Optimal Extragradient-Based Algorithms for Stochastic Variational Inequalities with Separable Structure

Huizhuo Yuan[◇] **Chris Junchi Li**[†] **Gauthier Gidel**[‡]
Michael I. Jordan[†,□] **Quanquan Gu**[◇] **Simon S. Du**[⋆]
[◇] Department of Computer Science, University of California, Los Angeles
{hzyuan, qgu}@cs.ucla.edu
[†] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
{junchili, jordan}@cs.berkeley.edu
[‡] DIRO, Université de Montréal and Mila
gauthier.gidel@umontreal.ca
[□] Department of Statistics, University of California, Berkeley
[⋆] Paul G. Allen School of Computer Science and Engineering, University of Washington
ssdu@cs.washington.edu

Abstract

We consider the problem of solving stochastic monotone variational inequalities with a separable structure using a stochastic first-order oracle. Building on standard extragradient for variational inequalities we propose a novel algorithm—stochastic *accelerated gradient-extragradient* (AG-EG)—for strongly monotone variational inequalities (VIs). Our approach combines the strengths of extragradient and Nesterov acceleration. By showing that its iterates remain in a bounded domain and applying scheduled restarting, we prove that AG-EG has an optimal convergence rate for strongly monotone VIs. Furthermore, when specializing to the particular case of bilinearly coupled strongly-convex-strongly-concave saddle-point problems, including bilinear games, our algorithm achieves fine-grained convergence rates that match the respective lower bounds, with the stochasticity being characterized by an additive statistical error term that is optimal up to a constant prefactor.

1 Introduction

The variational inequality (VI) problem plays a central role in a wide range of optimization problems with convex structure, including convex minimization, saddle-point problems, and games [Facchinei and Pang, 2003, Nemirovski, 2004, Nemirovski et al., 2009, Juditsky et al., 2011, Jordan et al., 2023]. A general VI problem aims to find a solution $\boldsymbol{z}^* \in \mathcal{Z}$ that satisfies:

$$\langle \mathcal{W}(\boldsymbol{z}^*), \boldsymbol{z}^* - \boldsymbol{z} \rangle \leq 0, \quad \forall \boldsymbol{z} \in \mathcal{Z}, \tag{1}$$

where $\mathcal{Z}$ is a finite-dimensional closed and convex feasible set and $\mathcal{W}(\cdot)$ is a monotone operator in the following form:

$$\mathcal{W}(\boldsymbol{z}) = \nabla \mathcal{F}(\boldsymbol{z}) + \mathcal{H}(\boldsymbol{z}) + J'(\boldsymbol{z}) \equiv \mathbb{E}_\xi[\nabla \widetilde{\mathcal{F}}(\boldsymbol{z}; \xi)] + \mathbb{E}_\zeta[\widetilde{\mathcal{H}}(\boldsymbol{z}; \zeta)] + J'(\boldsymbol{z}), \tag{2}$$

where $\mathcal{F}$ is continuously differentiable with L-Lipschitz continuous gradient and is μ-strongly convex, $\mathcal{H}$ is an M-Lipschitz monotone operator, $J' \in \partial J$ is the subgradient of a simple and convex function, ξ and ζ are drawn from distributions $\mathcal{D}_\xi$ and $\mathcal{D}_\zeta$, respectively. This formulation captures a separable structure in which $\mathcal{H}$ usually models the competing forces in a system, and J models

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

a nonsmooth factor. In addition, we consider the stochastic setting where we can only access $\nabla \mathcal{F}$ and $\mathcal{H}$ through their unbiased estimators $\nabla \widetilde{\mathcal{F}}(\boldsymbol{z}; \xi)$ and $\widetilde{\mathcal{H}}(\boldsymbol{z}; \zeta)$ respectively.

A notable instance of the VI problem (1) with separable structure (2) is the widely studied *bilinearly coupled strongly-convex-strongly-concave saddle-point problem*:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \max_{\boldsymbol{y} \in \mathbb{R}^m} \mathscr{F}(\boldsymbol{x}, \boldsymbol{y}) = F(\boldsymbol{x}) + H(\boldsymbol{x}, \boldsymbol{y}) - G(\boldsymbol{y}) \equiv \mathbb{E}_\xi \left[f(\boldsymbol{x}; \xi) \right] + \mathbb{E}_\zeta \left[h(\boldsymbol{x}, \boldsymbol{y}; \zeta) \right] - \mathbb{E}_\xi \left[g(\boldsymbol{y}; \xi) \right], \tag{3}$$

where $H(\boldsymbol{x}, \boldsymbol{y}) \equiv \boldsymbol{x}^\top \mathbf{B} \boldsymbol{y} - \boldsymbol{x}^\top \mathbf{u}_{\boldsymbol{x}} + \mathbf{u}_{\boldsymbol{y}}^\top \boldsymbol{y}$ is the bilinear coupling function with the coupling matrix $\mathbf{B} \in \mathbb{R}^{n \times m}$. Note that (3) is a special instance of (1) when taking $\mathcal{F}(\boldsymbol{z}) = F(\boldsymbol{x}) + G(\boldsymbol{y})$, $\mathcal{H}(\boldsymbol{z}) = [\nabla_{\boldsymbol{x}} H(\boldsymbol{x}, \boldsymbol{y}); -\nabla_{\boldsymbol{y}} H(\boldsymbol{x}, \boldsymbol{y})]$ and $J = 0$. In addition to a wide range of applications in economics, problems of form (3) are becoming increasingly important in machine learning. For instance, (3) appears in reinforcement learning, differentiable games, regularized empirical risk minimization, and robust optimization formulations. It can also be seen as a local approximation of the nonconvex-nonconcave minimax games—e.g., the generative adversarial network (GAN) [Goodfellow et al., 2020]—around a local Nash equilibrium [Mescheder et al., 2017, Nagarajan and Kolter, 2017].

In this paper, we aim to improve the efficiency of solving (1) by utilizing the structural information of the monotone operator in (2). More specifically, we consider the case when $\mathcal{F}$ is strongly monotone, or zero. Although optimal convergence results have been obtained for the monotone VI problem (1) [Chen et al., 2017] as well as the special case of convex-concave saddle-point problem with bilinear coupling (3) [Chen et al., 2014], it remains open how to design an optimal algorithm for the strongly monotone VI problem. Notably, for the special case (3) when F and/or G are strongly convex, several concurrent works have independently obtained the optimal convergence rates [Kovalev et al., 2022, Thekumparampil et al., 2022, Jin et al., 2022, Metelev et al., 2022, Li et al., 2022b]. On the other hand, when both F and G are zero, optimal convergence results have been obtained by Li et al. [2022a] and the accelerated-gradient optimistic gradient approach [Li et al., 2022b]. We defer a more complete overview of related work to the appendix.

1.1 Main Contributions

We start with the strongly monotone VI problem in an unbounded feasible set, extending the scope of recent work such as Jin et al. [2022] and going beyond earlier studies that focus on nonstrongly monotone VIs in a bounded feasible set [Juditsky et al., 2011, Chen et al., 2017].[1] We propose a class of algorithms named stochastic *accelerated gradient-extragradient* (AG-EG), which combine Nesterov acceleration with the extragradient method. By employing either a strong-convexity shifting technique or a scheduled restarting scheme, our algorithm achieves convergence rates that match the lower bounds for the general strongly monotone VI problem (1), the special SC-SC saddle-point problem (3), and bilinear games, in both deterministic and stochastic settings, thus providing a unified optimal solution. In sharp contrast to the accelerated mirror-prox (AMP) algorithm proposed by Chen et al. [2017], Jordan et al. [2023], our analysis does not rely on the boundedness of the feasible set $\mathcal{Z}$, which makes our algorithm projection-free. We also extend our algorithm to VIs with bounded feasible set and/or nondifferentiable convex regularization through proximal mapping. We summarize our contributions as follows:

(1) We present a direct approach for separable strongly monotone VIs, where the iteration complexity lower bound due to Zhang et al. [2022] is matched as $\mathcal{O}\left(\sqrt{\frac{L}{\mu}} + \frac{M}{\mu} + \frac{\sigma^2}{\mu^2 \varepsilon^2}\right) \log\left(\frac{L}{\mu} \frac{1}{\varepsilon}\right)$, which admits a sharp near-unity coefficient [§2.3, Theorem 2.3]. Here σ^2 is the weighted, uniform variance bound on the stochastic gradient and stochastic operator.

