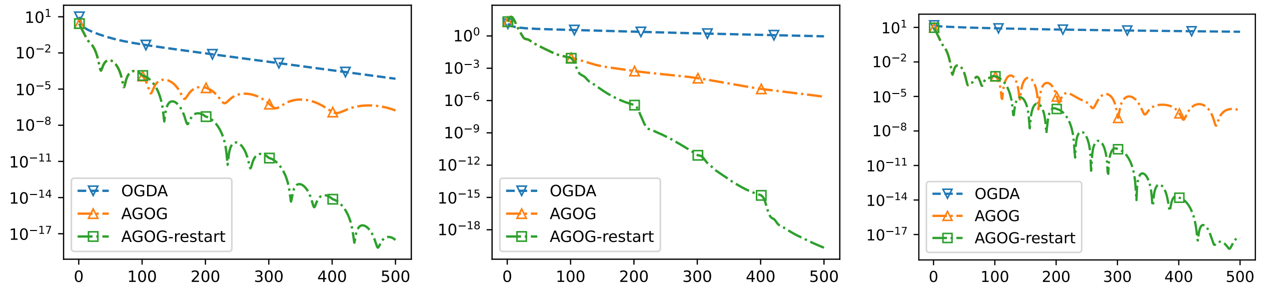
$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \; -\boldsymbol{y}^\top \mathbf{A}\boldsymbol{x} - \frac{1}{2}\|\boldsymbol{y}\|^2_{\mathbf{C}} + \boldsymbol{b}^\top \boldsymbol{y}.$$

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 quadratic regularizer term is imposed and $\mathbf{C}$ is strictly positive definite.

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

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}:\|\boldsymbol{y}-\boldsymbol{y}_0\| \leq \mathcal{R}} \frac{1}{2}||\mathbf{A}\boldsymbol{x} - \boldsymbol{y}||^2, \tag{A.2}$$

where $\mathbf{A}$ is a coefficient matrix and $\boldsymbol{y}$ 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 $\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \; \frac{1}{2}\|\mathbf{A}\boldsymbol{x} - \boldsymbol{y}\|^2 - \rho\|\boldsymbol{y} - \boldsymbol{y}_0\|^2$. When ρ is selected to be strictly greater than $\frac{1}{2}$, we get a strongly-convex-strongly-concave bilinearly coupled minimax optimization problem.



(a) $L_g = 64$, $\mu_g = 1$

(b) $L_g = 1$, $\mu_g = 1/64$

(c) $L_g = 4096$, $\mu_g = 64$

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

B. Experiments

In this section, we empirically study the performance of our AG-OG with restarting algorithm. In these experimental results, we study both deterministic [§B.1] and stochastic settings [§B.2], each of which we compare the state-of-the-art algorithms. Throughout this section, the x-axis represents the number of gradient queries while the y-axis represents the squared distance to the minimax point.

B.1. Deterministic Setting

We present results on synthetic quadratic game datasets:

$$\boldsymbol{x}^\top A_1 \boldsymbol{x} + \boldsymbol{y}^\top A_2 \boldsymbol{x} - \boldsymbol{y}^\top A_3 \boldsymbol{y}, \tag{B.1}$$

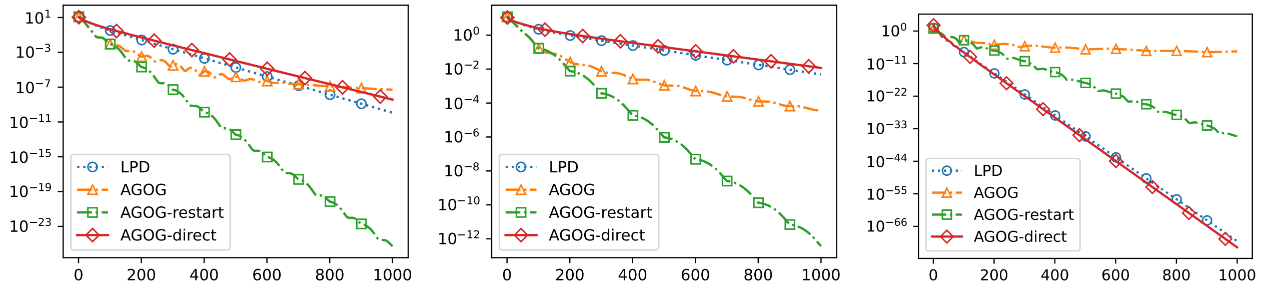
with various selections of the eigenvalues of A_1, A_2, A_3.

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 AG-OG algorithm and the AG-OG 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 AG-OG and the AG-OG with restarting algorithms and restart AG-OG with restarting once every 100 iterates. For the OGDA algorithm, we take stepsize $\eta = \frac{1}{2(L \vee L_H)}$ as is indicated by recent arts e.g. (Mokhtari et al., 2020b). For the parameters of the problem (B.1), we fix $L_H = 1, L_f = 64, \mu_f = 1$ and scatter various values of L_g, μ_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 Figures 1(a), 1(b) and 1(c) when the problem has different L_f, μ_f and L_g, μ_g, changing L_g, μ_g has larger impact on OGDA than on AG-OG, which matches our theoretical results.

Comparison with LPD. Next, we focus on comparison to the Lifted Primal-Dual (LPD) algorithm (Thekumparampil et al., 2022). We implement the AG-OG algorithm and its restarted version, the AG-OG 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 `AG-OG-Direct` that takes advantage of the strongly-convex-strongly-concave nature 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 AG-OG and AG-OG 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 AG-OG-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 AG-OG with restarting once every 100 iterates.

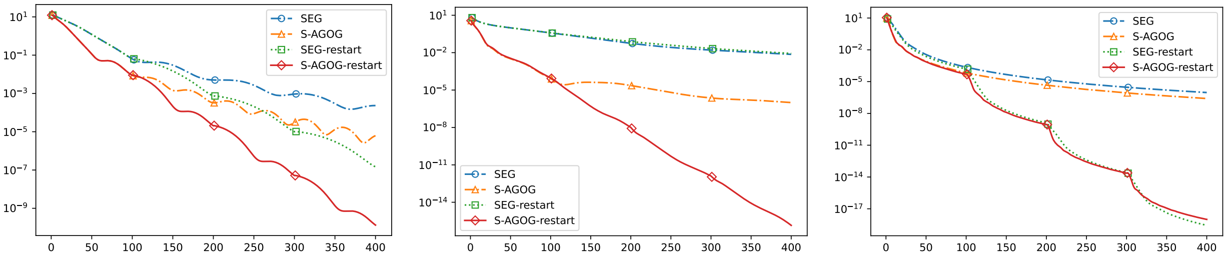
In Figure 2(a), the bilinear coupling component $\boldsymbol{y}^\top A_2 \boldsymbol{x}$ is the dominant part. In Figure 2(b), we set the eigenvalues of A_2 even larger than in Figure 2(a). In Figure 2(c), $\boldsymbol{x}^\top A_1 \boldsymbol{x}$ and $\boldsymbol{y}^\top A_3 \boldsymbol{y}$ are the dominant 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 AG-OG 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 AG-OG with restarting relative to other algorithms is enhanced. This is in accordance with our theoretical analysis. In addition, AG-OG



(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)

Algorithm 3 Stochastic AcceleratedGradient-OptimisticGradient (S-AG-OG)($\boldsymbol{z}_0^{\text{ag}}, \boldsymbol{z}_0, \boldsymbol{z}_{-1/2}, K$)

```
for $k = 0, 1, \ldots, K-1$ do
    $\boldsymbol{z}_k^{\text{md}} = (1-\alpha_k)\boldsymbol{z}_k^{\text{ag}} + \alpha_k \boldsymbol{z}_k$
    $\boldsymbol{z}_{k+\frac{1}{2}} = \boldsymbol{z}_k - \eta_k \left(\widetilde{H}(\boldsymbol{z}_{k-\frac{1}{2}}; \zeta_{k-\frac{1}{2}}) + \nabla \widetilde{F}(\boldsymbol{z}_k^{\text{md}}; \xi_k)\right)$
    $\boldsymbol{z}_{k+1}^{\text{ag}} = (1-\alpha_k)\boldsymbol{z}_k^{\text{ag}} + \alpha_k \boldsymbol{z}_{k+\frac{1}{2}}$
    $\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta_k \left(\widetilde{H}(\boldsymbol{z}_{k+\frac{1}{2}}; \zeta_{k+\frac{1}{2}}) + \nabla \widetilde{F}(\boldsymbol{z}_k^{\text{md}}; \xi_k)\right)$
end for
Output: $\boldsymbol{z}_K^{\text{ag}}$
```

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 AG-OG-direct (red line) slightly outperforms LPD. Moreover, AG-OG-direct and LPD almost overlap in 2(a) and 2(b).

