OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Nesterov Meets Optimism: Rate-Optimal Optimistic-Gradient-Based Method for Stochastic Bilinearly-Coupled Minimax Optimization

Chris Junchi Li[⋆] JUNCHILI@BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
Berkeley, CA 94720

Angela Yuan[⋆] HZYUAN@CS.UCLA.EDU
Department of Computer Sciences
University of California, Los Angeles
Los Angeles, CA 90095

Gauthier Gidel GAUTHIER.GIDEL@UMONTREAL.CA
DIRO, Université de Montréal and Mila
2920 Chemin de la Tour
Montréal, Québec H3T 1J4, Canada

Michael I. Jordan JORDAN@CS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences & Department of Statistics
University of California, Berkeley
Berkeley, CA 94720

Abstract

We provide a novel first-order optimization algorithm for bilinearly-coupled strongly-convex-concave minimax optimization called the AcceleratedGradient OptimisticGradient (AG-OG). The main idea of our algorithm is to leverage the structure of the considered minimax problem and operates Nesterov's acceleration on the individual part and optimistic gradient on the coupling part of the objective. We motivate our method by showing that its continuous-time dynamics corresponds to an organic combination of the dynamics of optimistic gradient and of Nesterov's acceleration. By discretizing the dynamics we conclude polynomial convergence behavior in discrete time. Further enhancement of AG-OG with proper restarting allows us to achieve rate-optimal (up to a constant) convergence rates with respect to the conditioning of the coupling and individual parts, which results in the first single-call algorithm achieving improved convergence in the deterministic setting and rate-optimality in the stochastic setting under bilinearly coupled minimax problem sets.

1. Introduction

Optimization is the workhorse of machine learning (ML) and artificial intelligence research; indeed, many basic ML learning tasks can be cast as a minimization problem. In an increasing number of applications, however, such as generative adversarial networks (GANs) [17], robust/adversarial optimization [3, 28], Markov games (MGs) [35], and reinforcement learning (RL) [9, 11, 37], the

. *Equal contribution.

goal is to solve instead a minimax problem of the form:

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

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

Unlike in the minimization setting, there are no general convergence guarantees when simple gradient methods are used on convex-concave objectives. Indeed, there are examples showing the divergence of gradient descent ascent (GDA) on bilinear objectives [14, 25]. This has led to the development of extrapolation-based methods that include the extra-gradient (EG) method [22] and the optimistic gradient descent ascent (OGDA) method [32], both of which can be shown to converge in the convex-concave setting. While the EG algorithm summons an additional gradient oracle at each step, the OGDA algorithm can be seen as a single-call version of EG [14, 19]. In this paper, we design algorithms based on the idea of the OGDA iteration, aiming to improve its performance while retaining the convergence guarantee.

We focus on a specific instance of the general minimax problem, which we call the *bilinearly coupled strongly convex-strongly concave saddle point problem* (bi-SC-SC), formulated as follows:

$$\min_{\boldsymbol{x}\in\mathcal{X}}\max_{\boldsymbol{y}\in\mathcal{Y}} \ \mathcal{L}(\boldsymbol{x},\boldsymbol{y}) \equiv f(\boldsymbol{x}) + \boldsymbol{x}^\top \mathbf{B}\boldsymbol{y} - g(\boldsymbol{y}), \tag{1.2}$$

where f is L_f-smooth and μ_f-strongly convex, and g is L_g-smooth and μ_g-strongly convex. Here $\boldsymbol{x}^\top\mathbf{B}\boldsymbol{y}$ is called the bilinear coupling term and is L_H-smooth where $L_H \equiv \sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}$. Moreover, we consider throughout this paper the unconstrained problem where $\mathcal{X}=\mathbb{R}^n$ and $\mathcal{Y}=\mathbb{R}^m$ except specified in some application instances. When considering $L := \max(L_H, L_f, L_g)$ and $\mu = \min(\mu_f, \mu_g)$, the standard OGDA algorithm can be shown to yield a coarse-grained complexity of $\frac{L_f\vee L_g\vee L_H}{\mu_f\wedge\mu_g}\log\left(\frac{1}{\epsilon}\right)$ when applied to Problem (1.2) [14, 29]. In fact this rate is optimal *in the coarse sense* , as has been shown by Azizian et al. [2] that the minimal complexity is $\Omega(\frac{L}{\mu}\log(\frac{1}{\epsilon}))$. Nevertheless, the above convergence rate has a dependency on the parameters L_f, L_g, L_H and also μ_f, μ_g as a whole. Moreover, when reducing to an individual optimization problem $\min_{\boldsymbol{x}}\max_{\boldsymbol{y}} f(\boldsymbol{x}) - g(\boldsymbol{y})$, it does not achieve accelerated rate. This motivates and illustrates the difficulty of deriving *fine-grained* convergence rates that depend on the condition numbers of f, g and H separately and for which a notion of acceleration is possible.

Providing fine-grained rates in the *unbalanced* strongly-convex-strongly-concave setting where $\mu_x \neq \mu_y$ is of particular importance when the constants μ_x, μ_y, L_H, L_x and L_y are drastically different. For instance when μ_g is significantly larger than μ_f, decoupling the dependencies on L_f, μ_f and L_g, μ_g from the coarsened smoothness and strong convexity parameters $L_f \vee L_g, \mu_f\wedge\mu_g$ would significantly improve the theoretical complexity. In this paper, we focus on accelerating OGDA when constrained to the specific problem (1.2), and show that it enjoys a *fine-grained*, *accelerated* convergence rate in the sense that it has sharp dependency on the aforementioned Lipschitz constants in an individual fashion. More precisely, one can reduce the complexity due to the condition number of f and g by a square root. Overall, the best rate that can be achieved in this setting is given by the lower bound $\Omega\left(\sqrt{\frac{L_f}{\mu_f}+\frac{L_g}{\mu_g}}+\sqrt{\frac{L_H^2}{\mu_f\mu_g}}\right)\log\left(\frac{1}{\epsilon}\right)$, established by Zhang et al. [43]. With the goal of matching such a first-order complexity, we are devoted to the problem of: *Designing rate-optimal single-call algorithms for deterministic and stochastic bilinearly coupled saddle point problems*.

1.1. Contributions

We list our contributions in this subsection. First, we present a novel algorithm that blends acceleration dynamics based on the single-call OGDA algorithm for the adversarial part and Nesterov's acceleration for the individual part. We refer to this blend as the Accelerated Gradient-Optimistic Gradient (AG-OG) algorithm. Second, we illustrate how the dynamics of AG-OG can be seen as the summation of the OGDA dynamics and Nesterov's acceleration dynamics, which simplifies the understanding of acceleration on OGDA and provides insights into the convergence proof. Additionally, equipped with a scheduled restarting technique, we derive an "accelerated optimistic gradient with Nesterov's acceleration and restarting" (AVATAR) method that achieves an upper bound that matches the lower bound on Problem (1.2). Finally, when it comes to stochastic settings, we present a stochastic version of AVATAR and establish its convergence at an optimal rate. Overall, our work provides novel acceleration schemes for deterministic/stochastic OGDA under bilinearly-coupled saddle point problems (1.2) (Thms 4 and 6). Our algorithm is simple, shows high interpretability in discrete and continuous dynamics, and is the first single-call algorithm with a rate-optimal convergence rate with known sharpest bias term in the stochastic settings.

2. Preliminaries

In minimax optimization the goal is to find the Nash equilibrium point of problem (1.1), 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 $\nabla_{\boldsymbol{x}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$, $\nabla_{\boldsymbol{y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$. Thus, finding the Nash equilibrium 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 $\boldsymbol{z} \in \mathbb{R}^{n+m}$ the concatenation of $\boldsymbol{x}, \boldsymbol{y}$, and analogously for $\boldsymbol{z}^*$:

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

Based on this formulation, our goal is to find the stationary point of the monotone operator (or vector field) $W(\boldsymbol{z})$, namely a $\boldsymbol{z}^* = [\boldsymbol{x}^*; \boldsymbol{y}^*] \in \mathbb{R}^{n+m}$ satisfying (in the unconstrained case)

$$W(\boldsymbol{z}^*) = 0. \tag{2.2}$$

Problem (2.2) is referred to as the *variational inequality (VI) formulation* of minimax optimization. The compact representation of the convex-concave saddle-point problem as a VI allows us to simplify the notation.

In the gradient field (2.1), there are individual parts that points towards the direction optimizing f, g cooperatively, and adversarial part which corresponds to the gradient vector field of a bilinear minimax problem. For the individual part, we let $F(\boldsymbol{z}) := f(\boldsymbol{x}) + g(\boldsymbol{y})$ and correspondingly

$\nabla F(\boldsymbol{z})^\top = \left[\nabla f(\boldsymbol{x})^\top, \nabla g(\boldsymbol{y})^\top\right]$. For the adversarial part, we define the operator $H(\boldsymbol{z})^\top = \left[(\mathbf{B}\boldsymbol{y})^\top, -(\mathbf{B}^\top \boldsymbol{x})^\top\right]$. Note that the representation allows us to write $W(\boldsymbol{z})$ in Problem (2.2) 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 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. Formally we have for $\forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$ and $\forall \boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^m$:*

$$\frac{\mu_f}{2}\left\|\boldsymbol{x}-\boldsymbol{x}'\right\|^2 \le f(\boldsymbol{x}) - f(\boldsymbol{x}') - \left\langle \nabla f(\boldsymbol{x}'), \boldsymbol{x}-\boldsymbol{x}'\right\rangle \le \frac{L_f}{2}\left\|\boldsymbol{x}-\boldsymbol{x}'\right\|^2,$$
$$\frac{\mu_g}{2}\left\|\boldsymbol{y}-\boldsymbol{y}'\right\|^2 \le g(\boldsymbol{y}) - g(\boldsymbol{y}') - \left\langle \nabla g(\boldsymbol{y}'), \boldsymbol{y}-\boldsymbol{y}'\right\rangle \le \frac{L_g}{2}\left\|\boldsymbol{y}-\boldsymbol{y}'\right\|^2.$$

This implies that $F(\boldsymbol{z})$ is $L_f \vee L_g$-smooth and μ-strongly convex, where $\mu = \mu_f \wedge \mu_g$.