(2) We also present a stochastic AG-EG algorithm equipped with scheduled restarting, which achieves the sharpest possible iteration complexity of $\mathcal{O}\left(\left(\sqrt{\frac{L}{\mu}} + \frac{M}{\mu}\right) \log\left(\frac{1}{\varepsilon}\right) + \frac{\sigma^2}{\mu^2 \varepsilon^2}\right)$ for finding an ε-optimal point. The deterministic part matches the complexity lower bound in Zhang et al. [2022], while the stochastic part matches the optimal statistical error.

[1] VIs in an unbounded feasible set is more difficult to solve because existing algorithms and analyses crucially rely on the boundedness of the feasible set.

When specializing the VI problem to bilinearly coupled SC-SC saddle-point problems, our results have the following implications:

Strongly-convex-strongly-concave (SC-SC) Saddle-Point Problem. For the class of SC-SC saddle-point problems, the stochastic AG-EG descent-ascent Algorithm 1, equipped with scaling reduction, achieves an iteration complexity of

$$\mathcal{O}\left(\left(\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}\right) \log\left(\frac{1}{\varepsilon}\right) + \frac{\sigma^2}{\mu_F^2 \varepsilon^2}\right), \tag{4}$$

where $F : \mathbb{R}^n \to \mathbb{R}$ is L_F-smooth and μ_F-strongly convex, $G : \mathbb{R}^m \to \mathbb{R}$ is L_G-smooth and μ_G-strongly convex. When the optimization problem is deterministic, the complexity upper bound matches the lower bound of Zhang et al. [2022][§3.1, Corollary 2.8].

Bilinear Games. For bilinear games ($\nabla f(\boldsymbol{x}; \xi) = \mathbf{0}$ and $\nabla g(\boldsymbol{y}; \xi) = \mathbf{0}$ almost surely), Algorithm 1, equipped with scheduled restarting achieves an iteration complexity of

$$\mathcal{O}\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}} \log\left(\frac{\sqrt[4]{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}}{\sigma_{\mathrm{Bil}}}\right) + \frac{\sigma_{\mathrm{Bil}}^2}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\varepsilon^2}\right), \tag{5}$$

where σ_{Bil}^2 is the variance of the stochastic gradient on the bilinear coupling term. When there is no randomness, this complexity result reduces to $\mathcal{O}\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}} \log\left(\frac{1}{\varepsilon}\right)\right)$ for bilinear games, matching the lower bound of Ibrahim et al. [2020] [§3.2, Corollary 3.3].[2]

Organization. The rest of this paper is organized as follows. Section 2 proposes the Accelerated Gradient-Extragradient Descent-Ascent algorithm for strongly monotone VIs, showing that it achieves an accelerated convergence rate, and extending to VIs with bounded domains with proximal operator. Section 3 discusses two specific instances of saddle-point problems, where our proposed AG-EG algorithm has a convergence rate that matches the corresponding lower bounds. Finally, Section 4 summarizes our results and suggests future directions.

Notation. Let $\lambda_{\max}(\mathbf{M})$ (resp. $\lambda_{\min}(\mathbf{M})$ be the largest (resp. smallest) eigenvalue of a real symmetric matrix $\mathbf{M}$. Let $a \vee b \equiv \max(a, b)$ (resp. $a \wedge b \equiv \min(a, b)$) denote the maximum (resp. minimum) value of two reals a, b. For two nonnegative real sequences (a_n) and (b_n), we write $a_n = \mathcal{O}(b_n)$ or $a_n \lesssim b_n$ (resp. $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$) to denote $a_n \leq C b_n$ (resp. $a_n \geq C b_n$) for all $n \geq 1$ for a positive, numerical constant C, and let $a_n \asymp b_n$ if both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold. We also let $a_n = \tilde{\mathcal{O}}(b_n)$ denote $a_n \leq C b_n$ where C hides a polylogarithmic factor in problem-dependent constants. We let $[\boldsymbol{x}; \boldsymbol{y}] \in \mathbb{R}^{n+m}$ concatenate two vectors $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$. Finally for two real symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, we denote $\mathbf{A} \preceq \mathbf{B}$ (resp. $\mathbf{A} \succeq \mathbf{B}$) when $\mathbf{v}^\top(\mathbf{A} - \mathbf{B})\mathbf{v} \leq 0$ (resp. $\mathbf{v}^\top(\mathbf{A} - \mathbf{B})\mathbf{v} \geq 0$) holds for all vectors $\mathbf{v}$.

2 Accelerated Gradient-Extragradient Descent-Ascent Algorithm

In this section, we focus on accelerating the extragradient algorithm for the strongly monotone VI problem in (1) with separable structure (2). Our algorithm design draws inspiration from the work of Chen et al. [2017] on the stochastic Accelerated MirrorProx (AMP) algorithm for nonstrongly monotone VIs. The AMP algorithm applies Nesterov-type acceleration on top of the mirror-prox method [Korpelevich, 1976, Nemirovski, 2004] and attains the optimal iteration complexity of $\mathcal{O}\left(\sqrt{\frac{L}{\varepsilon}} + \frac{M}{\varepsilon}\right)$. However, the big-O notation hides the diameter of the feasible set, and the existing theory for the AMP algorithm can only deal with VIs with bounded domain. Our algorithm not only achieves the optimal convergence rates for the strongly monotone VI problem with separable structure but we also remove the dependency on the diameter of the feasible set. Therefore, our algorithm can deal with VIs with unbounded domains.

[2]For the function class of bilinear games, we assume that $n = m$ where $\mathbf{B}$ is a nonsingular square matrix, so that $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top) > 0$ and the complexity makes sense. See §3.2 for more on this.

Throughout §2, we maintain conceptual simplicity by presenting all our algorithm designs in the deterministic setting, while presenting the convergence results in the more general stochastic setting. These results can be easily reduced to the deterministic setting when the stochastic noise vanishes.

2.1 Setting and Assumptions

In this section, we formally introduce our assumptions. We first state the smoothness and monotonicity assumptions that we impose on $\mathcal{F}$ and $\mathcal{H}$.

Assumption 2.1 (Monotonicity, strong convexity and smoothness) *We assume that function $\mathcal{F}(\cdot)$ is continuously differentiable with L-Lipschitz continuous gradient and is μ-strongly convex. That is, for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$,*

$$\tfrac{\mu}{2}\|\boldsymbol{z}-\boldsymbol{z}'\|^2 \le \mathcal{F}(\boldsymbol{z}) - \mathcal{F}(\boldsymbol{z}') - \nabla\mathcal{F}(\boldsymbol{z}')^\top(\boldsymbol{z}-\boldsymbol{z}') \le \tfrac{L}{2}\|\boldsymbol{z}-\boldsymbol{z}'\|^2.$$

Furthermore, operator $\mathcal{H}(\cdot)$ is monotone and M-Lipschitz in the sense that for any $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$,

$$\langle \mathcal{H}(\boldsymbol{z}) - \mathcal{H}(\boldsymbol{z}'), \boldsymbol{z}-\boldsymbol{z}'\rangle \ge 0, \quad \|\mathcal{H}(\boldsymbol{z}) - \mathcal{H}(\boldsymbol{z}')\| \le M\|\boldsymbol{z}-\boldsymbol{z}'\|.$$

Second, we impose assumptions on the noise variance.