B.2. Stochastic Setting

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

We use the same quadratic game setting as in (B.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 AG-OG with restarting achieves a more desirable convergence speed than SEG-restart. Also, the restarting technique significantly accelerates the convergence, validating our theory.

C. Proof of Main Convergence Results

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

C.1. Proof of Theorem 3.3

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

$$V(\boldsymbol{z}, \boldsymbol{z}') := F(\boldsymbol{z}) - F(\boldsymbol{z}') + \langle H(\boldsymbol{z}'), \boldsymbol{z} - \boldsymbol{z}' \rangle . \tag{C.1}$$

We first provide the following property for the primal-dual gap function:

Lemma C.1. *For L-smooth and μ-strongly convex $F(\boldsymbol{z})$, and for any $\boldsymbol{z} \in \mathbb{R}^{n+m}$ we have*

$$V(\boldsymbol{z}, \boldsymbol{z}^*) = F(\boldsymbol{z}) - F(\boldsymbol{z}^*) + \langle H(\boldsymbol{z}^*), \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq \frac{\mu}{2} \left\| \boldsymbol{z} - \boldsymbol{z}^* \right\|^2 . \tag{C.2}$$

Proof of Lemma C.1 is provided in §E.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 $\boldsymbol{z}_k^{\mathrm{ag}}$ and $\boldsymbol{z}^*$, whose proof is delayed to §E.4.

Lemma C.2. *For arbitrary $\alpha_k \in (0, 1]$ and any $\omega_{\boldsymbol{z}} \in \mathbb{R}^{n+m}$ the iterates of Algorithm 1 satisfy for $k = 1, \ldots, K$ almost surely*

$$V(\boldsymbol{z}_{k+1}^{ag}, \omega_{\boldsymbol{z}}) - (1 - \alpha_k) V(\boldsymbol{z}_k^{ag}, \omega_{\boldsymbol{z}}) \leq \alpha_k \underbrace{\left\langle \nabla F(\boldsymbol{z}_k^{md}) + H(\boldsymbol{z}_{k+\frac{1}{2}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}} \right\rangle}_{I} + \underbrace{\frac{L\alpha_k^2}{2} \left\| \boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k \right\|^2}_{II} . \tag{C.3}$$

Note that in Lemma C.2, the term I is an inner product that involves a gradient term.[5] The term II is brought by gradient evaluated at $\boldsymbol{z}_k^{\mathrm{md}}$.

Additionally, throughout the proof of Lemma C.2, we only leverage the convexity and L-smoothness of f and the monotonicity of H as in (2.2), as well as the update rules as in Line 2 and Line 4. The proof involves no update rules regarding the gradient updates and hence Lemma C.2 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 C.3 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 C.3 (Proposition 4.2 in Chen et al. (2017) and Lemma 7.1 in Du et al. (2022))**.** *Given an initial point $\boldsymbol{\theta} \in \mathbb{R}^d$, two update vectors $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2 \in \mathbb{R}^d$ and the corresponding outputs $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \in \mathbb{R}^d$ satisfying:*

$$\boldsymbol{\varphi}_1 = \boldsymbol{\theta} - \boldsymbol{\delta}_1, \qquad \boldsymbol{\varphi}_2 = \boldsymbol{\theta} - \boldsymbol{\delta}_2. \tag{C.4}$$

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

$$\langle \boldsymbol{\delta}_2, \boldsymbol{\varphi}_1 - \mathbf{z} \rangle \leq \frac{1}{2} \|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\|^2 + \frac{1}{2} \left[\|\boldsymbol{\theta} - \mathbf{z}\|^2 - \|\boldsymbol{\varphi}_2 - \mathbf{z}\|^2 - \|\boldsymbol{\theta} - \boldsymbol{\varphi}_1\|^2 \right] . \tag{C.5}$$

[5] In fact, this term reduces to $\langle \nabla f(\boldsymbol{z}_k), \boldsymbol{z}_k - \omega_{\boldsymbol{z}} \rangle$ of the vanilla gradient algorithm if the features of accelerations and optimistic gradients are removed.

Noting that the gradient term $\nabla F(\boldsymbol{z}_k^{\text{md}}) + H(\boldsymbol{z}_{k+\frac{1}{2}})$ in Term I of inequality (C.3) of Lemma C.2 has been used in updating $\boldsymbol{z}_k$ to $\boldsymbol{z}_{k+1}$ in Line 5 in Algorithm 1. Comparing Line 5 with Line 3 and by letting $\boldsymbol{\theta} = \boldsymbol{z}_k, \boldsymbol{\varphi}_1 = \boldsymbol{z}_{k+\frac{1}{2}}, \boldsymbol{\varphi}_2 = \boldsymbol{z}_{k+1}$ in Proposition C.3, we obtain an upper bound for the inner product term I:

$$\begin{aligned}
\eta_k \cdot \mathrm{I} &\le \frac{\eta_k^2}{2}||H(\boldsymbol{z}_{k+\frac{1}{2}}) - H(\boldsymbol{z}_{k-\frac{1}{2}})||^2 + \frac{1}{2}\left[||\boldsymbol{z}_k - \omega_{\boldsymbol{z}}||^2 - ||\boldsymbol{z}_{k+1} - \omega_{\boldsymbol{z}}||^2 - ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2\right] \\
&\le \frac{L_H^2\eta_k^2}{2}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 + \frac{1}{2}\left[||\boldsymbol{z}_k - \omega_{\boldsymbol{z}}||^2 - ||\boldsymbol{z}_{k+1} - \omega_{\boldsymbol{z}}||^2 - ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2\right],
\end{aligned} \tag{C.6}$$

where the last inequality is due to properties of H and the definition of L_H. Combining Eqs. (C.3) and (C.6) we obtain

$$\begin{aligned}
V(\boldsymbol{z}_{k+1}^{\text{ag}}, \omega_{\boldsymbol{z}}) - (1-\alpha_k)V(\boldsymbol{z}_k^{\text{ag}}, \omega_{\boldsymbol{z}}) &\le \frac{L_H^2\eta_k\alpha_k}{2}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 \\
+ \frac{\alpha_k}{2\eta_k}\left[||\boldsymbol{z}_k - \omega_{\boldsymbol{z}}||^2 - ||\boldsymbol{z}_{k+1} - \omega_{\boldsymbol{z}}||^2 - ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2\right] &+ \frac{L\alpha_k^2}{2}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2.
\end{aligned} \tag{C.7}$$

This finishes Step 1.

Step 2: Building and solving the recursion. We first apply the following lemma to build connections between $||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2$ and $||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2$, reducing Eq. (C.7) to the composition of sequences $\{||\boldsymbol{z}_k - \omega_{\boldsymbol{z}}||^2\}_{0\le k\le K-1}$ and $\{||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2\}_{0\le k\le K-1}$. The proof of Lemma C.4 is deferred to §E.5.