The above assumption adds convexity and smoothness constraints to the individual parts $f(\boldsymbol{x})$ and $g(\boldsymbol{y})$. For the adversarial part $\boldsymbol{x}^\top \mathbf{B}\boldsymbol{y}$, without loss of generality, we assume that $\mathbf{B} \in \mathbb{R}^{n\times m}$, $n \ge m > 0$ is a tall matrix. Note that as $\boldsymbol{x}$ and $\boldsymbol{y}$ are exchangeable, tall matrices cover all circumstances. Moreover, the bilinear structure of the coupling function yields the property that for all $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^{n+m}$:

$$\left\langle H(\boldsymbol{z}) - H(\boldsymbol{z}'), \boldsymbol{z}-\boldsymbol{z}'\right\rangle = 0. \tag{2.3}$$

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

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

Note that the bounded noise assumption is common in the stochastic optimization literature. Under the above assumptions, our goal is to find a ϵ-optimal solution z such that $||\boldsymbol{z} - \boldsymbol{z}^*||^2 \le \epsilon$. In the section that follows, we achieve this goal by adopting carefully crafted blending between Nesterov's acceleration, optimism, and scheduled restarting.

3. Accelerated Gradient Optimistic Gradient Descent Ascent

In this section, we discuss key elements of our algorithm design—the so-called *OptimisticGradient Descent-Ascent (OGDA)* and *Nesterov's acceleration method*—that together solves the bilinear saddle-point 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.

. We leave the generalization to models of unbounded noises to future work.

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) [17]. In the general variational inequality setting, the iteration of OGDA takes the following form [32]:

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

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 [14]:

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

Treating $W(\boldsymbol{z}_{k-\frac{1}{2}}) - W(\boldsymbol{z}_{t-\frac{3}{2}})$ as a prediction of the future, $W(\boldsymbol{z}_{k+\frac{1}{2}}) - W(\boldsymbol{z}_{k-\frac{1}{2}})$, 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 [22]: Although similar conceptually to OGDA (3.1), EG requires two gradient computations per iterate, which doubles the number of gradient computations of OGDA. Both OGDA and EG dynamics (3.1) alleviate the cyclic behavior by extrapolation from the past and exhibit a complexity of $(L/\mu)\log(1/\epsilon)$ [14, 29] in general setting (1.1) with L-smooth, μ-strongly-convex-μ-strongly-concave objectives.

3.2. Nesterov's Acceleration Scheme

Turning to the minimization problem, while vanilla gradient descent enjoys a gradient complexity of $\kappa\log(1/\epsilon)$ on L-smooth, μ-strongly convex problems, with $\kappa = L/\mu$ being the condition number, Nesterov's method [30], when equipped with proper restarting, achieves an improved gradient complexity of $\sqrt{\kappa}\log(1/\epsilon)$. We adopt the following version of the Nesterov acceleration, known as the "second scheme" [27, 39]:

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

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

$$\boldsymbol{z}_{k+1}^{\text{ag}} - \boldsymbol{z}_k^{\text{md}} = \frac{2}{k+2}(\boldsymbol{z}_{k+1} - \boldsymbol{z}_k) = -\eta_k \nabla F(\boldsymbol{z}_k^{\text{md}}) \Rightarrow \qquad \boldsymbol{z}_{k+1}^{\text{ag}} = \boldsymbol{z}_k^{\text{md}} - \eta_k \nabla F(\boldsymbol{z}_k^{\text{md}}). \tag{3.4}$$

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

$$\frac{k+2}{k+3}\boldsymbol{z}_{k+1}^{\text{ag}} - \boldsymbol{z}_{k+1}^{\text{md}} = \frac{k}{k+3}\boldsymbol{z}_k^{\text{ag}} - \frac{k+1}{k+3}\boldsymbol{z}_{k+1}^{\text{ag}} \Rightarrow \qquad \boldsymbol{z}_{k+1}^{\text{md}} = \boldsymbol{z}_{k+1}^{\text{ag}} + \frac{k}{k+3}\left(\boldsymbol{z}_{k+1}^{\text{ag}} - \boldsymbol{z}_k^{\text{ag}}\right). \tag{3.5}$$

. In fact an analogous result holds true for general smooth, strongly monotone variational inequalities. We refer to Mokhtari et al. [29] for background and related work.

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

1: **for** $k = 0, 1, \ldots, K-1$ **do**
2: $\quad \boldsymbol{z}_k^{\text{md}} = (1-\alpha_k)\boldsymbol{z}_k^{\text{ag}} + \alpha_k \boldsymbol{z}_k$
3: $\quad \boldsymbol{z}_{k+\frac{1}{2}} = \boldsymbol{z}_k - \eta_k \left(H(\boldsymbol{z}_{k-\frac{1}{2}}) + \nabla F(\boldsymbol{z}_k^{\text{md}}) \right)$
4: $\quad \boldsymbol{z}_{k+1}^{\text{ag}} = (1-\alpha_k)\boldsymbol{z}_k^{\text{ag}} + \alpha_k \boldsymbol{z}_{k+\frac{1}{2}}$
5: $\quad \boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta_k \left(H(\boldsymbol{z}_{k+\frac{1}{2}}) + \nabla F(\boldsymbol{z}_k^{\text{md}}) \right)$
6: **end for**
7: **Output:** $\boldsymbol{z}_K^{\text{ag}}$

Thus, by a simple notational transformation, (3.4) plus (3.5) (and hence the original update rule (3.3)) is exactly equivalent to the original updates of Nesterov's acceleration scheme [30]. Here, $\boldsymbol{z}_k^{\text{ag}}$ denotes a $\frac{2}{k}$-*weighted-averaged* iteration. In other words, compared with the vanilla gradient descent that proceeds as $\boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta_k \nabla F(\boldsymbol{z}_k)$, 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 also been discussed in recent work on the heavy-ball-based acceleration method [34, §1.3].

3.3. Accelerating OGDA on Bilinear Saddle Point Problems

In this subsection and §3.4, we show that an organic combination of the two algorithms in §3.1 and §3.2 achieves a desirable convergence rate in discrete time and when equipped with scheduled restarting, obtains a lower bound complexity of $\left(\sqrt{L_f/\mu_f} \vee \sqrt{L_g/\mu_g} + L_H/\sqrt{\mu_f \mu_g}\right) \log(1/\epsilon)$. 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 follows that in (3.3), while in Line (3) and (5) the updates follows the OGDA dynamics (3.1) with each step modified by (3.3b).

We first state an elementary lemma that shows the non-expansive property of $\boldsymbol{z}_k$, whose proof is presented in §G.3.

Lemma 3 (Bounded Iterates) *Under Assumptions 1, we set the parameters as* $L = L_f \vee L_g$, $L_H = \sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}$, $\eta_k = \frac{k+2}{2L + \sqrt{3+\sqrt{3}} L_H (k+2)}$ *and* $\alpha_k = \frac{2}{k+2}$ *in Algorithm 1, at any iterate* $k < K$, $\boldsymbol{z}_k$ *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 $\boldsymbol{z}^*$ *denotes the unique solution of Problem* (1.2).

Lemma 3 establishes the following last-iterate boundedness: the $\boldsymbol{z}_k$ iteration is bounded within the ball centered at $\boldsymbol{z}^*$ and radius $||\boldsymbol{z}_0 - \boldsymbol{z}^*||$ and is essential in proving convergence of iteration $\boldsymbol{z}_k^{\text{ag}}$, where handling additional recursions brought by gradient evaluated from a previous step is the main technical difficulty. With the parameter choice in Lemma 3, Line 4 can also be seen as an average step that makes the cycling last iterates shrink toward the center of convergence. Equipped with Lemma 3, we are ready to state the following convergence theorem for discrete-time AG-OG:

Algorithm 2 Accelerated optimistic gradient with nesteroV AcceleraTion And Restarting (AVATAR)

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

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

Theorem 4 *Under Assumption 1 and setting the parameters as in Lemma 3, the output of Algorithm 1 satisfies:*

$$||z_K^{ag} - z^*||^2 \leq \frac{4L + 2\sqrt{3+\sqrt{3}}L_H(K+1)}{\mu(K+1)^2}\|z_0 - z^*\|^2. \tag{3.6}$$

The proof of Theorem 4 is provided in §E.1. The choice of $\alpha_k = \frac{2}{k+2}$ is vital for Nesterov's accelerated gradient descent to achieve desirable convergence behavior [30]. The convergence rate in (3.6) for strongly convex problems is slow and is not even a linear convergence. However, in the next subsection we show how a simple restarting technique not only achieves the linear convergence rate, but also matches the lower bound in Zhang et al. [43].

3.4. Improving Convergence Rates via Restarting

Normally, as f and g has different strong convexity parameters (μ_f and μ_g), it is not ideal to use the same stepsize η_k for both the descent step on $f(x)$ and the ascent step on $g(y)$. We accordingly introduce a scaling reduction technique [13] that allows us to consider applying a single scaling for all parameters. Setting $\widehat{y} = \sqrt{\frac{\mu_g}{\mu_f}} y$, we have $\nabla_{\widehat{y}} H(z) = \sqrt{\frac{\mu_f}{\mu_g}} \nabla_y H(z)$ and $\nabla_{\widehat{y}} G(y) = \sqrt{\frac{\mu_f}{\mu_g}} \nabla G(y)$. Other scaling changes are listed as follows:

$$L = L_f \vee \frac{\mu_f}{\mu_g} L_g, \quad L_H = \sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B}) \cdot \frac{\mu_f}{\mu_g}}, \quad \eta_{k,y} = \frac{\eta_k \mu_f}{\mu_g}, \quad \mu = \mu_f, \tag{3.7}$$

where by $\eta_{k,y}$ we mean that when updating $z = [x; y] \in \mathbb{R}^{n+m}$, we adopt stepsize η_k on the first n coordinates and $\eta_{k,y}$ on the last m coordinates. With the new scaling, by scheduled restarting the AGOG algorithm 1, we call the restarting Algorithm 2 the AVATAR, the overall algorithm the AGOG-Avatar and have the following Corollary 5.