Assumption 2.2 (Unbiased gradients and variance bounds) *We assume that $\boldsymbol{z} \in \mathcal{Z}$, samples $\xi \sim \mathcal{D}_\xi$ and $\zeta \sim \mathcal{D}_\zeta$ are drawn from given distributions such that the following conditions hold: $\mathbb{E}_\xi[\nabla\tilde{\mathcal{F}}(\boldsymbol{z};\xi)] = \nabla\mathcal{F}(\boldsymbol{z})$, $\mathbb{E}_\zeta[\tilde{\mathcal{H}}(\boldsymbol{z};\zeta)] = \mathcal{H}(\boldsymbol{z})$, and*

$$\mathbb{E}_\xi\left[\|\nabla\tilde{\mathcal{F}}(\boldsymbol{z};\xi) - \nabla\mathcal{F}(\boldsymbol{z})\|^2\right] \le \sigma_{\mathrm{Str}}^2, \quad \mathbb{E}_\zeta\left[\|\tilde{\mathcal{H}}(\boldsymbol{z};\zeta) - \mathcal{H}(\boldsymbol{z})\|^2\right] \le \sigma_{\mathrm{Bil}}^2. \tag{6}$$

For all results in this work, we suppose that Assumptions 2.1 and 2.2 hold with appropriate parameter settings. Given a desired accuracy $\varepsilon > 0$, our goal is to find an *ε-optimal point* defined as:

Definition 2.1 (ε-Optimal point) *A point $\boldsymbol{z} \in \mathcal{Z}$ is called an* ε-optimal point *for the VI problem in* (1) *if $\|\boldsymbol{z} - \boldsymbol{z}^*\| \le \varepsilon$.*

2.2 The ExtraGradient (EG) Algorithm

We first consider the case where $\mathcal{Z}$ is the entire space $\mathbb{R}^n$ and the objective is smooth ($J = 0$). The extragradient (EG) algorithm, introduced by Korpelevich [1976], is designed to address cyclic behavior in saddle-point problems by introducing an extrapolated point for gradient evaluation. In the context of VI problems (1), let $\boldsymbol{z}_t$ represents the t-th iterate of the EG algorithm. The update rule of EG is as follows:

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \eta\mathcal{W}\big(\boldsymbol{z}_t - \eta\mathcal{W}(\boldsymbol{z}_t)\big), \tag{7}$$

where $\eta > 0$ is the step size. For a L-smooth and μ-strongly monotone operator $\mathcal{W}$, Tseng [1995], Mokhtari et al. [2020], Gidel et al. [2019a] have shown that the EG algorithm achieves an iteration complexity of $\mathcal{O}(\kappa\log(1/\varepsilon))$, where $\kappa = L/\mu$ denotes the condition number of the problem.

2.3 Accelerating the ExtraGradient Algorithm, Direct Approach

The convergence rate of the EG algorithm is far from optimal for the strongly monotone VI problem in (1) with separable structure (2). Firstly, the update rule in (7) takes $\mathcal{W}$ as a whole without utilizing the separable structure. This prevents us from exploiting the properties of $\nabla\mathcal{F}$. Secondly, in the case of bilinear games, the established lower bound for EG is $\Omega(\sqrt{\kappa}\log(1/\varepsilon))$ rather than $\Omega(\kappa\log(1/\varepsilon))$. This discrepancy highlights the potential for accelerating the EG algorithm in various directions. We first rewrite the EG update rule in (7) as follows:

$$\begin{aligned}\boldsymbol{z}_{t-\frac{1}{2}} &= \boldsymbol{z}_{t-1} - \eta\mathcal{W}(\boldsymbol{z}_{t-1}) = \boldsymbol{z}_{t-1} - \eta\big(\mathcal{H}(\boldsymbol{z}_{t-1}) + \nabla\mathcal{F}(\boldsymbol{z}_{t-1})\big),\\ \boldsymbol{z}_t &= \boldsymbol{z}_{t-1} - \eta\mathcal{W}(\boldsymbol{z}_{t-\frac{1}{2}}) = \boldsymbol{z}_t - \eta\big(\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}) + \nabla\mathcal{F}(\boldsymbol{z}_{t-\frac{1}{2}})\big).\end{aligned} \tag{8}$$

To accelerate the process based on $\nabla\mathcal{F}$, we consider Nesterov's second acceleration scheme on minimizing a single convex function $\mathcal{F}$ [Tseng, 2008, Lan and Zhou, 2018, Lin et al., 2020c]:

$$\boldsymbol{z}_{t-1}^{\mathrm{md}} = (1-\alpha_t)\boldsymbol{z}_{t-1}^{\mathrm{ag}} + \alpha_t\boldsymbol{z}_{t-1}, \boldsymbol{z}_t = \boldsymbol{z}_{t-1} - \frac{\eta}{\alpha_t}\nabla\mathcal{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}}), \boldsymbol{z}_t^{\mathrm{ag}} = (1-\alpha_t)\boldsymbol{z}_{t-1}^{\mathrm{ag}} + \alpha_t\boldsymbol{z}_t, \tag{9}$$

where α_t is the extrapolation step size in the standard three-line Nesterov scheme. Here we adopt the notation $\boldsymbol{z}^{\text{md}}$ and $\boldsymbol{z}^{\text{ag}}$ to indicate the middle point and the aggregated point [Chen et al., 2017], respectively. Next, to achieve acceleration, we replace the gradient of $\nabla\mathcal{F}$ evaluated at both $\boldsymbol{z}_{t-1}$ and $\boldsymbol{z}_{t-\frac{1}{2}}$ in (8) by the gradient evaluated at the extrapolated point $\boldsymbol{z}_{t-1}^{\text{md}}$ in (9). Furthermore, we shift the index of $\boldsymbol{z}^{\text{ag}}$ by $\frac{1}{2}$ to indicate the use of $\boldsymbol{z}_{t-\frac{1}{2}}$ instead of $\boldsymbol{z}_t$ in the $\boldsymbol{z}^{\text{ag}}$ update in (9). In addition, we take into account the μ-strong convexity of $\mathcal{F}$ and shift the gradient of the strongly convex part $\nabla_{\boldsymbol{z}}\left[\frac{\mu}{2}\|\boldsymbol{z}-\boldsymbol{z}_0\|^2\right] = \mu(\boldsymbol{z}-\boldsymbol{z}_0)$ from $\nabla\mathcal{F}(\boldsymbol{z})$ to $\mathcal{H}(\boldsymbol{z})$ as $\mathcal{W}(\boldsymbol{z}) = (\nabla\mathcal{F}(\boldsymbol{z}) - \mu(\boldsymbol{z}-\boldsymbol{z}_0)) + (\mathcal{H}(\boldsymbol{z}) + \mu(\boldsymbol{z}-\boldsymbol{z}_0))$, we obtain the following update rule for a direct version of an accelerated EG algorithm (different step size schemes for η_t are required for different algorithmic designs):

$$\begin{cases} \boldsymbol{z}_{t-1}^{\text{md}} = (1-\alpha_t)\boldsymbol{z}_{t-\frac{3}{2}}^{\text{ag}} + \alpha_t \boldsymbol{z}_{t-1}, \\ \boldsymbol{z}_{t-\frac{1}{2}} = \boldsymbol{z}_{t-1} - \eta_t\left(\mathcal{H}(\boldsymbol{z}_{t-1}) + \nabla\mathcal{F}(\boldsymbol{z}_{t-1}^{\text{md}}) - \mu(\boldsymbol{z}_{t-1}^{\text{md}} - \boldsymbol{z}_{t-1})\right), \\ \boldsymbol{z}_t = \boldsymbol{z}_{t-1} - \eta_t\left(\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}) + \nabla\mathcal{F}(\boldsymbol{z}_{t-1}^{\text{md}}) - \mu(\boldsymbol{z}_{t-1}^{\text{md}} - \boldsymbol{z}_{t-\frac{1}{2}})\right), \\ \boldsymbol{z}_{t-\frac{1}{2}}^{\text{ag}} = (1-\alpha_t)\boldsymbol{z}_{t-\frac{3}{2}}^{\text{ag}} + \alpha_t \boldsymbol{z}_{t-\frac{1}{2}}. \end{cases} \tag{10}$$

We call the algorithm in (10) the *accelerated gradient-extragradient, direct approach* (AG-EG-Direct), and postpone its full description to Algorithm 2 in §C.1. The final output of the direct approach is $\boldsymbol{z}_T$ after T iterates. The following theorem records the convergence rate and iteration complexity of AG-EG (direct approach).