Lemma C.4. *For any stepsize sequence $\{\eta_k\}_{0\le k\le K-1}$ satisfying for some positive constant $c > 0$ and the Lipschitz parameter L_H such that $L_H\eta_k \le \sqrt{\frac{c}{2}}$ holds for all k. Algorithm 1 with initialization $\boldsymbol{z}_{-\frac{1}{2}} = \boldsymbol{z}_0^{ag} = \boldsymbol{z}_0$ gives the following for any $k = 0, \ldots, K-1$:*

$$\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}\right\|^2 \le 2c^k \sum_{\ell=0}^{k} c^{-\ell}\left\|\boldsymbol{z}_{\ell+\frac{1}{2}} - \boldsymbol{z}_\ell\right\|^2. \tag{C.8}$$

Combining Eqs. (C.7) and (C.8), bringing in the stepsize choice $\alpha_k = \frac{2}{k+2}$ and rearranging the terms, we obtain the following relation:

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

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

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

Taking $\eta_k = \frac{k+2}{2L + \sqrt{\frac{2}{c}}L_H(k+2)}$, we have $\frac{k+2}{\eta_k} - 2L = \sqrt{\frac{2}{c}}L_H(k+2)$, and the previous inequality reduces to

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

Subtracting off term $V(\boldsymbol{z}_{k+1}^{\text{ag}}, \omega_{\boldsymbol{z}})$ on both sides and summing over k from 0 to $K-1$, we have

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

Simple algebra yields

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

Straightforward derivations give that if we choose $c = \frac{2}{3+\sqrt{3}}$, the inequality $\sqrt{\frac{2}{c}}(k+2) \ge \sqrt{2c}\left[\frac{k+2}{1-c} + \frac{c}{(1-c)^2}\right]$ holds for all $k \ge 0$. Thus, summing III_1 and III_2 terms we have

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

Finally, we solve the recursion and conclude

$$\begin{aligned}
&\left[(K+1)^2 - 1\right] V(\boldsymbol{z}_K^{\text{ag}}, \omega_{\boldsymbol{z}}) + \left(2L + \sqrt{\frac{2}{c}} L_H (K+1)\right) \|\boldsymbol{z}_K - \omega_{\boldsymbol{z}}\|^2 \\
&\le \left(2L + \sqrt{\frac{2}{c}} L_H\right) ||\boldsymbol{z}_0 - \omega_{\boldsymbol{z}}||^2 + \sqrt{\frac{2}{c}} L_H \sum_{k=0}^{K-1} \|\boldsymbol{z}_k - \omega_{\boldsymbol{z}}\|^2 - \sum_{k=0}^{K-1} V(\boldsymbol{z}_{k+1}^{\text{ag}}, \omega_{\boldsymbol{z}}). \qquad \text{(C.9)}
\end{aligned}$$

finishing Step 2.

Step 3: Proving $\boldsymbol{z}_k$ stays nonexpansive with respect to $\boldsymbol{z}^*$. In Lemma 3.1, we show that $\boldsymbol{z}_k$ always stays in the ball centered at $\boldsymbol{z}^*$ with radius $||\boldsymbol{z}_0 - \boldsymbol{z}^*||$. The proof of this lemma is presented in §E.2.

Lemma 3.1 (Nonexpansiveness, restated)**.** 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 $\alpha_k = \frac{2}{k+2}$ Algorithm 1 with initialization $\boldsymbol{z}_{-\frac{1}{2}} = \boldsymbol{z}_0^{\text{ag}} = \boldsymbol{z}_0$, at any iterate $k < K$ we have

$$||\boldsymbol{z}_k - \boldsymbol{z}^*|| \le ||\boldsymbol{z}_0 - \boldsymbol{z}^*||.$$

Step 4: Combining everything together. Bringing the nonexpansiveness result in Lemma 3.1 into the solved recursion (C.9), setting $\omega_{\boldsymbol{z}} = \boldsymbol{z}^*$ and rearranging, we obtain the following:

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

Dividing both sides by $(K+1)^2$ and noting that Lemma C.1 implies $V(\boldsymbol{z}_K^{\text{ag}}, \boldsymbol{z}^*) \ge \frac{\mu}{2}||\boldsymbol{z}_K^{\text{ag}} - \boldsymbol{z}^*||^2$. Hence, bringing in the choice of $c = \frac{2}{3+\sqrt{3}}$ concludes our proof of Theorem 3.3. □

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.

C.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 $||\boldsymbol{z}_K^{\text{ag}} - \boldsymbol{z}^*||^2 \le \frac{1}{e}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2$ to hold, we can choose K such that

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

This is equivalent to

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

For a given threshold $\epsilon > 0$, with the output of every epoch satisfying $||\boldsymbol{z}_K^{\text{ag}} - \boldsymbol{z}^*||^2 \le \frac{1}{e}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2$, the total epochs required to obtain an ϵ-optimal minimax point would be $\log\left(\frac{||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2}{\epsilon}\right)$. Thus, the total number of iterates required to get within the ϵ 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. □

C.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:

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

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

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

Similarly for the regularized problem we define

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

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

$$\mathcal{L}_\epsilon(\widehat{\boldsymbol{x}}, \boldsymbol{y}) - \mathcal{L}_\epsilon(\boldsymbol{x}, \widehat{\boldsymbol{y}}) \le V_\epsilon(\widehat{\boldsymbol{z}}, \boldsymbol{z}),$$

and moreover,

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

By applying the AG-OG algorithm (Algorithm 1) onto the regularized objective $\mathcal{L}_\epsilon$, if we can find a pair $\widehat{\boldsymbol{z}} = (\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})$ such that

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

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

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

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

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

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

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

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

$$
(K+2)(K+1)V_\epsilon(\boldsymbol{z}_K^{\mathrm{ag}}, \boldsymbol{\omega}_{\boldsymbol{z}}) + \left(2L + \sqrt{\frac{2}{c}}L_H K\right) \|\boldsymbol{z}_K - \boldsymbol{\omega}_{\boldsymbol{z}}\|^2 \le 2L||\boldsymbol{z}_0 - \boldsymbol{\omega}_{\boldsymbol{z}}||^2 + \sqrt{\frac{2}{c}}L_H \sum_{k=0}^{K-1} \|\boldsymbol{z}_k - \boldsymbol{\omega}_{\boldsymbol{z}}\|^2. \tag{C.10}
$$

Taking $\boldsymbol{\omega}_{\boldsymbol{z}} = \boldsymbol{z}_\epsilon^*$ where $\boldsymbol{z}_\epsilon^*$ 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 μ_ϵ being the strongly convexity parameter of V_ϵ, L_ϵ being the smoothness parameter of the regularized F, the following inequality

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

Applying the same restarting as in Corollary 3.4, the total number of iterates required to get within the ϵ threshold (in terms of $||\boldsymbol{z}_K^{\mathrm{ag}} - \boldsymbol{z}_\epsilon^*||^2$) 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 $\boldsymbol{z}_0$ such that $||\boldsymbol{z}_0 - \boldsymbol{z}_\epsilon^*||^2 \le \epsilon$. We then analyze at iteration $n = N-1$. Again from Equation (C.10), letting $\boldsymbol{\omega}_{\boldsymbol{z}} := \boldsymbol{z}_K^* = \arg\max_{\boldsymbol{z}\in\mathcal{Z}} V_\epsilon(\boldsymbol{z}_K^{\mathrm{ag}}, \boldsymbol{z})$, we have

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

As $V_\epsilon(\boldsymbol{z}_K^{\mathrm{ag}}, \boldsymbol{z}_K^*) \ge 0$, we can apply the same boostrapping argument and derive

$$
\frac{\mu_\epsilon}{2}||\boldsymbol{z}_K^{\mathrm{ag}} - \boldsymbol{z}_K^*||^2 \le V_\epsilon(\boldsymbol{z}_K^{\mathrm{ag}}, \boldsymbol{z}_K^*) \le \frac{2L_\epsilon + \sqrt{3+\sqrt{5}}L_H K}{K(K+1)}||\boldsymbol{z}_0 - \boldsymbol{z}_K^*||^2. \tag{C.11}
$$

On the other hand, we also have that

$$\frac{\mu_\epsilon}{2}||\boldsymbol{z}_K^{\text{ag}} - \boldsymbol{z}_\epsilon^*||^2 \le V_\epsilon(\boldsymbol{z}_K^{\text{ag}}, \boldsymbol{z}_\epsilon^*) \le \frac{2L + \sqrt{3+\sqrt{5}}L_H K}{K(K+1)}||\boldsymbol{z}_0 - \boldsymbol{z}_\epsilon^*||^2. \tag{C.12}$$