Corollary 5 *With scheduled restarting imposed on top of Algorithm 1. Algorithm 2 outputs a solution within an ϵ-ball of z^* within a number N of iterates, where N satisfies*

$$N \geq \mathcal{O}\left(\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \sqrt{\frac{L_H^2}{\mu_f \mu_g}}\right) \cdot \log\left(\frac{1}{\epsilon}\right)\right). \tag{3.8}$$

We defer the proof of the corollary to §E.2. Our rate in Eq. (3.8) of Corollary 5 exactly matches the lower bound result in Zhang et al. [43].

4. Stochastic Accelerated Gradient Optimistic Gradient Descent Ascent

While the complexity result in Corollary 5 has also been obtained in Jin et al. [20], Kovalev et al. [23], Thekumparampil et al. [38], in addition to conceptual simplicity our approach has the significant advantage that it yields a stochastic version of the algorithm and a convergence rate for the stochastic case. Indeed, the stochastic version of Algorithm 1 and 2 maintains an optimal convergence behavior. The stochastic AG-OG algorithm simply replaces each deterministic gradient with its stochastic counterpart, with noise represented by ζ_t, ξ_t. The full stochastic AG-OG algorithm is shown in Algorithm 3 in §D.2.

Based on a boundedness lemma (Lemma 17, presented in §E.3), which is the stochastic analogue of Lemma 3, we can proceed to our stochastic result. See §E.3 for the proof.

Theorem 6 *Under Assumption 1 and 2, set the parameters as in Lemma 17 and take* $\eta_k = \frac{k+2}{4L+D+4\sqrt{2+\sqrt{2}}L_H(k+2)}$*. Then the output of Algorithm 3 satisfies:*

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

Remark 7 *Without knowledge of* $\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2$*, we need a different scheme to set up the stepsize* η_k*. We assume an upper bound on* $||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2$ *defined as* Γ_0 *and let* $C = \frac{\Gamma_0}{\sqrt{\mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2}}$*. Then, we have that* $D = \frac{\sigma A(K)}{\Gamma_0}$ *which is known. Thus,*

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

As in §3.4, we restart the S-AG-OG algorithm and obtain the following optimal complexity:

Corollary 8 *With scheduled restarting imposed on top of Algorithm 1, Algorithm 2 outputs a solution within an* ϵ*-ball of* $\boldsymbol{z}^*$ *within* N *iterates, for* N *satisfying:*

$$N \geq \mathcal{O}\left(\left(\sqrt{\frac{L_f}{\mu_f} \vee \frac{L_g}{\mu_g}} + \sqrt{\frac{L_H^2}{\mu_f \mu_g}}\right) \cdot \log\left(\frac{1}{\epsilon}\right) + \frac{\sigma^2}{\mu_f^2 \epsilon^2}\right).$$

This result matches that of Du et al. [13] which matches Zhang et al. [43] in the nonrandom setting and admits worst-case optimality on the stochastic noise.

5. Conclusion

In this paper, we propose novel algorithms under both a deterministic setting (AG-OG) and a stochastic setting (S-AG-OG) with structural interpretability and simplicity. To illustrate the design, we provide a continuous-time dynamics approximation that blends optimism with Nesterov's acceleration and an intuitive proof of convergence by leveraging a novel Lyapunov analysis. When discretizing the dynamics using AG-OG, we conclude desirable polynomial convergence behavior in discrete time. By properly restarting the algorithm, we propose the corresponding AGOG-Avatar and S-AGOG-Avatar algorithms and prove theoretically that our restarted algorithms enjoys rate-optimal sample complexity for finding an ϵ-accurate solution. Future directions include generalizations to stochastic settings, to nonconvexity, and to improvement of statistical error from worst-case optimality to (near) instance-dependent optimality.

References

[1] Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pages 778–816. PMLR, 2022.

[2] Waïss Azizian, Damien Scieur, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. Accelerating smooth games by manipulating spectral shapes. In *International Conference on Artificial Intelligence and Statistics*, pages 1705–1715. PMLR, 2020.

[3] Yu Bai and Chi Jin. Provable self-play algorithms for competitive reinforcement learning. *arXiv preprint arXiv:2002.04017*, 2020.

[4] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

[5] Aleksandr Beznosikov, Boris Polyak, Eduard Gorbunov, Dmitry Kovalev, and Alexander Gasnikov. Smooth monotone stochastic variational inequalities and saddle point problems–survey. *arXiv preprint arXiv:2208.13592*, 2022.

[6] Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Accelerated schemes for a class of variational inequalities. *Mathematical Programming*, 165(1):113–149, 2017.

[8] Michael B Cohen, Aaron Sidford, and Kevin Tian. Relative lipschitzness in extragradient methods and a direct recipe for acceleration. *arXiv preprint arXiv:2011.06572*, 2020.

[9] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. In *International Conference on Machine Learning*, pages 1125–1134. PMLR, 2018.

[10] Simon S Du and Wei Hu. 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*, pages 196–205. PMLR, 2019.

[11] Simon S Du, Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. Stochastic variance reduction methods for policy evaluation. In *International Conference on Machine Learning*, pages 1049–1058. PMLR, 2017.

[12] Simon S Du, Sham M Kakade, Jason D Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl. *arXiv preprint arXiv:2103.10897*, 2021.

[13] Simon S Du, Gauthier Gidel, Michael I Jordan, and Chris Junchi Li. Optimal extragradient-based bilinearly-coupled saddle-point optimization. *arXiv preprint arXiv:2206.08573*, 2022.

[14] Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

[15] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Rémi Le Priol, Gabriel Huang, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1802–1811. PMLR, 2019.

[16] Noah Golowich, Sarath Pattathil, and Constantinos Daskalakis. Tight last-iterate convergence rates for no-regret learning in multi-player games. *Advances in neural information processing systems*, 2020.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[18] Eduard Gorbunov, Adrien Taylor, and Gauthier Gidel. Last-iterate convergence of optimistic gradient method for monotone variational inequalities. *Advances in neural information processing systems*, 2022.

[19] Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[20] Yujia Jin, Aaron Sidford, and Kevin Tian. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. *arXiv preprint arXiv:2202.04640*, 2022.

[21] Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

[22] Galina M Korpelevich. The extragradient method for finding saddle points and other problems. *Ekonomika i Matematicheskie Metody*, 12:747–756, 1976.

[23] Dmitry Kovalev, Alexander Gasnikov, and Peter Richtárik. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. *arXiv preprint arXiv:2112.15199*, 2021.

[24] Chris Junchi Li, Yaodong Yu, Nicolas Loizou, Gauthier Gidel, Yi Ma, Nicolas Le Roux, and Michael Jordan. On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging. In *International Conference on Artificial Intelligence and Statistics*, pages 9793–9826. PMLR, 2022.

[25] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 907–915. PMLR, 2019.

[26] Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020.

[27] Zhouchen Lin, Huan Li, and Cong Fang. Accelerated optimization for machine learning. *Nature Singapore: Springer*, 2020.

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[29] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

[30] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

[31] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems*, 29, 2016.

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

[33] Ernest K Ryu, Kun Yuan, and Wotao Yin. Ode analysis of stochastic gradient methods with optimism and anchoring for minimax problems. *arXiv preprint arXiv:1905.10899*, 2019.

[34] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pages 3935–3971. PMLR, 2021.

[35] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

[36] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: theory and insights. *Advances in neural information processing systems*, 27, 2014.

[37] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[38] Kiran K Thekumparampil, Niao He, and Sewoong Oh. Lifted primal-dual method for bilinearly coupled smooth minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 4281–4308. PMLR, 2022.

[39] Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2(3), 2008.

[40] Yuanhao Wang and Jian Li. Improved algorithms for convex-concave minimax optimization. *Advances in Neural Information Processing Systems*, 33:4800–4810, 2020.

[41] Guangzeng Xie, Yuze Han, and Zhihua Zhang. Dippa: An improved method for bilinear saddle point problems. *arXiv preprint arXiv:2103.08270*, 2021.

[42] Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33:5667–5678, 2020.

[43] Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, pages 1–35, 2021.

[44] Siqi Zhang, Junchi Yang, Cristóbal Guzmán, Negar Kiyavash, and Niao He. The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR, 2021.

[45] Renbo Zhao. Accelerated stochastic algorithms for convex-concave saddle-point problems. *Mathematics of Operations Research*, 47(2):1443–1473, 2022.

Appendix A. Related Work

Deterministic Case. Many works have studied the linear convergence rate of gradient-based methods for games in the context of strongly monotone operators (which is implied by strong convex-concavity) [29] and several works [8, 40–42, 44] have slowly 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 [8, 26, 29, 40, 41], and only very recently have optimal results that reach the lower bound been presented Jin et al. [20], Kovalev et al. [23], Thekumparampil et al. [38]. These three works proposed improved methods leveraging convex duality. However, it could be challenging to extend the proposed methods to the non-bilinear coupling term case or the stochastic case. In particular, none of these previous works provide result in the non-finite-sum stochastic case.