Theorem 2.3 (Convergence of stochastic AG-EG, direct approach) *Suppose Assumptions 2.1 and 2.2 hold. Fix any $r \in (0,1)$, $\beta \in (0,\infty)$, let $\kappa_\beta = \frac{L}{\mu} + \frac{(1+\beta)M^2}{\mu^2}$ and set the step size upper bound $\bar{\alpha} \equiv \frac{r}{1+\sqrt{1+r\kappa_\beta}}$. For any sequence of step sizes $\alpha_t \in (0, \bar{\alpha}]$ and $\eta_t = \frac{\alpha_t}{\mu}$, the iterates of stochastic AG-EG (direct approach) satisfy that for all $t = 1, \ldots, T$, we have*

$$\mathbb{E}\left\|\boldsymbol{z}_t - \boldsymbol{z}^*\right\|^2 \le \left\|\boldsymbol{z}_0 - \boldsymbol{z}^*\right\|^2 \left(\tfrac{L}{\mu}+1\right)\prod_{s=1}^{t}(1-\alpha_s) + \tfrac{3\sigma^2}{\mu^2}\sum_{s=1}^{t}\alpha_s^2\prod_{\tau=s+1}^{t}(1-\alpha_\tau), \tag{11}$$

where we define $\sigma = \frac{1}{\sqrt{3}}\sqrt{\frac{1}{1-r}\sigma_{\text{Str}}^2 + (2+\frac{1}{\beta})\sigma_{\text{Bil}}^2}$.

In the rest of the paper, we use the same definition σ as in Theorem 2.3. The proof of Theorem 2.3 is provided in §D.4. We further note that one possible choice of step size is to let $\alpha_t \equiv \alpha$, such that (11) reduces to

$$\mathbb{E}\left\|\boldsymbol{z}_t - \boldsymbol{z}^*\right\|^2 \le \left\|\boldsymbol{z}_0 - \boldsymbol{z}^*\right\|^2 \left(\tfrac{L}{\mu}+1\right) e^{-\alpha t} + \tfrac{3\sigma^2}{\mu^2}\alpha.$$

For any given $T \ge 1$, by choosing the optimal $\alpha = \frac{1}{T}\left(1 + \log\left(\frac{\mu^2 T}{3\sigma^2}\left(\frac{L}{\mu}+1\right)\left\|\boldsymbol{z}_0 - \boldsymbol{z}^*\right\|^2\right)\right) \wedge \bar{\alpha}$, (11) implies

$$\mathbb{E}\|\boldsymbol{z}_T - \boldsymbol{z}^*\|^2 \le \left\|\boldsymbol{z}_0 - \boldsymbol{z}^*\right\|^2 \left(\tfrac{L}{\mu}+1\right) e^{-\bar{\alpha}T} + \tfrac{3\sigma^2}{\mu^2 T}\left(1 + \log\left(\tfrac{\mu^2 T}{3\sigma^2}\left(\tfrac{L}{\mu}+1\right)\left\|\boldsymbol{z}_0 - \boldsymbol{z}^*\right\|^2\right)\right).$$

Prescribing the desired accuracy $\varepsilon > 0$, the iteration complexity to output an ε-optimal minimax point is [3]

$$\mathcal{O}\left(\left(\sqrt{\tfrac{L}{\mu}} + \tfrac{M}{\mu} + \tfrac{\sigma^2}{\mu^2\varepsilon^2}\right)\log\left(\left(\tfrac{L}{\mu}+1\right)\left\|\boldsymbol{z}_0 - \boldsymbol{z}^*\right\|^2/\varepsilon^2\right)\right).$$

We conjecture that the logarithmic factor in the optimal statistical rate $\frac{\sigma^2}{\mu^2\varepsilon^2}$ is removable using a proper diminishing step size, a possibility that we reserve for future study. In the setting of deterministic optimization, setting $\sigma = 0$ and $r \to 1^-$, $\beta \to 0^+$ in Theorem 2.3, we obtain the optimal iteration complexity bound as follows:

$$\left(1 + \sqrt{1 + \tfrac{L}{\mu} + \tfrac{M^2}{\mu^2}}\right)\log\left(\left(\tfrac{L}{\mu}+1\right)/\varepsilon^2\right). \tag{12}$$

[3] Throughout this work, we focus on the iteration complexity whereas the required number of queries to the stochastic gradient oracle is three times the iteration complexity.

Algorithm 1 Stochastic AcceleratedGradient-ExtraGradient (AG-EG) Descent-Ascent Algorithm, with Scheduled Restarting

Require: Initialization $\boldsymbol{z}_0^{[0]}$, total number of epochs $\mathscr{S} \geq 1$, total number of per-epoch iterates $(T_s : s = 1, \ldots, \mathscr{S})$, stepsizes $(\alpha_t, \eta_t : t = 1, 2, \ldots)$.
for $s = 1, 2, \ldots, \mathscr{S}$ **do**
  Set $\boldsymbol{z}_{-\frac{1}{2}}^{\mathrm{ag}} \leftarrow \boldsymbol{z}_0^{[s-1]}, \boldsymbol{z}_0 \leftarrow \boldsymbol{z}_0^{[s-1]}, \boldsymbol{z}_0^{\mathrm{md}} \leftarrow \boldsymbol{z}_0^{[s-1]}$
  for $t = 1, 2, \ldots, T_s$ **do**
    Draw samples $\xi_{t-\frac{1}{2}} \sim \mathcal{D}_\xi$ from oracle, and also $\zeta_{t-\frac{1}{2}}, \zeta_t \sim \mathcal{D}_\zeta$ independently from oracle
    $\boldsymbol{z}_{t-\frac{1}{2}} \leftarrow \boldsymbol{z}_{t-1} - \eta_t \left(\tilde{\mathcal{H}}(\boldsymbol{z}_{t-1}; \zeta_{t-\frac{1}{2}}) + \nabla \tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) \right)$
    $\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} \leftarrow (1 - \alpha_t) \boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}} + \alpha_t \boldsymbol{z}_{t-\frac{1}{2}}$
    $\boldsymbol{z}_t \leftarrow \boldsymbol{z}_{t-1} - \eta_t \left(\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t) + \nabla \tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) \right)$
    $\boldsymbol{z}_t^{\mathrm{md}} \leftarrow (1 - \alpha_{t+1}) \boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} + \alpha_{t+1} \boldsymbol{z}_t$
  end for
  Set $\boldsymbol{z}_0^{[s]} \leftarrow \boldsymbol{z}_{T_s - \frac{1}{2}}^{\mathrm{ag}}$ {//Warm-start using the output of the previous epoch}
end for
Output: $\boldsymbol{z}_0^{[\mathscr{S}]}$

Remark 2.4 *Our complexity bounds* fundamentally differs *from the previous analysis [Chen et al., 2017, Jordan et al., 2023] for separable smooth (strongly) monotone VIs. The convergence results in previous studies are dependent on the diameter of the domain, whereas our convergence rate is independent of the domain parameters and eliminates the need for projection onto a bounded domain. Moreover, our contributions go beyond those of Chen et al. [2017] by extending the analysis to the strongly monotone case. In comparison with Jordan et al. [2023], we design an algorithm where $\nabla \mathcal{F}$ is strongly monotone and resolve the open problem of extending the analysis to the stochastic case. Additionally, our complexity bound in* (12) *indicates a near-unity coefficient on the condition-number exponent, improving the corresponding coefficient in Chen et al. [2017, Theorem 15] by an asymptotic factor of* 4*.*

The direct approach, which reduces to EG when $\nabla \mathcal{F} = 0$ and $\mu = 0$, falls short of attaining optimality within the specific regime of bilinear games. In the next subsection, we will introduce a new algorithm that can overcome this limitation.