By analyzing the two inequalities (C.11) and (C.12), 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 $\frac{4L+2\sqrt{3+\sqrt{5}}L_H K}{\mu_\epsilon K(K+1)} \le \frac{1}{c_2}$,

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

Furthermore, we apply (C.11) again and derive

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

Taking $c_2 = 3$, and noting that we have obtained $||\boldsymbol{z}_0 - \boldsymbol{z}_\epsilon^*||^2 \le \epsilon$ 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 $\max_{y\in\mathcal{Y}} \mathcal{L}_\epsilon\left(\widehat{\boldsymbol{x}}, \boldsymbol{y}\right) - \min_{\boldsymbol{x}\in\mathcal{X}} \mathcal{L}_\epsilon\left(\boldsymbol{x}, \widehat{\boldsymbol{y}}\right) \le \max_{z\in\mathcal{Z}} V_\epsilon(\widehat{\boldsymbol{z}}, \boldsymbol{z}) \le \epsilon$ 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 ϵ after scaling reduction) gives a final complexity of

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

□

C.4. Proof of Theorem 4.1

Proof.[Proof of Theorem 4.1] For the stochastic case, we use the primal-dual gap function (C.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 C.2 holds for the stochastic case as well. Thus, we have

$$V(\boldsymbol{z}_{k+1}^{\text{ag}}, \omega_{\boldsymbol{z}}) - (1-\alpha_k)V(\boldsymbol{z}_k^{\text{ag}}, \omega_{\boldsymbol{z}}) \le \alpha_k \underbrace{\left\langle \nabla F(\boldsymbol{z}_k^{\text{md}}) + H(\boldsymbol{z}_{k+\frac{1}{2}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}} \right\rangle}_{\text{I}} + \underbrace{\frac{L\alpha_k^2}{2}\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k\right\|^2}_{\text{II}}. \tag{C.3}$$

By applying Proposition C.3 to the iterates of Algorithm 3. Taking $\boldsymbol{x} = \boldsymbol{z}_k, \boldsymbol{\phi}_1 = \boldsymbol{z}_{k+\frac{1}{2}}, \boldsymbol{\phi}_2 = \boldsymbol{z}_{k+1}$ in Proposition C.3 and recalling the update rules in Algorithm 3, we obtain the following stochastic version of inequality (C.6):

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

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 $||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2$ with $||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2$ 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

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

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 §E.6.

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

$$\mathbb{E}||\widetilde{H}(\boldsymbol{z}_{k+\frac{1}{2}};\zeta_{k+\frac{1}{2}}) - \widetilde{H}(\boldsymbol{z}_{k-\frac{1}{2}};\zeta_{k-\frac{1}{2}})||^2 \le (1+\beta)L_H^2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 + \left(2 + \frac{1}{\beta}\right)\sigma_H^2. \tag{C.13}$$

Taking $\beta = 1$ in Lemma C.5 and bringing the result into the expectation of (C.3), we obtain that

$$\begin{aligned}
&\mathbb{E}V(\boldsymbol{z}_{k+1}^{\mathrm{ag}}, \omega_{\boldsymbol{z}}) - (1-\alpha_k)\mathbb{E}V(\boldsymbol{z}_k^{\mathrm{ag}}, \omega_{\boldsymbol{z}}) \\
&\le \frac{\alpha_k\eta_k}{2}\left[2L_H^2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 + 3\sigma_H^2\right] + \alpha_k\mathbb{E}\left\langle \Delta_f^k + \Delta_h^{k+\frac{1}{2}}, \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}}\right\rangle \\
&\quad + \frac{L\alpha_k^2}{2}\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + \frac{\alpha_k}{2\eta_k}\mathbb{E}\left[||\boldsymbol{z}_k - \omega_{\boldsymbol{z}}||^2 - ||\boldsymbol{z}_{k+1} - \omega_{\boldsymbol{z}}||^2 - ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2\right].
\end{aligned} \tag{C.14}$$

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 C.6, whose proof is delayed to §E.7.

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

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

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

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

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

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

Telescoping over $k = 0, 1, \dots K-1$ and using the same techniques as in the proof of Theorem 3.3, we have for $\frac{k+2}{2\eta_k} \ge 2L + \frac{1}{\sqrt{c}}L_H(k+2)$ 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

$$\begin{aligned}
&\left[(K+1)^2 - 1\right]\mathbb{E}V(\boldsymbol{z}_K^{\mathrm{ag}}, \boldsymbol{z}^*) + \frac{K+1}{\eta_{K-1}}\mathbb{E}||\boldsymbol{z}_K - \boldsymbol{z}^*||^2 \\
&\le \frac{2}{\eta_0}\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{2}{\sqrt{c}}L_H\sum_{k=1}^{K-1}\mathbb{E}\left\|\boldsymbol{z}_k - \boldsymbol{z}^*\right\|^2 + \sum_{k=0}^{K-1}(k+2)\eta_k\sigma^2 - \sum_{k=0}^{K-1}\mathbb{E}V(\boldsymbol{z}_{k+1}^{\mathrm{ag}}, \boldsymbol{z}^*).
\end{aligned} \tag{C.15}$$

Step 3: Proving z_k stays within a neighbourhood of $\boldsymbol{z}^*$. We introduce the following Lemma C.7, whose proof is in §E.3

Lemma C.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 $\forall k \in [K-1]$:

$$||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \le ||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{\eta_0}{2} \sum_{k=0}^{K-1} (k+2)\eta_k \sigma^2.$$

(b) In addition if $\eta_k \le \frac{k+2}{D}$ 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

$$||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \le ||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{A(K)^2 \sigma^2}{D^2}. \tag{C.16}$$

(c) Taking $D = \frac{\sigma}{C} \frac{A(K)}{\sqrt{\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2}}$ for some absolute constant $C > 0$, bound (C.16) *reduces to*

$$||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \le \left(1 + C^2\right) ||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2. \tag{C.17}$$

Step 4: Combining everything together. Combining the choice of stepsize η_k in (a), (b) in Lemma C.7 and $\frac{k+2}{2\eta_k} \ge 2L + \frac{1}{\sqrt{c}} L_H(k+2)$, and bound (C.15) with Eq. (C.17), by rearranging the terms again, we conclude that for $\eta_k = \frac{k+2}{4L+D+4\sqrt{2+\sqrt{2}}L_H(k+2)}$,

$$(K+1)^2 \mathbb{E} V(\boldsymbol{z}_K^{\text{ag}}, \boldsymbol{z}^*) \le \left(4L + 2\sqrt{2+\sqrt{2}}(K+1)\left(1+C^2\right) L_H\right) \mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2$$
$$+ \left(C + \frac{1}{C}\right) \sigma A(K) \sqrt{\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2}.$$

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

$$\mathbb{E}||\boldsymbol{z}_K^{\text{ag}} - \boldsymbol{z}^*||^2 \le \left[\frac{8L}{\mu(K+1)^2} + \frac{7.4(1+C^2)L_H}{\mu(K+1)}\right] \mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{2(C+\frac{1}{C})\sigma}{\mu\sqrt{K+1}} \sqrt{\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2},$$

hence concluding the entire proof of Theorem 4.1.

□

D. Proof of Bilinear Game Cases

D.1. Proof of Theorem 3.6

Proof.[Proof of Theorem 3.6]

Step 1: Non-expansiveness of OGDA last-iterate. We start by the non-expansiveness Lemma, whose proof is in §E.8.