Stochastic Case. There exists a rich literature on stochastic variational inequalities with application to solving stochastic minimax problems [1, 5, 6, 19, 21, 45]. However, only a few works have proposed fined-grained bounds suited to the (Bil-)SC-SC setting. To the best of our knowledge, most fined-grained bounds have been proposed in the finite-sum setting [20, 31]. Two closely related works are Li et al. [24], who provide a convergence rate for extra-gradient in the purely bilinear setting and Du et al. [13], who study an accelerated version of extra-gradient, dubbed as AcceleratedGradient-ExtraGradient (AG-EG) in the Bil-SC-SC setting. Our work provides results in the same vein as Du et al. [13] but instead employs the optimistic gradient instead of extra-gradient to handle the bilinear coupling part. Optimistic-gradient-based methods have been considered extensively in literature due to its requiring fewer gradient oracle calls per iteration than standard extra-gradient and can be applied to the online learning setting [16]. Note that, in general, EG and OG methods often share some similarities in their analyses, but also acknowledge sigfinicant differences [16, §3.1], [18, §2]. More specifically in our case, using a *single-call algorithm* that reuses previously calculated gradients alters the recursion (Eq. (E.6)). Although the main part of the proof follows the streamline 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. (E.7)) underneath the original iterative rule essential for proving bounded iterates. In addition, due to the accumulated error across iterates, the maximum stepsize allowed in single-call algorithms is forced to be smaller.

Appendix B. Examples

In this subsection, we use two examples to showcase the applications of formulation (1.2). In the first example, we demonstrate how the parameters of a linear state-value function can be estimated by solving (1.2). In the second example, we illustrate that turning the contraint in a robust learning problem into penalty and we get an objective in the form of (1.2).

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

- s_t, s_{t+1} are the current state (at time t) and future state (at time $t+1$), respectively;
- a_t is the action at time t generated by policy π, that is, $a_t = \pi(s_t)$;

. Limited by space, we refer readers to §E.1 and §E.3 for technical details.

- $r_t = r(s_t, a_t)$ 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

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

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)*:

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

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 (B.1) would necessitate computing the inversion (and storing) of the matrix $\mathbf{C}$, or alternatively computing the matrix-vector product $\mathbf{C}^{-1}\boldsymbol{v}$ for a vector $\boldsymbol{v}$ at each step, and would be computationally costly or even prohibited. To circumvent inverting matrix $\mathbf{C}$ a reformulation via *conjugate function* can be resorted to; that is, solving (B.1) is equivalent to solving the following saddle-point (or minimax) problem [10, 11]:

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

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

Robust Learning. A robust learning or robust optimization problem targets to minimize an objective function (here the least-square) formulated as a minimax (saddle-point) optimization problem [4, 10, 38]

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

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 (B.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 saddle-point optimization problem.

Appendix C. Motivations of AG-OG from Continuous-Time Perspectives

C.1. Motivations of AG-OG from Continuous-Time Perspectives

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

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

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

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

$$\dot{\widehat{Z}} + H(\widehat{Z}) = 0. \tag{C.1}$$

We note here that ODE (C.1) is a coarse-grained dynamics and its discretization can be either EG or GDA as well. However, as can be seen later in Proposition 9 and Theorem 10, the intuition of blending (C.1) with another dynamics motivates a strongly interpretable algorithm together with its theoretical analysis.

Dynamics of Nesterov's Acceleration. Su et al. [36] derived the following ODE for the standard form of Nesterov's scheme:

$$\ddot{Z} + \frac{3}{t}\dot{Z} + \nabla F(Z) = 0. \tag{C.2}$$

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

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

where $\widehat{Z} = Z + \frac{t}{2}\dot{Z}$. Intuitively, while Z represents the continuous dynamics of $\boldsymbol{z}_k^{\mathrm{ag}}$ (averaging point), $\widehat{Z}$ represents the dynamics of $\boldsymbol{z}_k$. The equality $\widehat{Z} = Z + \frac{t}{2}\dot{Z}$ exactly matches the relationship between $\boldsymbol{z}_k^{\mathrm{ag}}$ and $\boldsymbol{z}_k$ shown in the update rule (3.3c). More specifically, starting from the lens of discrete time

. In fact, a direct acceleration via momentum leads to sub-optimality for the refined case (1.2) where the condition number of the three parts are entangled [15].

updates (3.3c) and (3.3b), where $\boldsymbol{z}_{k+1} = \frac{k}{2}(\boldsymbol{z}_{k+1}^{\text{ag}} - \boldsymbol{z}_k^{\text{ag}}) + \boldsymbol{z}_{k+1}^{\text{ag}}, \quad \boldsymbol{z}_{k+1} = \boldsymbol{z}_k - \eta_k \nabla F(\boldsymbol{z}_k^{\text{md}})$, by choosing $L' > L$ where L is the Lipschitz constant of f, $\eta_k = \frac{k}{2L'}$, $t = k\Delta t$, $\Delta t = \frac{1}{\sqrt{L'}}$ and finally let $L' \to \infty$, we conclude that (also noting that $\boldsymbol{z}_k^{\text{md}} - \boldsymbol{z}_{k+1}^{\text{ag}}$ is $o(\Delta t)$):

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

which again leads to the ODE (C.2). On a related note, the dynamics of Z is the averaged iteration and is highly related to the idea of anchoring [33].

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

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

Proposition 9 *We let $\eta_k = \frac{k+1}{2L' + \sqrt{3+\sqrt{3}}L'_H(k+1)}$, where we assume arbitrarily large $L' > L$ and $L'_H > L_H$. Furthermore by letting $\delta = \frac{1}{\sqrt{L' \vee L'_H}}$, as $\delta \to 0^+$ and $\frac{L'_H}{\sqrt{L'}} \to c \in [0, \infty)$ the ODE for the AG-OG Algorithm 1 under this scaling condition is*

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

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

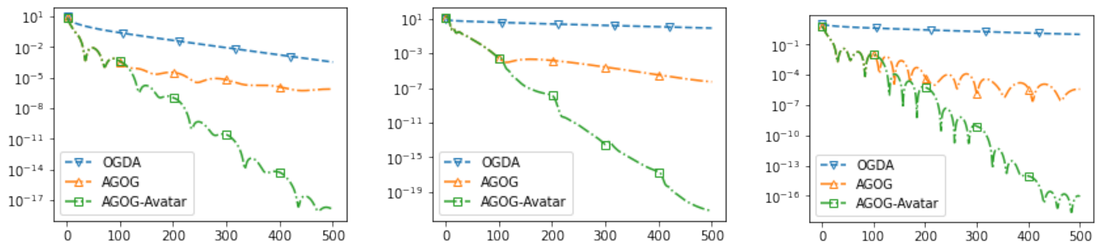
$$\frac{t}{2}\ddot{Z} + \frac{3}{2}\dot{Z} + c_t\left(\nabla F(Z) + HZ + \frac{t}{2}H\dot{Z}\right) = 0. \tag{C.5}$$

We finalize this subsection with a convergence rate result for the combined dynamics in ODE (C.4). Throughout the paper we use $\boldsymbol{z}^*$ to denote a solution to Problem (1.2). We conduct a standard continuous-time analysis based on the following Lyapunov energy function: $\mathcal{E} = t^2(F(Z) - F(\boldsymbol{z}^*)) + \frac{t}{c_t}\|\widehat{Z} - \boldsymbol{z}^*\|^2$, and compute the time-derivative $\dot{\mathcal{E}}$. In contrast with existing analysis for the optimization setting [36], our analysis reposes on the use of a continuous-time Gronwall-inequality-based analysis. The overall result is the following theorem, whose proof is deferred to §F.1:

Theorem 10 *We define the gap function as $V(Z) = F(Z) - F(\boldsymbol{z}^*) + \langle Z - \boldsymbol{z}^*, H\boldsymbol{z}^*\rangle$ and have*

$$V(Z(T)) \le \frac{1}{T}\left(\frac{2}{(1\vee c)^2 T} + \frac{\sqrt{3+\sqrt{3}}c}{1\vee c}\right)\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2.$$

Note in the special case where $c = 0$, the result recovers the standard continuous-time convergence rate as in Su et al. [36]. Later in §3.3 we will prove a similar result (Theorem 4) in the discrete-time context. We note in the meantime that the continuous-time analysis is much simpler and more informative than the (comparatively more complicated) discrete-time analysis.



(*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)

Appendix D. Experiments

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

D.1. Deterministic Setting

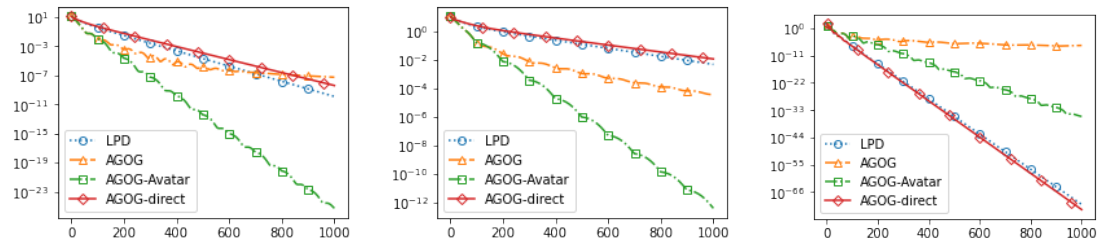
We present results on synthetic quadratic game datasets:

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

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

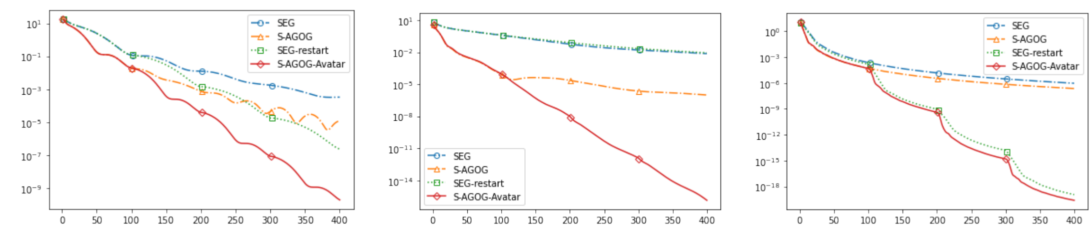
Comparison with OGDA We use the single-call OGDA algorithm [14, 19] as the baseline. In Figure 1 we plot the AGOG algorithm and the AGOG-Avatar algorithm under three different instances. We use stepsize $\eta_k = \frac{k+2}{2L+\sqrt{3+\sqrt{3}}L_H(k+2)}$ in both the AGOG and the AGOG-Avatar algorithms and restart AGOG-Avatar once every 100 iterates. For the OGDA algorithm, we take stepsize $\eta = \frac{1}{2(L \vee L_H)}$. For the parameters of the problem (D.1), we fix $L_H = 1, L_f = 64, \mu_f = 1$ and change different choices of L_g, μ_g. In Figure 1(*a*)subfigure we take $L_g = 64$, $\mu_g = 1$. In Figure 1(*b*)subfigure we take $L_g = 1$, $\mu_g = 1/64$ and in Figure 1(*c*)subfigure we take $L_g = 4096$, $\mu_g = 64$. We see from 1(*a*)subfigure, 1(*b*)subfigure and 1(*c*)subfigure 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 AGOG, matching our theoretical observations.