2.4 Accelerating the ExtraGradient Algorithm with Scheduled Restarting

In this subsection, we solve problem (1) by further accelerating the stochastic EG algorithm. Rather than directly relying on the strong monotonicity of $\nabla \mathcal{F}$, the inner updates of our new algorithm are identical to the updates in (10) with $\mu = 0$. Due to the domain-independent nature of our analysis, we can apply the scheduled restarting technique [O'donoghue and Candes, 2015, Roulet and d'Aspremont, 2017, Renegar and Grimmer, 2022] to the outer loop, accelerating the algorithm from sublinear convergence to linear convergence. In addition, the output of our algorithm is the aggregated point $\boldsymbol{z}_{T-\frac{1}{2}}^{\mathrm{ag}}$ after T iterates. We present the full algorithm in Algorithm 1.

We first present the convergence rate of a single epoch (i.e., the inner loop) of Algorithm 1 in Theorem 2.5. To accommodate more flexibility in the choice of parameters, we introduce three constants r, β, and C in the theorem statement.

Theorem 2.5 (Convergence of stochastic AG-EG, one epoch) *Suppose Assumptions 2.1 and 2.2 hold. For any fixed epoch length $T \geq 1$, any constant $r \in (0, 1)$, $\beta \in (0, \infty)$, $C \in (0, \infty)$, choose step sizes $\alpha_t = \frac{2}{t+1}$ and η_t such that*

$$\frac{t}{\eta_t} = \frac{2}{r} L \vee B + \sqrt{\frac{1+\beta}{r}} M t, \tag{13}$$

where $B = \frac{\sigma\sqrt{T}(T+1)}{C\sqrt{\mathbb{E}\|\boldsymbol{z}_0-\boldsymbol{z}^*\|^2}}$. *The output* $\boldsymbol{z}^{ag}_{T-\frac{1}{2}}$ *of a single epoch of Algorithm 1 satisfies*

$$\mathbb{E}\left\|\boldsymbol{z}^{ag}_{T-\frac{1}{2}}-\boldsymbol{z}^*\right\|^2 \le \frac{2}{\mu(T+1)}\left(\frac{2L}{rT}+A\sqrt{\frac{1+\beta}{r}}M\right)\mathbb{E}\left\|\boldsymbol{z}_0-\boldsymbol{z}^*\right\|^2+\frac{2(\frac{1}{C}+C)\sigma}{\mu\sqrt{T}}\sqrt{\mathbb{E}\left\|\boldsymbol{z}_0-\boldsymbol{z}^*\right\|^2}, \tag{14}$$

where the prefactor $A \equiv 1+C^2B\eta_1 \le 1+C^2$ *reduces to 1 when* $\sigma = 0$.

The proof of Theorem 2.5 is provided in §D.3. We make a few remarks on Theorem 2.5 as follows:

Remark 2.6 *In the setting of deterministic optimization, by taking* $\sigma = 0$, $r \to 1^-$, $\beta \to 0^+$ *in our analysis, with step size choice* $\eta_t = \frac{t}{2L+Mt}$, *we obtain that*

$$\|\boldsymbol{z}^{ag}_{T-\frac{1}{2}}-\boldsymbol{z}^*\|^2 \le \frac{2}{\mu(T+1)}\left(\frac{2L}{T}+M\right)\|\boldsymbol{z}_0-\boldsymbol{z}^*\|^2, \tag{15}$$

In this setting, the algorithm is independent of B *and requires no knowledge of* $\|\boldsymbol{z}_0-\boldsymbol{z}^*\|^2$. *In the face of stochasticity, we choose* $C = 1$ *when the initial distance to the optimal point is known. Alternatively, when only an over-estimate* Γ_0 *of* $\sqrt{\mathbb{E}\|\boldsymbol{z}_0-\boldsymbol{z}^*\|^2}$ *is available, we can set (large enough)* $C = \frac{\Gamma_0}{\sqrt{\mathbb{E}\|\boldsymbol{z}_0-\boldsymbol{z}^*\|^2}} \ge 1$ *to obtain*

$$\mathbb{E}\|\boldsymbol{z}^{ag}_{T-\frac{1}{2}}-\boldsymbol{z}^*\|^2 \le \frac{2}{\mu(T+1)}\left(\frac{2L}{rT}+2\sqrt{\frac{1+\beta}{r}}M\right)\Gamma_0^2+\frac{4\sigma}{\mu\sqrt{T}}\Gamma_0. \tag{16}$$

Remark 2.7 *When the constants are not a concern, the coarse-grained choices of* $r = \frac{1}{2}$ *and* $\beta = 1$ *would suffice. Nevertheless, to optimize the constants, the tradeoff between the deviation of* r *from* 1 *and* β *from* 0 *is crucial, as it determines a balance between the stochastic gradient noise variance and the convergence rate coefficients.*

To prepare for our multi-epoch result with the help of scheduled restarting, we perform an induction based on (16) as follows. Supposing that $\mathbb{E}\|\boldsymbol{z}_0^{[s-1]}-\boldsymbol{z}^*\|^2 \le \Gamma_0^2e^{1-s}$ hold, by taking $r = \frac{1}{2}$ and $\beta = 1$, we have

$$\mathbb{E}\|\boldsymbol{z}_0^{[s]}-\boldsymbol{z}^*\|^2 \lesssim \frac{L}{\mu T_s^2}\Gamma_0^2e^{1-s}+\frac{M}{\mu T_s}\Gamma_0^2e^{1-s}+\frac{\sigma}{\mu\sqrt{T_s}}\Gamma_0e^{\frac{1-s}{2}}.$$

Setting the right-hand side of the above inequality to satisfy $\le \Gamma_0^2e^{-s}$, and solving for T_s, we need the epoch length satisfies $T_s \asymp \sqrt{\frac{L}{\mu}}+\frac{M}{\mu}+\frac{\sigma^2}{\mu^2\Gamma_0^2e^{1-s}}$. Thus, we can obtain the total iteration complexity as

$$\sum_{s=1}^{S}\left[\sqrt{\frac{L}{\mu}}+\frac{M}{\mu}+\frac{\sigma^2}{\mu^2\Gamma_0^2e^{1-s}}\right] = \left(\sqrt{\frac{L}{\mu}}+\frac{M}{\mu}\right)S+\frac{\sigma^2}{\mu^2\Gamma_0^2}\cdot\frac{e^S-1}{e-1},$$

where $S \equiv \left\lceil \log\frac{\Gamma_0^2}{\varepsilon^2}\right\rceil$. This yields the following multi-epoch iteration complexity bound:

Corollary 2.8 (Iteration complexity of stochastic AG-EG with scheduled restarting) *Under the same condition of Theorem 2.5, the stochastic AG-EG with scheduled restarting in Algorithm 1 with epoch length* $T_s \asymp \sqrt{\frac{L}{\mu}}+\frac{M}{\mu}+\frac{\sigma^2}{\mu^2\Gamma_0^2e^{1-s}}$ *has a total iteration complexity of*

$$\mathcal{O}\left(\left(\sqrt{\frac{L}{\mu}}+\frac{M}{\mu}\right)\log\left(\frac{1}{\varepsilon}\right)+\frac{\sigma^2}{\mu^2\varepsilon^2}\right). \tag{17}$$

Note that the hard instance constructed by Zhang et al. [2022] can be modified in a straightforward way to establish a lower bound of $\Omega\left(\left(\sqrt{\frac{L}{\mu}}+\frac{M}{\mu}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$ for our monotone VI (1), demonstrating the optimality of Corollary 2.8 in the deterministic separable setting. An alternative optimality argument proceeds as follows: the first term $\sqrt{\frac{L}{\mu}}$ matches the lower bound for the minimization of a strongly convex function $\mathcal{F}$ [Nesterov, 2004], and the second term $\frac{M}{\mu}$ matches the lower bound for VI for non-strongly monotone operator when $\nabla\mathcal{F} = 0$ [Ouyang and Xu, 2021]. This together gives a lower bound for solving monotone VI (1) via a similar argument by Thekumparampil et al. [2022].