Lemma D.1 (Bounded Iterates). *Following* (3.11)*, at any iterate $k < K$, $\boldsymbol{z}_k$ stays within the region defined by the initialization $\boldsymbol{z}_0$:*

$$||\boldsymbol{z}_k - \boldsymbol{z}^*|| \le ||\boldsymbol{z}_0 - \boldsymbol{z}^*||,$$

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

By Lemma D.1, for any $0 \le k < K$, we have

$$||\boldsymbol{z}_k - \boldsymbol{z}^*|| \le ||\boldsymbol{z}_0 - \boldsymbol{z}^*||.$$

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

$$\boldsymbol{z}_{k+1}^{\text{ag}} = \frac{k}{k+2}\boldsymbol{z}_k^{\text{ag}} + \frac{2}{k+2}\boldsymbol{z}_{k+\frac{1}{2}}.$$

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

$$(k+1)(k+2)\left(\boldsymbol{z}_{k+1}^{\text{ag}} - \boldsymbol{z}^*\right) = k(k+1)\left(\boldsymbol{z}_k^{\text{ag}} - \boldsymbol{z}^*\right) + 2(k+1)\left(\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}^*\right).$$

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

$$K(K+1)\left(\boldsymbol{z}_K^{\text{ag}} - \boldsymbol{z}^*\right) = 2\sum_{k=0}^{K-1}(k+1)\left(\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}^*\right). \tag{D.1}$$

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

$$Kz_K - \sum_{k=0}^{K-1} z_k = \sum_{k=0}^{K-1}(k+1)\boldsymbol{z}_{k+1} - (k+1)\boldsymbol{z}_k = \sum_{k=0}^{K-1}\eta_k(k+1)H(\boldsymbol{z}_{k+\frac{1}{2}}). \tag{D.2}$$

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

$$\begin{aligned}||K(K+1)\left(\boldsymbol{z}_K^{\text{ag}} - \boldsymbol{z}^*\right)||^2 &= ||2\sum_{k=0}^{K-1}(k+1)\left(\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}^*\right)||^2 \le \frac{1}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}||2\sum_{k=0}^{K-1}(k+1)H(\boldsymbol{z}_{k+\frac{1}{2}})||^2 \\ &= \frac{16L_H^2}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}||K\boldsymbol{z}_K - \sum_{k=0}^{K-1}\boldsymbol{z}_k||^2 = \frac{16L_H^2}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}||\sum_{k=0}^{K-1}\left[(\boldsymbol{z}_K - \boldsymbol{z}^*) - (\boldsymbol{z}_k - \boldsymbol{z}^*)\right]||^2 \\ &\le \frac{16L_H^2}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}K\cdot\sum_{k=0}^{K-1}\left[2||\boldsymbol{z}_K - \boldsymbol{z}^*||^2 + 2||\boldsymbol{z}_k - \boldsymbol{z}^*||^2\right].\end{aligned} \tag{D.3}$$

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

$$||\boldsymbol{z}_K^{\text{ag}} - \boldsymbol{z}^*||^2 \le \frac{64\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})(K+1)^2}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2.$$

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 $\log\left(\frac{||\boldsymbol{z}_0 - \boldsymbol{z}^*||}{\epsilon}\right)$ times, we obtain the final sample complexity of

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

for the nonstochastic setting. □

D.2. Stochastic Bilinear Game Case

For the stochastic AG-OG with restarting for bilinear case, our iteration spells

$$\begin{cases} \boldsymbol{z}_{k+\frac{1}{2}} &= \boldsymbol{z}_k - \eta_k\widetilde{H}(\boldsymbol{z}_{k-\frac{1}{2}};\zeta_{k-\frac{1}{2}}), & \text{(D.4a)} \\ \boldsymbol{z}_{k+1}^{\text{ag}} &= (1-\alpha_k)\boldsymbol{z}_k^{\text{ag}} + \alpha_k\boldsymbol{z}_{k+\frac{1}{2}}, & \text{(D.4b)} \\ \boldsymbol{z}_{k+1} &= \boldsymbol{z}_k - \eta_k\widetilde{H}(\boldsymbol{z}_{k+\frac{1}{2}};\zeta_{k+\frac{1}{2}}). & \text{(D.4c)} \end{cases}$$

We are able to derive the following theorem.

Theorem D.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* (D.4) *satisfies for any $\gamma > 0$,*

$$\mathbb{E}||\boldsymbol{z}_K^{ag} - \boldsymbol{z}^*||^2 \le (1+\gamma)\left(\frac{128L_H^2}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})(K+1)^2}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{48}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})(K+1)}\sigma_H^2\right) + \frac{4(1+\frac{1}{\gamma})\sigma_H^2}{3\lambda_{\min}(\mathbf{B}^\top\mathbf{B})K}.$$

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

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

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

Lemma D.3 (Bounded Iterates (Stochastic))**.** *Following* (D.4)*, for any $\beta > 0$, taking $\eta_k = \frac{1}{2L_H\sqrt{(1+\beta)}}$ at any iterate $k < K$, $\boldsymbol{z}_k$ stays within the region defined by the initialization $\boldsymbol{z}_0$:*

$$||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \le ||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \eta_k^2\left(2 + \frac{1}{\beta}\right)K\sigma_H^2,$$

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

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

$$\left\langle H(\boldsymbol{z}_{k+\frac{1}{2}}) - H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}}\right\rangle \ge 0, \qquad \forall \omega_{\boldsymbol{z}} \in \mathcal{Z}. \tag{D.5}$$

Let $\boldsymbol{\varphi}_1 = \boldsymbol{z}_{k+\frac{1}{2}}$, $\boldsymbol{\varphi}_2 = \boldsymbol{z}_{k+1}$, $\boldsymbol{\theta} = \boldsymbol{z}_k$, $\boldsymbol{\delta}_1 = \eta_k\widetilde{H}(\boldsymbol{z}_{k-\frac{1}{2}};\zeta_{k-\frac{1}{2}})$, $\boldsymbol{\delta}_2 = \eta_k\widetilde{H}(\boldsymbol{z}_{k+\frac{1}{2}};\zeta_{k+\frac{1}{2}})$ and $\boldsymbol{z} = \omega_z$ in Proposition C.3, we have

$$\begin{aligned}&\eta_k\left\langle \widetilde{H}(\boldsymbol{z}_{k+\frac{1}{2}};\zeta_{k+\frac{1}{2}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_z\right\rangle \\ &\le \frac{\eta_k^2}{2}||\widetilde{H}(\boldsymbol{z}_{k+\frac{1}{2}};\zeta_{k+\frac{1}{2}}) - \widetilde{H}(\boldsymbol{z}_{k-\frac{1}{2}};\zeta_{k-\frac{1}{2}})||^2 + \frac{1}{2}\left[||\boldsymbol{z}_k - \omega_z||^2 - ||\boldsymbol{z}_{k+1} - \omega_z||^2 - ||\boldsymbol{z}_k - \boldsymbol{z}_{k+\frac{1}{2}}||^2\right].\end{aligned} \tag{D.6}$$

Recalling the results in Lemma C.5 that

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

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

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

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

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

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

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

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

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

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

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

Rearranging the above inequality and we conclude that

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

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

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

which concludes our proof of Lemma D.3. □

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

$$\boldsymbol{z}_{k+1}^{\mathrm{ag}} = \frac{k}{k+2}\boldsymbol{z}_k^{\mathrm{ag}} + \frac{2}{k+2}\boldsymbol{z}_{k+\frac{1}{2}}.$$

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

$$(k+1)(k+2)\left(\boldsymbol{z}_{k+1}^{\mathrm{ag}} - \boldsymbol{z}^*\right) = k(k+1)\left(\boldsymbol{z}_k^{\mathrm{ag}} - \boldsymbol{z}^*\right) + 2(k+1)\left(\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}^*\right).$$

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

$$K(K+1)\left(\boldsymbol{z}_K^{\mathrm{ag}} - \boldsymbol{z}^*\right) = 2\sum_{k=0}^{K-1}(k+1)\left(\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}^*\right). \tag{D.8}$$

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

$$\begin{aligned}
Kz_K - \sum_{k=0}^{K-1} z_k = \sum_{k=0}^{K-1}(k+1)\boldsymbol{z}_{k+1} - (k+1)\boldsymbol{z}_k &= \sum_{k=0}^{K-1}\eta_k(k+1)\widetilde{H}(\boldsymbol{z}_{k+\frac{1}{2}}, \zeta_{k+\frac{1}{2}}) \\
&= \sum_{k=0}^{K-1}\eta_k(k+1)\left[H(\boldsymbol{z}_{k+\frac{1}{2}}) + \Delta_h^{k+\frac{1}{2}}\right].
\end{aligned} \tag{D.9}$$