Comparison with LPD Next, we focus on comparison to the Lifted Primal-Dual (LPD) algorithm [38]. We implement the AGOG algorithm and its restarted version, the AGOG-Avatar. Additionally, inspired by the technique of a single-loop direct-approach in Du et al. [13], we consider a single-loop algorithm named AGOG-Direct that takes advantage of the strongly-convex strongly-concave property of the problem. We refer readers to Du et al. [13] for the "direct" method. The parameters of LPD are chosen as described in Thekumparampil et al. [38]. For our AGOG and AGOG-Avatar Algorithm, 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. (3.7). For the AGOG-direct algorithm, we take $\eta = \frac{1}{(1+\sqrt{L/\mu_f+(\sqrt{3+\sqrt{3}}L_H)^2/\mu_f^2})\mu_f}$ with the same set of scaling parameters. We restart AGOG-Avatar once every 100 iterates.



(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)

In Figure 2(*a*)subfigure, the bilinear coupling part $\boldsymbol{y}^\top A_2 \boldsymbol{x}$ is the dominant part. In Figure 2(*b*)subfigure, we set the eigenvalues of A_2 even larger than in Figure 2(*a*)subfigure. In Figure 2(*c*)subfigure, $\boldsymbol{x}^\top A_1 \boldsymbol{x}$ 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*)subfigure and 2(*b*)subfigure that AGOG-Avatar (green line) outperforms LPD and MP in regimes where the bilinear term dominates, and when the eigenvalues of the coupling matrix increase, the performance of AGOG-Avatar relative to other algorithms is enhanced. This is in accordance with our theoretical analysis. In addition, AGOG-Avatar outperforms its non-restarted version (orange line) which has a gentle slope at the end. On the other hand, when the individual part dominates, our AGOG-direct (red line) slightly outperforms LPD. Moreover, AGOG-direct and LPD almost overlap in 2(*a*)subfigure and 2(*b*)subfigure.

D.2. Stochastic Setting

We compared stochastic AGOG and its restarted version S-AGOG-Avatar with Stochastic Extragradient (SEG) SEG with restarting, respectively [cf. 24]. 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 AGOG and AGOG-Avatar, 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 AGOG-Avatar.

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

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

We use the same quadratic game setting as in (D.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*)subfigure we consider larger eigenvalues for A_2 than A_1, A_3. In 3(*b*)subfigure, we let A_1, A_2, A_3 to be approximately of the same scale. In 3(*c*)subfigure, as the scale of the eigenvalues shrinks, the noise is relatively larger than in 3(*a*)subfigure and 3(*b*)subfigure. The specific choice of parameters are shown in the caption of the corresponding figures. We see from 3(*a*)subfigure, 3(*c*)subfigure and 3(*c*)subfigure that S-AGOG-Avatar achieves a more desirable convergence speed than SEG-restart. Also, the restarting technique significantly accelerates the convergence, validating our theory.

Appendix E. Proof of Main Convergence Results

This section collects the proofs of our main results, Theorem 4 [§E.1], Corollary 5 [§E.2], and Theorem 6 [§E.3].

E.1. Proof of Theorem 4

Proof.[Proof of Theorem 4] We define the point-wise primal-dual gap function as:

$$V(\boldsymbol{z}, \boldsymbol{z}') := f(\boldsymbol{z}) - f(\boldsymbol{z}') + \left\langle H(\boldsymbol{z}'), \boldsymbol{z} - \boldsymbol{z}' \right\rangle \tag{E.1}$$

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 §G.5.

Lemma 11 *For arbitrary $\alpha_k \in (0,1]$ the iterates of Algorithm 1 satisfy for $t = 1, \ldots, T$ almost surely*

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

Note that in Lemma 11, the term I is an inner product with the gradient term and reduces to $\langle \nabla f(\boldsymbol{z}_k), \boldsymbol{z}_k - \omega_{\boldsymbol{z}} \rangle$ of the vanilla gradient algorithm if acceleration and optimistic-gradient are removed. The squared term II is brought by gradient evaluated at $\boldsymbol{z}_k^{\mathrm{md}}$.

Additionally, throughout the proof of Lemma 11, we only leverage the convexity and L-smoothness of f and the property of H (2.3), as well as the update rules in Line (2) and Line (4). The proof involves no update rules regarding the gradient updates and hence Lemma 11 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 12 is a slight modification from the proof of Proposition 4.2 in Chen et al. [7] and analogous to Lemma 7.1 in Du et al. [13]. We omit the proof here as the argument comes from simple algebraic tricks. Readers can refer to Du et al. [13] for more details.

Proposition 12 (Proposition 4.2 in Chen et al. 7 and Lemma 7.1 in Du et al. [12]) *Given an initial point $\boldsymbol{\theta} \in \mathbb{R}^d$, two update directions $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2 \in \mathbb{R}^d$, and the corresponding results $\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \in \mathbb{R}^d$ satisfying:*

$$\boldsymbol{\varphi}_1 = \boldsymbol{\theta} - \boldsymbol{\delta}_1, \qquad \boldsymbol{\varphi}_2 = \boldsymbol{\theta} - \boldsymbol{\delta}_2. \tag{E.3}$$

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

$$\langle \boldsymbol{\delta}_2, \boldsymbol{\varphi}_1 - \mathbf{z} \rangle \leq \frac{1}{2}\|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\|^2 + \frac{1}{2}\left[\|\boldsymbol{\theta} - \mathbf{z}\|^2 - \|\boldsymbol{\varphi}_2 - \mathbf{z}\|^2 - \|\boldsymbol{\theta} - \boldsymbol{\varphi}_1\|^2\right]. \tag{E.4}$$

Noting that the gradient term $\nabla F(\boldsymbol{z}_k^{\mathrm{md}}) + H(\boldsymbol{z}_{k+\frac{1}{2}})$ in I of inequality (E.2) of Lemma 11 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 12, we obtain an upper bound for the inner product term I:

$$\begin{aligned}
\eta_k \cdot I &\leq \frac{1}{2}\eta_k^2 ||H(\boldsymbol{z}_{k+\frac{1}{2}}) - H(\boldsymbol{z}_{k-\frac{1}{2}})||^2 + \frac{1}{2}\left[||\boldsymbol{z}_k - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+1} - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2\right] \\
&\leq \frac{\eta_k^2 L_H^2}{2}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 + \frac{1}{2}\left[||\boldsymbol{z}_k - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+1} - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2\right]
\end{aligned} \tag{E.5}$$

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

$$\begin{aligned}
V(\boldsymbol{z}_{k+1}^{\mathrm{ag}}, \omega_{\boldsymbol{z}}) - (1-\alpha_k)V(\boldsymbol{z}_k^{\mathrm{ag}}, \omega_{\boldsymbol{z}}) &\leq \frac{\eta_k \alpha_k L_H^2}{2}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 \\
&+ \frac{\alpha_k}{2\eta_k}\left[||\boldsymbol{z}_k - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+1} - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2\right] + \frac{L\alpha_k^2}{2}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2
\end{aligned} \tag{E.6}$$

Step 2: Building and solving the recursion To build a iterative rule that connects the previous iterate with the current iterate, we first apply the following lemma to connect $||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2$ and $||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2$ and reduce Eq. (E.6) to compositions of $\{||\boldsymbol{z}_k - \boldsymbol{z}||^2\}_{k=0,\ldots,K-1}$ and $\{||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2\}_{k=0,\ldots,K-1}$ terms. The proof of Lemma 13 is delayed to §G.6.