It is worth noting that while both complexity bounds in Corollary 2.8 and Theorem 2.3 match the lower bound in Zhang et al. [2022] for strongly monotone VIs with separable structure, the direct approach in §2.3 reduces to the *last-iterate* independent-sample stochastic extragradient (SEG) algorithm in *bilinear games*. Consequently, the deterministic part ($\sigma = 0$) fails to match the lower bound in Ibrahim et al. [2020]. In the stochastic case with noise variance bounded away from zero, the direct approach in §2.3 can exhibit *nonconvergence behavior* [Hsieh et al., 2020, §3]. The AG-EG algorithm in §2.4 resolves this issue by restarting the *average-iterate* SEG, matching the lower bound results (see §3.2 for more details). In addition, the complexity bound in (17) also eliminates the $\log$ prefactor of the statistical error term $\frac{\sigma^2}{\mu^2\varepsilon^2}$ compared to Theorem 2.3. The optimality of our algorithm lies in not only the optimization complexity but also the statistical error rate $\frac{\sigma^2}{\mu^2\epsilon^2}$. Here the ε-optimal point $\boldsymbol{z}$ is defined as $\|\boldsymbol{z} - \boldsymbol{z}^*\| \leq \varepsilon$.[4]

2.5 Extension of AG-EG to Proximal Algorithms

In previous subsections, we have focused on the case where the feasible set $\mathcal{Z}$ represents the entire space and the nondifferentiable convex function J is dropped. We now extend the AG-EG algorithm and its analysis to the more general setting that has a bounded feasible set (via Euclidean projection onto the feasible set) as well as a nondifferentiable convex regularization term (via a proximal operator). These settings are useful in various applications, such as the variational inequality on the Lorentz cone where projection onto $\mathcal{Z} = \left\{(\boldsymbol{x}, t) \in \mathbb{R}^{(n+1)} : \|\boldsymbol{x}\| \leq t\right\}$ is required [Chen et al., 2017], and the two-player game that involves projection onto the probability simplex, among others. To deal with bounded feasible set $\mathcal{Z}$, we adopt a variant of the EG algorithm, where we project the extrapolated point and the main iterates back onto the feasible set $\mathcal{Z}$ of $\mathcal{W}$:

$$\begin{aligned}
\boldsymbol{z}_{t-\frac{1}{2}} &= P_{\mathcal{Z}}\left[\boldsymbol{z}_{t-1} - \eta\mathcal{W}(\boldsymbol{z}_{t-1})\right] = \arg\min_{\boldsymbol{z}\in\mathcal{Z}} \langle \boldsymbol{z} - \boldsymbol{z}_{t-1}, \eta\mathcal{W}(\boldsymbol{z}_{t-1})\rangle + \frac{1}{2}\left\|\boldsymbol{z} - \boldsymbol{z}_{t-1}\right\|^2, \\
\boldsymbol{z}_t &= P_{\mathcal{Z}}\left[\boldsymbol{z}_{t-1} - \eta\mathcal{W}(\boldsymbol{z}_{t-\frac{1}{2}})\right] = \arg\min_{\boldsymbol{z}\in\mathcal{Z}} \left\langle \boldsymbol{z} - \boldsymbol{z}_{t-1}, \eta\mathcal{W}(\boldsymbol{z}_{t-\frac{1}{2}})\right\rangle + \frac{1}{2}\left\|\boldsymbol{z} - \boldsymbol{z}_{t-1}\right\|^2, \quad (18)
\end{aligned}$$

where $P_{\mathcal{Z}}(\boldsymbol{z}) = \arg\min_{\boldsymbol{z}'\in\mathcal{Z}} \|\boldsymbol{z} - \boldsymbol{z}'\|^2$ is the Euclidean projection operator. To handle the nondifferentiable simple convex function J, we replace the projection operator in (18) by the following proximal mapping defined via a Bregman divergence $\mathcal{B}(\cdot,\cdot)$:

$$\text{prox}^J_{\boldsymbol{z}}(\mathbf{v}) \equiv \operatorname*{argmin}_{\mathbf{u}\in\mathcal{Z}} \langle \mathbf{v}, \mathbf{u} - \boldsymbol{z}\rangle + \mathcal{B}(\boldsymbol{z}, \mathbf{u}) + J(\mathbf{u}). \quad (19)$$

In fact, (18) can be seen as a special case of (19) when choosing the Bregman divergence $\mathcal{B}(\boldsymbol{z}, \mathbf{u}) = \frac{1}{2}\|\boldsymbol{z} - \mathbf{u}\|^2$ and $J(\mathbf{u})$ as the set indicator function of the feasible set $\mathcal{Z}$. Therefore, by substituting the prox-mapping (19) into the AG-EG updates introduced in §2.4, we obtain the more general proximal AG-EG algorithm in Algorithm 3 (See in §C.2), which reduces to Algorithm 1 when $J = 0$, $\mathcal{B}(\boldsymbol{z}, \mathbf{u}) = \frac{1}{2}\|\boldsymbol{z} - \mathbf{u}\|^2$ and $\mathcal{Z} = \mathbb{R}^n$. Moreover, we assume that $\mathcal{B}(\cdot,\cdot)$ is $\mu_{\mathcal{B}}$-strongly convex. Without loss of generality, in contrast to the previous assumption of μ-strong convexity for $\mathcal{F}$, we instead assume that $\mathcal{F}$ is μ-strongly convex with respect to the Bregman divergence $\mathcal{B}(\cdot,\cdot)$ (See, for example, Hazan and Kale [2014], Xu et al. [2018]). Similar to Corollary 2.8, we have the following iteration complexity result, whose proof is deferred to §D.5:

Corollary 2.9 (Iteration complexity of stochastic proximal AG-EG with scheduled restarting) *Under the same condition of Theorem 2.5, the stochastic proximal AG-EG with scheduled restarting in Algorithm 3, with epoch length* $T_s \asymp \sqrt{\frac{L}{\mu\mu_{\mathcal{B}}}} + \frac{M}{\mu\mu_{\mathcal{B}}} + \frac{\sigma^2\mathcal{B}(\boldsymbol{z}_0,\boldsymbol{z}^*)}{\mu^2\mu_{\mathcal{B}}\Gamma_0^2 e^{1-s}}$*, has a total iteration complexity of*

$$\mathcal{O}\left(\left(\sqrt{\frac{L}{\mu\mu_{\mathcal{B}}}} + \frac{M}{\mu\mu_{\mathcal{B}}}\right)\log\left(\frac{1}{\varepsilon}\right) + \frac{\sigma^2\mathcal{B}(\boldsymbol{z}_0,\boldsymbol{z}^*)}{\mu^2\mu_{\mathcal{B}}\varepsilon^2}\right).$$

For the deterministic case, proximal AG-EG with scheduled restarting has a total iteration complexity of $\mathcal{O}\left(\left(\sqrt{\frac{L}{\mu\mu_{\mathcal{B}}}} + \frac{M}{\mu\mu_{\mathcal{B}}}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$ to output an ε-optimal point of (1).

[4] The optimal statistical error rate $\frac{\sigma^2}{\mu^2 T}$ has been achieved by a multistage algorithm in Fallah et al. [2020], where the ε-optimal point is defined by $\left\|\boldsymbol{z} - \boldsymbol{z}^*\right\|^2 \leq \varepsilon$. In our paper, the ε-optimal point is defined by $\left\|\boldsymbol{z} - \boldsymbol{z}^*\right\| \leq \varepsilon$. Therefore, our statistical error rate can be translated into $\frac{\sigma^2}{\mu^2 T}$ using their definition, which matches their result.