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

$$\begin{aligned}
&||K(K+1)\left(\boldsymbol{z}_K^{\text{ag}} - \boldsymbol{z}^*\right)||^2 \\
&= ||2\sum_{k=0}^{K-1}(k+1)\left(\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}^*\right)||^2 \le \frac{1}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}||2\sum_{k=0}^{K-1}(k+1)H(\boldsymbol{z}_{k+\frac{1}{2}})||^2 \\
&= \frac{16(1+\beta)(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}||K\boldsymbol{z}_K - \sum_{k=0}^{K-1}\boldsymbol{z}_k||^2 + \frac{4(1+\frac{1}{\gamma})}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}||\sum_{k=0}^{K-1}(k+1)\Delta_h^{k+\frac{1}{2}}||^2 \\
&\le \frac{16(1+\beta)(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}K\cdot\sum_{k=0}^{K-1}\left[2||\boldsymbol{z}_K - \boldsymbol{z}^*||^2 + 2||\boldsymbol{z}_k - \boldsymbol{z}^*||^2\right] + \frac{4(1+\frac{1}{\gamma})}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})}||\sum_{k=0}^{K-1}(k+1)\Delta_h^{k+\frac{1}{2}}||^2. \qquad \text{(D.10)}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}||\boldsymbol{z}_K^{\text{ag}} - \boldsymbol{z}^*||^2 &\le \frac{16(1+\beta)(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})K(K+1)^2}\sum_{k=0}^{K-1}\mathbb{E}\left[2||\boldsymbol{z}_K - \boldsymbol{z}^*||^2 + 2||\boldsymbol{z}_k - \boldsymbol{z}^*||^2\right] \\
&\quad + \frac{4(1+\frac{1}{\gamma})}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})K^2(K+1)^2}\sum_{k=0}^{K-1}(k+1)^2\mathbb{E}||\Delta_h^{k+\frac{1}{2}}||^2 \\
&\le \frac{64(1+\beta)(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})(K+1)^2}\left[||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \eta_k^2(2+\frac{1}{\beta})K\sigma_H^2\right] + \frac{4(1+\frac{1}{\gamma})\sigma_H^2}{3\lambda_{\min}(\mathbf{B}^\top\mathbf{B})K} \\
&= \frac{64(1+\beta)(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})(K+1)^2}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{16(1+\gamma)(2+\frac{1}{\beta})}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})(K+1)}\sigma_H^2 + \frac{4(1+\frac{1}{\gamma})\sigma_H^2}{3\lambda_{\min}(\mathbf{B}^\top\mathbf{B})K},
\end{aligned}$$

Minimize over β, we have $\beta = \frac{\sigma_H\sqrt{K+1}}{2L_H||\boldsymbol{z}_0 - \boldsymbol{z}^*||}$ and the first two terms become

$$\frac{64(1+\gamma)L_H^2}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})(K+1)^2}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{32(1+\gamma)}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})(K+1)}\sigma_H^2 + \frac{32(1+\gamma)L_H\sigma_H}{\lambda_{\min}(\mathbf{B}^\top\mathbf{B})(K+1)^{3/2}}||\boldsymbol{z}_0 - \boldsymbol{z}^*||$$

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

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

Further minimizing over γ, we conclude that so

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

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

E. Proof of Auxiliary Lemmas

E.1. Proof of Lemma C.1

Proof.[Proof of Lemma C.1]

Since $F(\boldsymbol{z})$ is L-smooth and μ-strongly convex. For the rest of this proof, we observe that the saddle definition of $\boldsymbol{z}^*$ satisfies the first-order stationary condition:

$$\nabla F(\boldsymbol{z}^*) + H(\boldsymbol{z}^*) = 0. \tag{E.1}$$

Furthermore, we have

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

where in both of the two displays, the inequality holds due to the μ-strong convexity of F, and the equality holds due to the first-order stationary condition (E.1). This completes the proof. □

E.2. Proof of Lemma 3.1

Proof.[Proof of Lemma 3.1] Following (C.9), we let $\omega_{\boldsymbol{z}} = \boldsymbol{z}^*$. Due to the non-negativity of $V(\cdot, \boldsymbol{z}^*)$, we can eliminate the V terms and have:

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

We adopt a "bootstrapping" argument.We define $M_K = \max_{0 \leq k \leq K-1} ||\boldsymbol{z}_k - \boldsymbol{z}^*||^2$ and taking a maximum on each term on the right hand side of the above inequality, we conclude that

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

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

E.3. Proof of Lemma C.7

Proof.[Proof of Lemma C.7] Starting from (C.15) that

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

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

$$\frac{K+1}{\eta_{K-1}} \mathbb{E}||\boldsymbol{z}_K - \boldsymbol{z}^*||^2 \leq \frac{2}{\eta_0} \mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{2}{\sqrt{c}} L_H \sum_{k=1}^{K-1} ||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 + \sum_{k=0}^{K-1} (k+2) \eta_k \sigma^2. \tag{E.2}$$

Rewrite $||\boldsymbol{z}_K - \boldsymbol{z}^*||^2$ as the difference between two summations, we obtain:

$$\frac{K+1}{\eta_{K-1}}\left(\sum_{k=1}^{K} - \sum_{k=1}^{K-1}\right)\mathbb{E}||\boldsymbol{z}_k - \omega_{\boldsymbol{z}}||^2 \le \frac{2}{\eta_0}\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{2}{\sqrt{c}}L_H\sum_{k=1}^{K-1}\mathbb{E}||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 + \sum_{k=0}^{K-1}(k+2)\eta_k\sigma^2.$$

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:

$$\frac{K+1}{\eta_{K-1}}\sum_{k=1}^{K}\mathbb{E}||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \le \frac{2}{\eta_0}\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{K+2}{\eta_K}\sum_{k=1}^{K-1}\mathbb{E}||\boldsymbol{z}_k - \omega_{\boldsymbol{z}}||^2 + \sum_{k=0}^{K-1}(k+2)\eta_k\sigma^2.$$

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:

$$\frac{\eta_K}{K+2}\sum_{k=1}^{K}\mathbb{E}||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \le \frac{\eta_{K-1}}{K+1}\sum_{k=1}^{K-1}\mathbb{E}||\boldsymbol{z}_k - \omega_{\boldsymbol{z}}||^2 + \frac{\eta_{K-1}\eta_K}{(K+1)(K+2)}\left[\frac{2}{\eta_0}\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \sum_{k=0}^{K-1}(k+2)\eta_k\sigma^2\right].$$

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, $\sum_{k=0}^{K-1}(k+2)\eta_k\sigma^2 \le \sum_{k=0}^{\mathcal{K}-1}(k+2)\eta_k\sigma^2$ is an upper bound that does not change with the choice of K. It follows that:

$$\begin{aligned}\frac{\eta_K}{K+2}\sum_{k=1}^{K}\mathbb{E}||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 &\le \frac{\eta_{K-1}}{K+1}\sum_{k=1}^{K-1}\mathbb{E}||\boldsymbol{z}_k - \omega_{\boldsymbol{z}}||^2 + \frac{\sqrt{c}}{2L_H}\left[\frac{\eta_{K-1}}{K+1} - \frac{\eta_K}{K+2}\right]\left[\frac{2}{\eta_0}\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \sum_{k=0}^{\mathcal{K}-1}(k+2)\eta_k\sigma^2\right]\\ &\le \frac{\sqrt{c}}{2L_H}\left[\frac{\eta_0}{2} - \frac{\eta_K}{K+2}\right]\left[\frac{2}{\eta_0}\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \sum_{k=0}^{\mathcal{K}-1}(k+2)\eta_k\sigma^2\right].\end{aligned}$$

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

$$\sum_{k=1}^{K}\mathbb{E}||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \le \frac{\sqrt{c}}{2L_H}\left[\frac{\eta_0(K+2)}{2\eta_K} - 1\right]\left[\frac{2}{\eta_0}\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \sum_{k=0}^{\mathcal{K}-1}(k+2)\eta_k\sigma^2\right].$$