Lemma 13 *For any stepsize sequence $\{\eta_k\}_{k=0,\ldots,K-1}$ satisfying for some positive constant $c > 0$ and the Lipchitz parameter L_H that $\eta_k L_H \leq \sqrt{c/2}$ holds for all k . In Algorithm 1, the following holds for any $k \in [K-1]$:*

$$||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 \leq 2c^k \sum_{\ell=0}^{k} c^{-\ell}||\boldsymbol{z}_{\ell+\frac{1}{2}} - \boldsymbol{z}_\ell||^2. \tag{E.7}$$

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

$$V(\boldsymbol{z}_{k+1}^{\mathrm{ag}}, \omega_{\boldsymbol{z}}) - \frac{k}{k+2} V(\boldsymbol{z}_k^{\mathrm{ag}}, \omega_{\boldsymbol{z}}) \leq \frac{1}{\eta_k (k+2)} \left[||\boldsymbol{z}_k - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+1} - \boldsymbol{z}||^2 \right]$$
$$- \left(\frac{1}{\eta_k (k+2)} - \frac{2L}{(k+2)^2} \right) ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + \frac{2\eta_k L_H^2}{k+2} \sum_{\ell=0}^{k} c^{k-\ell} ||\boldsymbol{z}_{\ell+\frac{1}{2}} - \boldsymbol{z}_\ell||^2$$

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

$$(k+2)^2 V(\boldsymbol{z}_{k+1}^{\mathrm{ag}}, \omega_{\boldsymbol{z}}) - [(k+1)^2 - 1] V(\boldsymbol{z}_k^{\mathrm{ag}}, \omega_{\boldsymbol{z}}) \leq \frac{k+2}{\eta_k} \left[||\boldsymbol{z}_k - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+1} - \boldsymbol{z}||^2 \right]$$
$$- \left(\frac{k+2}{\eta_k} - 2L \right) ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + 2(k+2)\eta_k L_H^2 \sum_{\ell=0}^{k} c^{k-\ell} ||\boldsymbol{z}_{\ell+\frac{1}{2}} - \boldsymbol{z}_\ell||^2$$

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

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

and the previous inequality reduces to

$$(k+2)^2 V(\boldsymbol{z}_{k+1}^{\mathrm{ag}}, \omega_{\boldsymbol{z}}) - [(k+1)^2 - 1] V(\boldsymbol{z}_k^{\mathrm{ag}}, \omega_{\boldsymbol{z}})$$
$$\leq \left(2L + \sqrt{2/c} L_H (k+2) \right) \left[||\boldsymbol{z}_k - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+1} - \boldsymbol{z}||^2 \right]$$
$$- \sqrt{2/c}(k+2) L_H ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + (k+2)\sqrt{2c} L_H \sum_{\ell=0}^{k} c^{k-\ell} ||\boldsymbol{z}_{\ell+\frac{1}{2}} - \boldsymbol{z}_\ell||^2$$

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

$$\left[(K+1)^2 - 1 \right] V(\boldsymbol{z}_K^{\mathrm{ag}}, \boldsymbol{z}^*) + \left(2L + \sqrt{2/c} L_H (K+1) \right) \|\boldsymbol{z}_K - \boldsymbol{z}^*\|^2$$
$$\leq \left(2L + \sqrt{2/c} L_H \right) ||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \sqrt{2/c} L_H \sum_{k=0}^{K-1} \|\boldsymbol{z}_k - \boldsymbol{z}^*\|^2 - \sum_{k=0}^{K-1} V(\boldsymbol{z}_{k+1}^{\mathrm{ag}}, \boldsymbol{z}^*)$$
$$- \sqrt{2/c} L_H \underbrace{\sum_{k=0}^{K-1} (k+2) ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2}_{\mathrm{III}_1} + \sqrt{2c} L_H \underbrace{\sum_{k=0}^{K-1} (k+2) \sum_{\ell=0}^{k} c^{k-l} ||\boldsymbol{z}_{\ell+\frac{1}{2}} - \boldsymbol{z}_\ell||^2}_{\mathrm{III}_2}$$

Simple algebra yields:

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

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

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

holds for all $k \geq 0$. Thus, summing III_1 and III_2 terms we have

$$-\left(\sqrt{2/c} - \sqrt{2c}\right) L_H \mathrm{III}_1 + 2\sqrt{c/2} L_H \mathrm{III}_2 \leq 0.$$

Finally, we solve the recursion as:

$$\begin{aligned}
&\left[(K+1)^2 - 1\right] V(\boldsymbol{z}_K^{\mathrm{ag}}, \boldsymbol{z}^*) + \left(2L + \sqrt{2/c} L_H (K+1)\right) \|\boldsymbol{z}_K - \boldsymbol{z}^*\|^2 \\
&\leq \left(2L + \sqrt{2/c} L_H\right) ||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \sqrt{2/c} L_H \sum_{k=0}^{K-1} \|\boldsymbol{z}_k - \boldsymbol{z}^*\|^2 - \sum_{k=0}^{K-1} V(\boldsymbol{z}_{k+1}^{\mathrm{ag}}, \boldsymbol{z}^*) \qquad \text{(E.8)}
\end{aligned}$$

Step 3: Proving $\boldsymbol{z}_k$ stays within a neighbourhood of $\boldsymbol{z}^*$ In Lemma 3 we show that $\boldsymbol{z}_k$ is always bounded within the ball centered at $\boldsymbol{z}^*$ with radius $||\boldsymbol{z}_0 - \boldsymbol{z}^*||$.

Lemma 14 (Lemma 3) *Under Assumption 1. Set the parameters as* $L = L_f \vee L_g$, $L_H = \sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}$, $\eta_k = \frac{k+2}{2L + \sqrt{3+\sqrt{3}} L_H (k+2)}$ *and* $\alpha_k = \frac{2}{k+2}$ *in Algorithm 1, at any iterate* $k < K$, $\boldsymbol{z}_k$ *stays within the region defined by the initialization* $\boldsymbol{z}_0$*:*

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

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

Step 4: Combining everything together Bringing the bounded iterates results in Lemma 3 into the recursion (E.8) and rearranging, we obtain the following:

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

Dividing both sides by $(K+1)^2$ and noting that $V(\boldsymbol{z}_K^{\mathrm{ag}}, \boldsymbol{z}^*) \geq \frac{\mu}{2}\mathbb{E}||\boldsymbol{z}_K^{\mathrm{ag}} - \boldsymbol{z}^*||^2$, bringing in the choice of $c = \frac{2}{3+\sqrt{3}}$ and we conclude our proof of Theorem 4. □

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.

E.2. Proof of Corollary 5

Proof.[Proof of Corollary 5] The proof of restarting argument is direct. By Eq. (3.6), if we want $||\boldsymbol{z}_K^{\mathrm{ag}} - \boldsymbol{z}^*||^2 \leq \frac{1}{e}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2$ to hold, we can choose K such that

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

This is equivalent to

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

For a given threshold $\epsilon > 0$, with the output of every epoch satisfying $||\boldsymbol{z}_K^{\mathrm{ag}} - \boldsymbol{z}^*||^2 \leq \frac{1}{e}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2$, the total epochs required to get within the ϵ-ball centered at $\boldsymbol{z}^*$ would be $\log(\frac{||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2}{\epsilon})$. Thus, the total number of iterates required to get within the ϵ threshold would be:

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

Bringing the choice of scaling parameters in (3.7) and we conclude our proof of Corollary 5. □

E.3. Proof of Theorem 6

Proof.[Proof of Theorem 6] For the stochastic case, we use the same definition of the primal-dual gap function, rewritten as:

$$V(\boldsymbol{z}, \boldsymbol{z}') := f(\boldsymbol{z}) - f(\boldsymbol{z}') + \left\langle H(\boldsymbol{z}'), \boldsymbol{z} - \boldsymbol{z}' \right\rangle \tag{E.1}$$

Step 1: Estimating weighted temporal difference in squared norms We mentioned in the proof of Theorem 4 that Lemma 11 holds for the stochastic case as well. Thus, we have

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

By applying Proposition 12 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 12, we obtain the following stochastic version of inequality (E.5):

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

Step 2: Building and solving the recursion Note that in the stochastic case, unlike Step 2 in the proof of Theorem 4, 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 6, we denote

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

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

Lemma 15 *For any $\beta > 0$, under Assumption 2, we have*

$$\mathbb{E}||\widetilde{H}(\boldsymbol{z}_{k+\frac{1}{2}};\zeta_{k+\frac{1}{2}}) - \widetilde{H}(\boldsymbol{z}_{k-\frac{1}{2}};\zeta_{k-\frac{1}{2}})||^2 \le (1+\beta)L_H^2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 + \left(2+\frac{1}{\beta}\right)\sigma_H^2. \tag{E.9}$$

Taking $\beta = 1$ in Lemma 15 and bringing the result into the expectation of (E.2), we obtain that

$$\begin{aligned}
&\mathbb{E}V(\boldsymbol{z}_{k+1}^{\text{ag}},\omega_{\boldsymbol{z}}) - (1-\alpha_k)\mathbb{E}V(\boldsymbol{z}_k^{\text{ag}},\omega_{\boldsymbol{z}}) \\
&\le \frac{\alpha_k\eta_k}{2}\left[2L_H^2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 + 3\sigma_H^2\right] + \alpha_k\mathbb{E}\left\langle \Delta_f^k + \Delta_h^{k+\frac{1}{2}}, \boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}^*\right\rangle \\
&\quad + \frac{L\alpha_k^2}{2}\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + \frac{\alpha_k}{2\eta_k}\mathbb{E}\left[||\boldsymbol{z}_k - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+1} - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2\right].
\end{aligned} \tag{E.10}$$

Following the above inequality and following similar techniques as in Step 2 of the proof of Theorem 4, we can derive the following Lemma 16, whose proof is delayed to §G.8.