3 Implications for Specific Instances

In this section, we discuss the implications of our AG-EG algorithm and its convergence rates when applying to two instances of saddle-point problems.

3.1 Strongly-Convex-Strongly-Concave Saddle-Point Problem

For the stochastic bilinearly-coupled SC-SC saddle-point problem (3), we note that the smoothness and strong convexity parameters L_F, L_G, μ_F, and μ_G of F and G may differ. To accommodate these variations in curvature information, we employ a scaling reduction technique. This technique enables us to convert the SC-SC with equal strong convexity parameters for F and G by reparametrizing the objective function. The same argument is also applicable to the direct approach.

In lieu of (3), we consider

$$\min_{\hat{\boldsymbol{x}}} \max_{\hat{\boldsymbol{y}}} \hat{\mathscr{F}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = F(\hat{\boldsymbol{x}}) + \hat{H}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) - \hat{G}(\hat{\boldsymbol{y}}),$$

where $\hat{\mathscr{F}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \mathscr{F}(\boldsymbol{x}, \boldsymbol{y})$ with the symbolic reparametrization $\hat{\boldsymbol{x}} = \boldsymbol{x}$, $\hat{\boldsymbol{y}} = \sqrt{\frac{\mu_G}{\mu_F}}\boldsymbol{y}$, $\hat{H}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = H(\boldsymbol{x}, \boldsymbol{y})$, $\hat{G}(\hat{\boldsymbol{y}}) = G(\boldsymbol{y})$ and also their derivatives $\nabla_{\hat{\boldsymbol{y}}} \hat{H}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \sqrt{\frac{\mu_F}{\mu_G}} \nabla_{\boldsymbol{y}} H(\boldsymbol{x}, \boldsymbol{y}), \nabla \hat{G}(\hat{\boldsymbol{y}}) = \sqrt{\frac{\mu_F}{\mu_G}} \nabla G(\boldsymbol{y})$ (the stochastic oracles $\hat{h}, \hat{g}$ follow the same rule). It is straightforward to verify that $\hat{\mathscr{F}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ is μ-strongly-convex-μ-strongly-concave. The essence of our update rules can be summarized by the rescaled updates on $\boldsymbol{y}$:

$$\begin{aligned}
&\hat{\boldsymbol{y}}_t = \hat{\boldsymbol{y}}_{t-1} - \eta_t \left(-\nabla_{\hat{\boldsymbol{y}}} h(\hat{\boldsymbol{x}}_{t-\frac{1}{2}}, \hat{\boldsymbol{y}}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\hat{\boldsymbol{y}}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) \right) \\
&\Leftrightarrow \ \boldsymbol{y}_t = \boldsymbol{y}_{t-1} - \eta_t \cdot \tfrac{\mu_F}{\mu_G} \left(-\nabla_{\boldsymbol{y}} h(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\hat{\boldsymbol{y}}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) \right).
\end{aligned}$$

Therefore, it suffices to analyze Algorithm 3 for $\hat{\mathscr{F}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ and due to this scaling reduction, we only need to prove all results for the case of $\mu_F = \mu_G = \mu$. It is also straightforward to justify corresponding scaling changes as: $L = L_F \vee \frac{\mu_F}{\mu_G} L_G$, $M = \sqrt{\frac{\mu_F}{\mu_G}} \lambda_{\max}(\mathbf{B}^\top \mathbf{B})$, and $\mu = \mu_F$. The following corollary is recovered by reverting the scaling reduction from $\hat{\mathscr{F}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ to $\mathscr{F}(\boldsymbol{x}, \boldsymbol{y})$.

Corollary 3.1 (Iteration complexity of stochastic AG-EG on SC-SC saddle-point problem) *For solving* (3)*, Algorithm 1 with an epoch length* $T_s \asymp \sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}} + \frac{\sigma^2}{\mu_F^2 \Gamma_0^2 e^{1-s}}$ *has a total iteration complexity of*

$$\mathcal{O}\left(\left(\sqrt{\tfrac{L_F}{\mu_F} \vee \tfrac{L_G}{\mu_G}} + \sqrt{\tfrac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}} \right) \log\left(\tfrac{1}{\varepsilon}\right) + \tfrac{\sigma^2}{\mu_F^2 \varepsilon^2} \right).$$

In the deterministic case, the iteration complexity in Theorem 2.8 matches the lower bound established by Zhang et al. [2022], i.e., $\Omega\left(\left(\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}} \right) \log\left(\frac{1}{\varepsilon}\right) \right)$. Moreover, our algorithm achieves the optimal statistical rate of $\frac{\sigma^2}{\mu_F^2 \varepsilon^2}$ up to a constant prefactor.

Remark 3.2 *A well-known finding regarding the second scheme of Nesterov acceleration is its connection to the primal-dual method [Lan and Zhou, 2018, Lin et al., 2020c]. This finding has been incorporated into the design of the LPD algorithm [Thekumparampil et al., 2022], where a Chambolle-Pock-style primal-dual method is utilized as an approximation of proximal point methods, instead of the extragradient used in this paper. The LPD algorithm [Thekumparampil et al., 2022] also achieves the optimal complexity for the deterministic bilinearly-coupled saddle-point problem.*

3.2 Bilinear Games

In this subsection, we consider the particular case of bilinear games. *We assume* $n = m$ *such that* $\mathbf{B}$ *is a nonsingular square matrix,* $\nabla f(\boldsymbol{x}; \xi) = \mathbf{0}$ and $\nabla g(\boldsymbol{y}; \xi) = \mathbf{0}$ a.s., so (3) reduces to

$$\min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \mathscr{F}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_\zeta \left[h(\boldsymbol{x}, \boldsymbol{y}; \zeta) \right] = H(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top \mathbf{B} \boldsymbol{y} - \boldsymbol{x}^\top \mathbf{u}_{\boldsymbol{x}} + \mathbf{u}_{\boldsymbol{y}}^\top \boldsymbol{y}, \tag{20}$$

and Algorithm 3 reduces to the independent-sample extragradient descent-ascent algorithm for (20). The saddle point $[\boldsymbol{z}^*; \boldsymbol{\omega}_{\boldsymbol{y}}^\star]$ in this case is the unique solution to the linear equation

$$\begin{bmatrix} \mathbf{0} & \mathbf{B} \\ -\mathbf{B}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}^* \\ \boldsymbol{\omega}_{\boldsymbol{y}}^\star \end{bmatrix} = \begin{bmatrix} \mathbf{u}_{\boldsymbol{x}} \\ \mathbf{u}_{\boldsymbol{y}} \end{bmatrix}, \quad \text{which has a closed-form solution } \begin{bmatrix} \boldsymbol{z}^* \\ \boldsymbol{\omega}_{\boldsymbol{y}}^\star \end{bmatrix} = \begin{bmatrix} -(\mathbf{B}^\top)^{-1}\mathbf{u}_{\boldsymbol{y}} \\ \mathbf{B}^{-1}\mathbf{u}_{\boldsymbol{x}} \end{bmatrix}.$$

Our results imply the following iteration complexity for solving stochastic bilinear games.