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

$$\frac{K+1}{\eta_{K-1}}\mathbb{E}||\boldsymbol{z}_K - \boldsymbol{z}^*||^2 \le \frac{\eta_0(K+1)}{2\eta_{K-1}}\left[\frac{2}{\eta_0}\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \sum_{k=0}^{\mathcal{K}-1}(k+2)\eta_k\sigma^2\right].$$

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

$$\mathbb{E}||\boldsymbol{z}_K - \boldsymbol{z}^*||^2 \le \frac{\eta_0}{2}\left[\frac{2}{\eta_0}\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \sum_{k=0}^{\mathcal{K}-1}(k+2)\eta_k\sigma^2\right].$$

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

$$\mathbb{E}||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \le \mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{\eta_0}{2}\sum_{k=0}^{K-1}(k+2)\eta_k\sigma^2,$$

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

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

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

$$\mathbb{E}||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \le \mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{A(K)^2\sigma^2}{D^2},$$

which concludes our proof of (b). And (c) follows by straightforward calculations. □

E.4. Proof of Lemma C.2

Proof.[Proof of Lemma C.2] 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 $\boldsymbol{z}_k^{\mathrm{md}}$. For any given $\boldsymbol{u} \in \mathcal{Z}$, the convexity and L-smoothness of $F(\cdot)$ implies that

$$\begin{aligned} F(\boldsymbol{z}_{k+1}^{\mathrm{ag}}) - F(\boldsymbol{u}) &= F(\boldsymbol{z}_{k+1}^{\mathrm{ag}}) - F(\boldsymbol{z}_k^{\mathrm{md}}) - \big(F(\boldsymbol{u}) - F(\boldsymbol{z}_k^{\mathrm{md}})\big) \\ &\le \left\langle \nabla F(\boldsymbol{z}_k^{\mathrm{md}}), \boldsymbol{z}_{k+1}^{\mathrm{ag}} - \boldsymbol{z}_k^{\mathrm{md}} \right\rangle + \frac{L}{2}\left\| \boldsymbol{z}_{k+1}^{\mathrm{ag}} - \boldsymbol{z}_k^{\mathrm{md}} \right\|^2 - \left\langle \nabla F(\boldsymbol{z}_k^{\mathrm{md}}), \boldsymbol{u} - \boldsymbol{z}_k^{\mathrm{md}} \right\rangle . \end{aligned}$$

Taking $\boldsymbol{u} = \omega_{\boldsymbol{z}}$ and $\boldsymbol{u} = \boldsymbol{z}_k^{\mathrm{ag}}$ respectively, we conclude that

$$F(\boldsymbol{z}_{k+1}^{\mathrm{ag}}) - F(\omega_{\boldsymbol{z}}) \le \left\langle \nabla F(\boldsymbol{z}_k^{\mathrm{md}}), \boldsymbol{z}_{k+1}^{\mathrm{ag}} - \boldsymbol{z}_k^{\mathrm{md}} \right\rangle + \frac{L}{2}\left\| \boldsymbol{z}_{k+1}^{\mathrm{ag}} - \boldsymbol{z}_k^{\mathrm{md}} \right\|^2 - \left\langle \nabla F(\boldsymbol{z}_k^{\mathrm{md}}), \omega_{\boldsymbol{z}} - \boldsymbol{z}_k^{\mathrm{md}} \right\rangle , \tag{E.3}$$

$$F(\boldsymbol{z}_{k+1}^{\mathrm{ag}}) - F(\boldsymbol{z}_k^{\mathrm{ag}}) \le \left\langle \nabla F(\boldsymbol{z}_k^{\mathrm{md}}), \boldsymbol{z}_{k+1}^{\mathrm{ag}} - \boldsymbol{z}_k^{\mathrm{md}} \right\rangle + \frac{L}{2}\left\| \boldsymbol{z}_{k+1}^{\mathrm{ag}} - \boldsymbol{z}_k^{\mathrm{md}} \right\|^2 - \left\langle \nabla F(\boldsymbol{z}_k^{\mathrm{md}}), \boldsymbol{z}_k^{\mathrm{ag}} - \boldsymbol{z}_k^{\mathrm{md}} \right\rangle . \tag{E.4}$$

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

$$\begin{aligned} F(\boldsymbol{z}_{k+1}^{\mathrm{ag}}) - \alpha_k F(\omega_{\boldsymbol{z}}) - (1-\alpha_k) F(\boldsymbol{z}_k^{\mathrm{ag}}) &\le \left\langle \nabla F(\boldsymbol{z}_k^{\mathrm{md}}), \boldsymbol{z}_{k+1}^{\mathrm{ag}} - (1-\alpha_k)\boldsymbol{z}_k^{\mathrm{ag}} - \alpha_k \omega_{\boldsymbol{z}} \right\rangle + \frac{L}{2} || \boldsymbol{z}_{k+1}^{\mathrm{ag}} - \boldsymbol{z}_k^{\mathrm{md}} ||^2 \\ &= \underbrace{\alpha_k \left\langle \nabla F(\boldsymbol{z}_k^{\mathrm{md}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}} \right\rangle}_{\text{I(a)}} + \underbrace{\frac{L\alpha_k^2}{2}\left\| \boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k \right\|^2}_{\text{II}}, \end{aligned} \tag{E.5}$$

where by substracting Line 2 from Line 4 of Algorithm 1 and by Line 4 itself, the last equality of (E.5) follows.

Recalling that $\boldsymbol{z}_k^{\mathrm{ag}}$ corresponds to regular iterates and $\boldsymbol{z}_k^{\mathrm{md}}$ corresponds to the extrapolated iterates of Nesterov's acceleration scheme. The squared error term II in (E.5) is brought by gradient evaluated at the extrapolated point instead of the regular point. Note that if we do an implicit version of Nesterov such that $\boldsymbol{z}_{k-1}^{\mathrm{md}} = \boldsymbol{z}_k^{\mathrm{ag}}$, 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

$$\left\langle H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_{k+1}^{\mathrm{ag}} - \omega_{\boldsymbol{z}} \right\rangle - (1-\alpha_k)\left\langle H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_k^{\mathrm{ag}} - \omega_{\boldsymbol{z}} \right\rangle = \alpha_k \left\langle H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}} \right\rangle \le \underbrace{\alpha_k \left\langle H(\boldsymbol{z}_{k+\frac{1}{2}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}} \right\rangle}_{\text{I(b)}}, \tag{E.6}$$

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

$$\left\langle H(\boldsymbol{z}_{k+\frac{1}{2}}) - H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}} \right\rangle \ge 0.$$

Summing both sides of Eq. (E.5) and Eq. (E.6) we obtain the following:

$$\begin{aligned} &F(\boldsymbol{z}_{k+1}^{\mathrm{ag}}) - \alpha_k F(\omega_{\boldsymbol{z}}) - (1-\alpha_k) F(\boldsymbol{z}_k^{\mathrm{ag}}) + \left\langle H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_{k+1}^{\mathrm{ag}} - \omega_{\boldsymbol{z}} \right\rangle - (1-\alpha_k)\left\langle H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_k^{\mathrm{ag}} - w_z \right\rangle \\ &\le \underbrace{\alpha_k \left\langle \nabla F(\boldsymbol{z}_k^{\mathrm{md}}) + H(\boldsymbol{z}_{k+\frac{1}{2}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}} \right\rangle}_{\text{I}} + \underbrace{\frac{L\alpha_k^2}{2}\left\| \boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k \right\|^2}_{\text{II}}, \end{aligned}$$

where I is the summation of I(a) and I(b). This concludes our proof of Lemma C.2 by bringing in the definitions of $V(\boldsymbol{z}_{k+1}^{\mathrm{ag}}, \boldsymbol{z}^*)$ and $V(\boldsymbol{z}_k^{\mathrm{ag}}, \boldsymbol{z}^*)$. □