Lemma 16 *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*

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

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

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

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

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

Telescoping over $k = 0, 1, \ldots K-1$ and using the same techniques as in the proof of Theorem 4, we have for $\frac{k+2}{2\eta_k} \geq 2L + \frac{1}{\sqrt{c}} L_H(k+2)$ and $c = \frac{1}{2+\sqrt{2}}$ $(c/(1-c) = \sqrt{2}-1$, write $\sigma^2 = 3\sqrt{2}\sigma_H^2 + 2\sigma_F^2$ that

$$\begin{aligned}&\left[(K+1)^2 - 1\right] \mathbb{E}V(\boldsymbol{z}_K^{\text{ag}}, \boldsymbol{z}^*) + \frac{K+1}{\eta_{K-1}} \mathbb{E}||\boldsymbol{z}_K - \boldsymbol{z}^*||^2 \\ &\leq \frac{2}{\eta_0} \mathbb{E}||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{2}{\sqrt{c}} L_H \sum_{k=1}^{K-1} \mathbb{E}\left\|\boldsymbol{z}_k - \boldsymbol{z}^*\right\|^2 + \sum_{k=0}^{K-1} (k+2)\eta_k \sigma^2 - \sum_{k=0}^{K-1} \mathbb{E}V(\boldsymbol{z}_{k+1}^{\text{ag}}, \boldsymbol{z}^*)\end{aligned} \tag{E.11}$$

Step 3: Proving z_k stays within a neighbourhood of $\boldsymbol{z}^*$ We introduce the following Lemma 17, whose proof is in §G.4

Lemma 17 *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 $\forall k \in [K-1]$:

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

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

$$||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \leq ||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 + \frac{A(K)^2 \sigma^2}{D^2} \tag{E.12}$$

(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 (E.12) reduces to

$$||\boldsymbol{z}_k - \boldsymbol{z}^*||^2 \leq \left(1 + C^2\right) ||\boldsymbol{z}_0 - \boldsymbol{z}^*||^2 \tag{E.13}$$

Step 4: Combining everything together Combining the choice of stepsize η_k in (a), (b) in Lemma 17 and $\frac{k+2}{2\eta_k} \geq 2L + \frac{1}{\sqrt{c}} L_H(k+2)$, and bound (E.11) with Eq. (E.13), by rearranging the terms again, we conclude that for $\eta_k = \frac{k+2}{4L + D + 4\sqrt{2+\sqrt{2}} L_H(k+2)}$,

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

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

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

□

Appendix F. Understanding the Continuous-Time Dynamics of AG-OG

Due to the complexity of discrete analysis, in §C.1 we describe the continuous dynamics of AG-OG as a hybrid dynamics of OGDA and Nesterov's acceleration. After introducing the discrete dynamics of AG-OG in §3.3, we are now ready to present a formal proposition that connects Algorithm 1 with its ODE. We adopt similar stepsize scheme as in Theorem 4 by letting $\eta_k = \frac{k+1}{2L' + \sqrt{3+\sqrt{3}}L'_H(k+1)}$ where L' and L'_H can be arbitrarily large. Furthermore, we set $\Delta t = \frac{1}{\sqrt{L' \vee L'_H}}$, let $\frac{L'_H}{\sqrt{L'}} \to c$ and use the notation $c_t = \frac{t(1\vee c)^2}{2+\sqrt{3+\sqrt{3}}c(1\vee c)t}$. We derive the following proposition:

Proposition 18 (Proposition 9) *The ODE for the AG-OG Algorithm 1 is*

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

Proof.[Proof of Proposition 9] We recall that in Algorithm 1, Line 3 and 5 yield:

$$\begin{aligned}
\boldsymbol{z}_{k+\frac{1}{2}} &= \boldsymbol{z}_k - \eta_k\left(H(\boldsymbol{z}_{k-\frac{1}{2}}) + \nabla F(\boldsymbol{z}_k^{\mathrm{md}})\right)\\
\boldsymbol{z}_{k+1} &= \boldsymbol{z}_k - \eta_k\left(H(\boldsymbol{z}_{k+\frac{1}{2}}) + \nabla F(\boldsymbol{z}_k^{\mathrm{md}})\right)
\end{aligned}$$

Subtracting the first line from the second line, we have

$$\boldsymbol{z}_{k+1} = \boldsymbol{z}_{k+\frac{1}{2}} - \eta_k\left(H(\boldsymbol{z}_{k+\frac{1}{2}}) - H(\boldsymbol{z}_{k-\frac{1}{2}})\right) \tag{F.2}$$

Furthermore, by shifting indices, we have

$$\boldsymbol{z}_{k+\frac{3}{2}} = \boldsymbol{z}_{k+1} - \eta_{k+1}\left(H(\boldsymbol{z}_{k+\frac{1}{2}}) + \nabla F(\boldsymbol{z}_{k+1}^{\mathrm{md}})\right) \tag{F.3}$$

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

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

Moreover, Line 2 and 4 in Algorithm 1 imply that

$$\boldsymbol{z}_{k+1}^{\mathrm{md}} = \boldsymbol{z}_{k+1}^{\mathrm{ag}} + \frac{k}{k+3}\left(\boldsymbol{z}_{k+1}^{\mathrm{ag}} - \boldsymbol{z}_k^{\mathrm{ag}}\right).$$

Thus, we obtain

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

Recalling that in Theorem 4 we take the stepsize $\eta_k = \frac{k+2}{2L+\sqrt{3+\sqrt{3}}L_H(k+2)}$ and noting that the stepsize can actually be chosen for any $L' > L, L'_H > L_H$ as $\eta_k = \frac{k+2}{2L'+\sqrt{3+\sqrt{3}}L'_H(k+2)}$. We let $t = k\Delta t$ where $\Delta t = \frac{1}{\sqrt{L' \vee L'_H}}$. Thus, we obtain

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

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

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

Combining this with the value of η_k and we obtain that

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

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

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

With this choice of η_k, by Taylor expansion we have:

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

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

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

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

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

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

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

which is equivalent to

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

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

$$\widehat{Z}(t) = Z(t) + \frac{t}{2}\dot{Z}(t). \tag{F.7}$$

Combining (F.6) with (F.7) we conclude our proof.

□

F.1. Proof of Theorem 10

Proof.[Proof of Theorem 10] We first provide an estimate of the time derivative $\dot{\mathcal{E}}$ of the Lyapunov function corresponding to (C.4), and the result is shown is Lemma 19:

Lemma 19 *We set the Lyapunov function as defined in the following* (F.8)*:*

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

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

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

We postpone the proof of Lemma 19 to §G.1. Note that both sides of (F.9) in Lemma 19 presents the quantity $\frac{t}{c_t}\|\widehat{Z} - \boldsymbol{z}^*\|^2$ in its original and gradient forms, respectively. By integrating on both sides and applying a Gronwall-type technique, we obtain the following Lemma 20 which shows that the continuous-time dynamics of AG-EG-ODE are non-expansive with respect to saddle $\boldsymbol{z}^*$.

Lemma 20 *We have*

$$\|\widehat{Z} - \boldsymbol{z}^*\| \le \|z_0 - \boldsymbol{z}^*\|. \tag{F.10}$$

We postpone the rigorous proof of Lemma 20 to §G.2. Now bringing (F.10) in Lemma 20 into (F.9), we conclude that

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

Integrating both sides gives

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

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

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

which concludes our proof.

□

Appendix G. Proof of Auxiliary Lemmas

G.1. Proof of Lemma 19

Proof.[Proof of Lemma 19] Since $\mathcal{E}(t)$ is set as

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

we have its time derivative

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

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

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

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

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

Denote

$$\begin{aligned} &2t\left[F(Z) - F(\boldsymbol{z}^*) + \langle Z - \boldsymbol{z}^*, H\boldsymbol{z}^*\rangle\right] + t^2\langle \nabla F(Z) + H\boldsymbol{z}^*, \dot{Z}\rangle + \frac{2t}{c_t}\langle \widehat{Z} - \boldsymbol{z}^*, \dot{\widehat{Z}}\rangle \\ &\equiv \mathrm{I} + \mathrm{II} + \mathrm{III}. \end{aligned}$$

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

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

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

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

and

$$\begin{aligned} \mathrm{I} + \mathrm{II} + \mathrm{III} &= 2t\left[F(Z) - F(\boldsymbol{z}^*) + \langle Z - \boldsymbol{z}^*, H\boldsymbol{z}^*\rangle\right] - \langle 2t(Z - \boldsymbol{z}^*), \nabla F(Z) + H\boldsymbol{z}^*\rangle \\ &= 2t\left[F(Z) - F(\boldsymbol{z}^*) - \langle Z - \boldsymbol{z}^*, \nabla F(Z)\rangle\right] \le 0, \end{aligned}$$

where the last step uses the convexity of f. This concludes the desired result of Lemma 19. $\square$

G.2. Proof of Lemma 20

Proof.[Proof of Lemma 20] To proceed with proof we adopt a Gronwall-type argument. Note

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

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

$$T^2 V(Z(T)) + \frac{T}{c_t}\|\widehat{Z}(T) - \boldsymbol{z}^*\|^2 - \frac{2}{(1\vee c)^2}\|z_0 - \boldsymbol{z}^*\|^2 \le \sqrt{3+\sqrt{3}}\frac{c}{1\vee c}\int_0^T \|\widehat{Z}(t) - \boldsymbol{z}^*\|^2 dt$$

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

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

which gives, via quotient rule,

$$\frac{d}{dT}\left(\frac{c_t y(T)}{T}\right) = \frac{d}{dT}\left(\frac{(1\vee c)^2 y(T)}{2+\sqrt{3+\sqrt{3}}c(1\vee c)T}\right)$$
$$= (1\vee c)^2\frac{(2+\sqrt{3+\sqrt{3}}c(1\vee c)T)y'(T) - c(1\vee c)y(T)}{(2+\sqrt{3+\sqrt{3}}c(1\vee c)T)^2} \le \frac{2(1\vee c)^2 y'(0)}{(2+\sqrt{3+\sqrt{3}}c(1\vee c)T)^2}$$

so

$$\frac{(1\vee c)^2 y(T)}{2+\sqrt{3+\sqrt{3}}c(1\vee c)T} - \frac{(1\vee c)^2 y(0)}{2} \le 2(1\vee c)^2 y'(0)\int_0^T \frac{1}{(2+\sqrt{3+\sqrt{3}}c(1\vee c)t)^2}dt$$
$$= 2\frac{1\vee c}{\sqrt{3+\sqrt{3}}c}y'(0)\left(\frac{1}{2} - \frac{1}{2+\sqrt{3+\sqrt{3}}c(1\vee c)T}\right)$$