E.5. Proof of Lemma C.4

Proof.[Proof of Lemma C.4] We focus on $k = 1, \dots, K-1$ since the $k = 0$ case holds automatically. By Young's inequality and Cauchy-Schwarz inequality, we have that

$$\begin{aligned}\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}\right\|^2 &\le 2\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k\right\|^2 + 2\left\|\boldsymbol{z}_k - \boldsymbol{z}_{k-\frac{1}{2}}\right\|^2 \\ &\overset{(a)}{\le} 2\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k\right\|^2 + 2\eta_{k-1}^2 L_H^2 \left\|\boldsymbol{z}_{k-\frac{1}{2}} - \boldsymbol{z}_{k-\frac{3}{2}}\right\|^2 \overset{(b)}{\le} 2\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k\right\|^2 + c\left\|\boldsymbol{z}_{k-\frac{1}{2}} - \boldsymbol{z}_{k-\frac{3}{2}}\right\|^2,\end{aligned} \tag{E.7}$$

where (a) is due to Lines 3 and 5 of Algorithm 1 and the definition of L_H, and (b) is due to the condition in Lemma C.4 that $\eta_k L_H \le \sqrt{\frac{c}{2}}$. Recursively applying the above gives (C.8) which is repeated as:

$$\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}\right\|^2 \le 2c^k \sum_{\ell=0}^{k} c^{-\ell} \left\|\boldsymbol{z}_{\ell+\frac{1}{2}} - \boldsymbol{z}_\ell\right\|^2. \tag{C.8}$$

Indeed, from (E.7)

$$c^{-k}\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}\right\|^2 - c^{-(k-1)}\left\|\boldsymbol{z}_{k-\frac{1}{2}} - \boldsymbol{z}_{k-\frac{3}{2}}\right\|^2 \le 2c^{-k}\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k\right\|^2,$$

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

$$c^{-K}\left\|\boldsymbol{z}_{K+\frac{1}{2}} - \boldsymbol{z}_{K-\frac{1}{2}}\right\|^2 - \left\|\boldsymbol{z}_{\frac{1}{2}} - \boldsymbol{z}_{-\frac{1}{2}}\right\|^2 \le 2\sum_{k=1}^{K} c^{-k}\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k\right\|^2,$$

which simply reduces to (due to $\boldsymbol{z}_0 = \boldsymbol{z}_{-\frac{1}{2}}$)

$$c^{-K}\left\|\boldsymbol{z}_{K+\frac{1}{2}} - \boldsymbol{z}_{K-\frac{1}{2}}\right\|^2 \le 2\sum_{k=1}^{K} c^{-k}\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k\right\|^2 + \left\|\boldsymbol{z}_{\frac{1}{2}} - \boldsymbol{z}_0\right\|^2 \le 2\sum_{k=0}^{K} c^{-k}\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k\right\|^2.$$

This gives (C.8) and the entire Lemma C.4. □

E.6. Proof of Lemma C.5

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

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

Then, we have

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

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

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

Recalling that by Assumption 2.2, $\mathbb{E}||\Delta_h^{k+\frac{1}{2}}||^2 \le \sigma_H^2$ and $\mathbb{E}||\Delta_h^{k-\frac{1}{2}}||^2 \le \sigma_H^2$, we conclude our proof of Lemma C.5. □

E.7. Proof of Lemma C.6

Proof.[Proof of Lemma C.6] By inequality (C.14), we have

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

The inner product term can be decomposed into

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

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

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

For any $r > 0$, we pair up

$$-\frac{(1-r)\alpha_k}{2\eta_k}\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + \alpha_k \mathbb{E}\left\langle \Delta_f^k, \boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k \right\rangle \le \frac{\alpha_k \eta_k}{2(1-r)}\mathbb{E}||\Delta_f^k||^2,$$

and thus

$$\begin{aligned}
&\mathbb{E}V(\boldsymbol{z}_{k+1}^{\text{ag}}, \omega_{\boldsymbol{z}}) - (1-\alpha_k)\mathbb{E}V(\boldsymbol{z}_k^{\text{ag}}, \omega_{\boldsymbol{z}}) \\
&\le \frac{\alpha_k \eta_k}{2}\left[2L_H^2 \mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 + 3\sigma_H^2\right] + \frac{\alpha_k \eta_k}{2(1-r)}\mathbb{E}||\Delta_f^k||^2 \\
&\quad + \frac{\alpha_k}{2\eta_k}\mathbb{E}\left[||\boldsymbol{z}_k - \omega_{\boldsymbol{z}}||^2 - ||\boldsymbol{z}_{k+1} - \omega_{\boldsymbol{z}}||^2\right] - \left(\frac{r\alpha_k}{2\eta_k} - \frac{L\alpha_k^2}{2}\right)\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2. \qquad \text{(E.8)}
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}\right\|^2 \le 2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + 2\mathbb{E}||\boldsymbol{z}_k - \boldsymbol{z}_{k-\frac{1}{2}}||^2 \\
&= 2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + 2\eta_{k-1}^2\mathbb{E}||\widetilde{H}(\boldsymbol{z}_{k-\frac{1}{2}}) - \widetilde{H}(\boldsymbol{z}_{k-\frac{3}{2}})||^2 \\
&= 2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + 2\eta_{k-1}^2\mathbb{E}||H(\boldsymbol{z}_{k-\frac{1}{2}}) - H(\boldsymbol{z}_{k-\frac{3}{2}}) + \Delta_h^{k-\frac{3}{2}}||^2 + 2\eta_{k-1}^2\mathbb{E}||\Delta_h^{k-\frac{1}{2}}||^2 \\
&\le 2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + 4\eta_{k-1}^2 L_H^2\mathbb{E}||\boldsymbol{z}_{k-\frac{1}{2}} - \boldsymbol{z}_{k-\frac{3}{2}}||^2 + 6\eta_{k-1}^2\sigma_H^2 \\
&= 2\sum_{\ell=0}^{k} c^{k-\ell}\mathbb{E}||\boldsymbol{z}_{\ell+\frac{1}{2}} - \boldsymbol{z}_\ell||^2 + 6\sum_{\ell=0}^{k} c^{k-\ell}\eta_{\ell-1}^2\sigma_H^2.
\end{aligned} \qquad \text{(E.9)}$$

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

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

which concludes our proof of Lemma C.6. □

E.8. Proof of Lemma D.1

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

$$\left\langle H(\boldsymbol{z}_{k+\frac{1}{2}}) - H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}}\right\rangle \geq 0, \quad \forall \omega_{\boldsymbol{z}} \in \mathcal{Z}. \tag{E.10}$$

Let $\boldsymbol{\varphi}_1 = \boldsymbol{z}_{k+\frac{1}{2}}$, $\boldsymbol{\varphi}_2 = \boldsymbol{z}_{k+1}$, $\boldsymbol{\theta} = \boldsymbol{z}_k$, $\boldsymbol{\delta}_1 = \eta_k H(\boldsymbol{z}_{k-\frac{1}{2}})$, $\boldsymbol{\delta}_2 = \eta_k H(\boldsymbol{z}_{k+\frac{1}{2}})$ and $\boldsymbol{z} = \omega_z$ in Proposition C.3, we have

$$\begin{aligned}
\eta_k\left\langle H(\boldsymbol{z}_{k+\frac{1}{2}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_z\right\rangle &\leq \frac{\eta_k^2}{2}||H(\boldsymbol{z}_{k+\frac{1}{2}}) - H(\boldsymbol{z}_{k-\frac{1}{2}})||^2 + \frac{1}{2}\left[||\boldsymbol{z}_k - \omega_z||^2 - ||\boldsymbol{z}_{k+1} - \omega_z||^2 - ||\boldsymbol{z}_k - \boldsymbol{z}_{k+\frac{1}{2}}||^2\right] \\
&\leq \frac{\eta_k^2 L_H^2}{2}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 + \frac{1}{2}\left[||\boldsymbol{z}_k - \omega_z||^2 - ||\boldsymbol{z}_{k+1} - \omega_z||^2 - ||\boldsymbol{z}_k - \boldsymbol{z}_{k+\frac{1}{2}}||^2\right],
\end{aligned} \tag{E.11}$$

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

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

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

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

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

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

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

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

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

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

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

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

which concludes our proof of Lemma D.1. □