Therefore

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

Plugging in back (G.1) gives

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

so

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

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

$$\|\widehat{Z}(T) - \boldsymbol{z}^*\|^2 \le \|z_0 - \boldsymbol{z}^*\|^2$$

completing the proof. □

G.3. Proof of Lemma 3

Proof.[Proof of Lemma 3] Following (E.8), we drop the V terms and have:

$$\begin{aligned}&\left(2L+\sqrt{2/c}L_H(K+1)\right)\|\boldsymbol{z}_K-\boldsymbol{z}^*\|^2\\&\le\left(2L+\sqrt{2/c}L_H\right)||\boldsymbol{z}_0-\boldsymbol{z}^*||^2+\sqrt{2/c}L_H\sum_{k=0}^{K-1}\|\boldsymbol{z}_k-\boldsymbol{z}^*\|^2.\end{aligned}$$

We adopt a "bootstrapping" argument, similar as the Gronwall-type analysis in the Proof of Theorem 10. We define $M_K=\max_{0\le k\le K-1}||\boldsymbol{z}_k-\boldsymbol{z}^*||^2$ and taking a maximum on each term on the right hand side of the above inequality, we conclude that

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

Thus, we know that $||\boldsymbol{z}_K-\boldsymbol{z}^*||^2\le M_{K-1}$ and hence $M_K=M_{K-1}$ always holds. That yields $M_K=M_0$, and we conclude the proof of Lemma 3. $\square$

G.4. Proof of Lemma 17

Proof.[Proof of Lemma 17] Starting from (E.11) that

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

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

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

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

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

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:

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

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:

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

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. (G.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 17. 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 direct calculation. □

G.5. Proof of Lemma 11

Proof.[Proof of Lemma 11] Recalling that F is L-smooth. To upper-bound the difference in pointwise primal-dual gap between iterates, we first estimate the difference in function values of f via gradients at the extrapolation point. For all $\boldsymbol{u} \in \mathcal{Z}$, the convexity and L-smoothness of $F(\cdot)$ implies that:

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

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

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

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

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

$$\begin{aligned} &F(\boldsymbol{z}_{k+1}^{\text{ag}}) - \alpha_k F(\omega_{\boldsymbol{z}}) - (1-\alpha_k)F(\boldsymbol{z}_k^{\text{ag}}) && \text{(G.5)} \\ &\le \left\langle \nabla F(\boldsymbol{z}_k^{\text{md}}), \boldsymbol{z}_{k+1}^{\text{ag}} - (1-\alpha_k)\boldsymbol{z}_k^{\text{ag}} - \omega_{\boldsymbol{z}} \right\rangle + \frac{L}{2}||\boldsymbol{z}_{k+1}^{\text{ag}} - \boldsymbol{z}_k^{\text{md}}||^2 \\ &= \underbrace{\alpha_k \left\langle \nabla F(\boldsymbol{z}_k^{\text{md}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}} \right\rangle}_{\text{I(a)}} + \underbrace{\frac{L\alpha_k^2}{2}\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k\right\|^2}_{\text{II}} && \text{(G.6)} \end{aligned}$$

where by substracting Line (2) from Line (4) of Algorithm 1 and by Line (4) it self, the last equality of (G.6) follows.

Recalling that $\boldsymbol{z}_k^{\text{ag}}$ corresponds to regular iterates and $\boldsymbol{z}_k^{\text{md}}$ corresponds to the extrapolated iterates of Nesterov's acceleration scheme. The squared error term II in (G.6) is brought by gradient calculation at the extrapolated point instead of the regular point. Note that if we do an implicit version of Nesterov such that $\boldsymbol{z}_{k-1}^{\text{md}} = \boldsymbol{z}_k^{\text{ag}}$, this squared term goes to zero, and the convergence analysis would be the same as in OGDA. This could potentially result in a new implicit algorithm with better convergence guarantee.

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

$$\begin{aligned} &\left\langle H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_{k+1}^{\text{ag}} - \omega_{\boldsymbol{z}} \right\rangle - (1-\alpha_k)\left\langle H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_k^{\text{ag}} - \omega_{\boldsymbol{z}} \right\rangle \\ &= \alpha_k \left\langle H(\omega_{\boldsymbol{z}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}} \right\rangle = \underbrace{\alpha_k \left\langle H(\boldsymbol{z}_{k+\frac{1}{2}}), \boldsymbol{z}_{k+\frac{1}{2}} - \omega_{\boldsymbol{z}} \right\rangle}_{\text{I(b)}} && \text{(G.7)} \end{aligned}$$

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

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

Summing up Eq. (G.7) with Eq. (G.6), we obtain the following:

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

where I is the summation of $I(a)$ and $I(b)$. This concludes our proof of Lemma 11 by bringing in the definitions of $V(\boldsymbol{z}_{k+1}^{\text{ag}}, \boldsymbol{z}^*)$, $V(\boldsymbol{z}_k^{\text{ag}}, \boldsymbol{z}^*)$. □

G.6. Proof of Lemma 13

Proof.[Proof of Lemma 13] By Young's inequality and Cauchy-Schwarz inequality, we have that

$$\begin{aligned} \left\| \boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}} \right\|^2 &\leq 2 \left\| \boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k \right\|^2 + 2 \left\| \boldsymbol{z}_k - \boldsymbol{z}_{k-\frac{1}{2}} \right\|^2 \\ &\overset{(a)}{\leq} 2 \left\| \boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k \right\|^2 + 2\eta_{k-1}^2 L_H^2 \left\| \boldsymbol{z}_{k-\frac{1}{2}} - \boldsymbol{z}_{k-\frac{3}{2}} \right\|^2 \\ &\overset{(b)}{\leq} 2 \left\| \boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k \right\|^2 + c \left\| \boldsymbol{z}_{k-\frac{1}{2}} - \boldsymbol{z}_{k-\frac{3}{2}} \right\|^2 \end{aligned} \tag{G.8}$$

where (a) is by Line (3) and (5) of Algorithm 1 and the definition of L_H, and (b) is by the condition in Lemma 13 that $\eta_k L_H \leq \sqrt{c/2}$.

Recursively applying the above gives

$$\left\| \boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}} \right\|^2 \leq 2c^k \sum_{\ell=0}^{k} c^{-\ell} \left\| z_{\ell+\frac{1}{2}} - z_\ell \right\|^2 \tag{G.9}$$

Indeed, from (G.8)

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

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

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

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

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

which gives (G.9). □

G.7. Proof of Lemma 15

Proof.[Proof of Lemma 15] We recall that we denote

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

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

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

Recalling that by Assumption 2, $\mathbb{E}||\Delta_h^{k+\frac{1}{2}}||^2 \leq \sigma_H^2$ and $\mathbb{E}||\Delta_h^{k-\frac{1}{2}}||^2 \leq \sigma_H^2$, we conclude our proof of Lemma 15. □

G.8. Proof of Lemma 16

Proof.[Proof of Lemma 16] By inequality (E.10), we have

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

The inner product term can be decomposed into

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

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

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

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

$$
\begin{aligned}
&\mathbb{E}V(\boldsymbol{z}_{k+1}^{\text{ag}}, \omega_{\boldsymbol{z}}) - (1-\alpha_k)\mathbb{E}V(\boldsymbol{z}_k^{\text{ag}}, \omega_{\boldsymbol{z}}) \\
&\le \frac{\alpha_k \eta_k}{2}\left[2L_H^2 \mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}||^2 + 3\sigma_H^2\right] + \frac{\alpha_k \eta_k}{2(1-r)}\mathbb{E}||\Delta_f^k||^2 \\
&\quad + \frac{\alpha_k}{2\eta_k}\mathbb{E}\left[||\boldsymbol{z}_k - \boldsymbol{z}||^2 - ||\boldsymbol{z}_{k+1} - \boldsymbol{z}||^2\right] - \left(\frac{r\alpha_k}{2\eta_k} - \frac{L\alpha_k^2}{2}\right)\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2.
\end{aligned}
\tag{G.10}
$$

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

$$
\begin{aligned}
&\mathbb{E}\left\|\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_{k-\frac{1}{2}}\right\|^2 \le 2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + 2\mathbb{E}||\boldsymbol{z}_k - \boldsymbol{z}_{k-\frac{1}{2}}||^2 \\
&= 2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + 2\eta_{k-1}^2 \mathbb{E}||\widetilde{H}(\boldsymbol{z}_{k-\frac{1}{2}}) - \widetilde{H}(\boldsymbol{z}_{k-\frac{3}{2}})||^2 \\
&= 2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + 2\eta_{k-1}^2 \mathbb{E}||H(\boldsymbol{z}_{k-\frac{1}{2}}) - H(\boldsymbol{z}_{k-\frac{3}{2}}) + \Delta_h^{k-\frac{3}{2}}||^2 + 2\eta_{k-1}^2 \mathbb{E}||\Delta_h^{k-\frac{1}{2}}||^2 \\
&\le 2\mathbb{E}||\boldsymbol{z}_{k+\frac{1}{2}} - \boldsymbol{z}_k||^2 + 4\eta_{k-1}^2 L_H^2 \mathbb{E}||\boldsymbol{z}_{k-\frac{1}{2}} - \boldsymbol{z}_{k-\frac{3}{2}}||^2 + 6\eta_{k-1}^2 \sigma_H^2 \\
&= 2\sum_{\ell=0}^{k} c^{k-\ell}\mathbb{E}||\boldsymbol{z}_{\ell+\frac{1}{2}} - \boldsymbol{z}_\ell||^2 + 6\sum_{\ell=0}^{k} c^{k-\ell}\eta_{\ell-1}^2 \sigma_H^2
\end{aligned}
\tag{G.11}
$$

Bringing Eq. (G.11) into (G.10), we have

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

and that concludes our proof of Lemma 16. □