Corollary 3.3 (Iteration complexity of stochastic AG-EG, bilinear games) *For solving (20), choose the step sizes $\alpha_t = \frac{2}{t+1}$ and $\eta_t \equiv \frac{1}{\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$, in which case Algorithm 1 with an epoch length $T_s \asymp \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}$ has the total iteration complexity of*

$$\mathcal{O}\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}\log\left(\frac{\sqrt[4]{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}{\sigma_{\mathrm{Bil}}}\right) + \frac{\sigma_{\mathrm{Bil}}^2}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\varepsilon^2}\right). \tag{21}$$

Note that our choice of the step size is maximal and is independent of the noise. In the deterministic setting, letting $\sigma_{\mathrm{Bil}} \asymp \varepsilon\sqrt[4]{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}$, the complexity bound in Corollary 3.3 reduces to $\mathcal{O}\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}\log\left(\frac{1}{\varepsilon}\right)\right)$, which matches the lower bound in Ibrahim et al. [2020]. Notably, Azizian et al. [2020b] proposed an algorithm achieving an upper bound that matches the lower bound in Ibrahim et al. [2020].[5] Li et al. [2022a] also proposed a lower-bound matching SEG algorithm that uses a shared sample in both steps under an unbounded noise assumption. In contrast, our algorithm is in the independent-sample setting with bounded noise variance.

Remark 3.4 *Standard acceleration techniques do not attain the optimal nonasymptotic convergence rate for bilinear games [Gidel et al., 2019b]. This limitation applies to various algorithms, including the direct approach [§2.3], as well as several other acceleration techniques [Thekumparampil et al., 2022, Kovalev et al., 2022, Jin et al., 2022], all of which fall short of achieving optimal acceleration for bilinear games. Therefore, matching both lower bounds in a single algorithm in the general stochastic setting has been an open problem. While Li et al. [2022b] present an algorithm that achieves both lower bounds in a single algorithm, it relies on the use of optimistic gradients rather than extragradients on the bilinear coupling function. Furthermore, our algorithm and analysis is more general than those in Li et al. [2022b] as we can handle the general variational inequality with proximal operators.*

4 Conclusions

We have presented a stochastic extragradient-based acceleration algorithm, AG-EG, for solving stochastic monotone variational inequalities with separable structure. The iteration complexity of our algorithm matches the lower bound and is independent of the size of the feasible set. When specialized to solving the bilinearly coupled saddle-point problem (3), our AG-EG algorithm simultaneously matches lower bounds due to Zhang et al. [2022] and Ibrahim et al. [2020] for strongly-convex-strongly-concave and bilinear games, respectively. To the best of our knowledge, this is the first time that all three lower bounds have been met by a single algorithm. There are some remaining issues to be addressed, however, including the case of one-sided nonstrong convexity, the setting of unbounded noise variance, and the characterization of the full parameter regime dependency on $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$. These are left as important directions for future research.

Acknowledgments and Disclosure of Funding

This work is supported in part by Canada CIFAR AI Chair to GG, by the Mathematical Data Science program of the Office of Naval Research under grant number N00014-18-1-2764 and also the Vannevar Bush Faculty Fellowship program under grant number N00014-21-1-2941 and National Science Foundation (NSF) grant IIS-1901252 to MIJ. This work is also supported in part by the NSF CAREER Award 1906169 to QG and by NSF Award's IIS-2110170 and DMS-2134106 to SSD.

[5] Azizian et al. [2020b], while achieving optimality in bilinear games, has a flavor of "aggressive extrapolation" [Hsieh et al., 2020] that results in a convergence behavior resembling accelerated Hamiltonian gradient descent. Thus, it cannot be tuned to match the complexity lower bound for the bilinearly coupled SC-SC case.

References

Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games. In *International Conference on Artificial Intelligence and Statistics*, pages 2863–2873. PMLR, 2020a.

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, 2020b.

Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization*, 24(4):1779–1814, 2014.

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

Ziyi Chen, Qunwei Li, and Yi Zhou. Finding local minimax points via (stochastic) cubic-regularized gda: Global convergence and complexity. *arXiv preprint arXiv:2110.07098*, 2021.

Michael B Cohen, Aaron Sidford, and Kevin Tian. Relative Lipschitzness in extragradient methods and a direct recipe for acceleration. In *12th Innovations in Theoretical Computer Science Conference (ITCS 2021)*, volume 185, page 62, 2021.

Christoph Dann, Gerhard Neumann, Jan Peters, et al. Policy evaluation with temporal differences: A survey and comparison. *Journal of Machine Learning Research*, 15:809–883, 2014.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with optimism. In *International Conference on Learning Representations*, 2018.

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.

Francisco Facchinei and Jong-Shi Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*. Springer, 2003.

Alireza Fallah, Asuman Ozdaglar, and Sarath Pattathil. An optimal multistage stochastic gradient method for minimax problems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 3573–3579. IEEE, 2020.

Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019a.

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, 2019b.

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.

Zhishuai Guo, Zhuoning Yuan, Yan Yan, and Tianbao Yang. Fast objective & duality gap convergence for nonconvex-strongly-concave min-max problems. *arXiv preprint arXiv:2006.06889*, 2020.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research*, 15(1): 2489–2512, 2014.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. In *Advances in Neural Information Processing Systems*, volume 33, pages 16223–16234, 2020.

Adam Ibrahim, Waıss Azizian, Gauthier Gidel, and Ioannis Mitliagkas. Linear lower bounds and conditioning of differentiable games. In *International Conference on Machine Learning*, pages 4583–4593. PMLR, 2020.

Alfredo N Iusem, Alejandro Jofré, Roberto Imbuzeiro Oliveira, and Philip Thompson. Extragradient method with variance reduction for stochastic variational inequalities. *SIAM Journal on Optimization*, 27(2):686–724, 2017.

Samy Jelassi, Carles Domingo-Enrich, Damien Scieur, Arthur Mensch, and Joan Bruna. Extragradient with player sampling for faster convergence in n-player games. In *International Conference on Machine Learning*, pages 4736–4745. PMLR, 2020.

Yujia Jin, Aaron Sidford, and Kevin Tian. Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. In *Conference on Learning Theory*, pages 4362–4415. PMLR, 2022.

Michael I Jordan, Tianyi Lin, and Manolis Zampetakis. First-order algorithms for nonlinear generalized Nash equilibrium problems. *Journal of Machine Learning Research*, 24(38):1–46, 2023.

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

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

Georgios Kotsalis, Guanghui Lan, and Tianjiao Li. Simple and optimal methods for stochastic variational inequalities, I: Operator extrapolation. *arXiv preprint arXiv:2011.02987*, 2020.

Dmitry Kovalev, Alexander Gasnikov, and Peter Richtárik. Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. *Advances in Neural Information Processing Systems*, 35:21725–21737, 2022.

Guanghui Lan and Yuyuan Ouyang. Mirror-prox sliding methods for solving a class of monotone variational inequalities. *arXiv preprint arXiv:2111.00996*, 2021.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171:167–215, 2018.

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, 2022a.

Chris Junchi Li, Angela Yuan, Gauthier Gidel, and Michael I Jordan. Nesterov meets optimism: Rate-optimal optimistic-gradient-based method for stochastic bilinearly-coupled minimax optimization. *arXiv preprint arXiv:2210.17550*, 2022b.

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

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020a.

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

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

Nicolas Loizou, Hugo Berard, Alexia Jolicoeur-Martineau, Pascal Vincent, Simon Lacoste-Julien, and Ioannis Mitliagkas. Stochastic hamiltonian gradient methods for smooth games. In *International Conference on Machine Learning*, pages 6370–6381. PMLR, 2020.

Luo Luo, Yujun Li, and Cheng Chen. Finding second-order stationary points in nonconvex-strongly-concave minimax optimization. *Advances in Neural Information Processing Systems*, 35:36667–36679, 2022.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*, 2018.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of GANs. *Advances in Neural Information Processing Systems*, 30, 2017.

Dmitriy Metelev, Alexander Rogozin, Alexander Gasnikov, and Dmitry Kovalev. Decentralized saddle-point problems with different constants of strong convexity and strong concavity. *arXiv preprint arXiv:2206.00090*, 2022.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

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.

Vaishnavh Nagarajan and J Zico Kolter. Gradient descent GAN optimization is locally stable. *Advances in Neural Information Processing Systems*, 30, 2017.

Arkadi Nemirovski. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2004.

Yurii Nesterov and Laura Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Discrete & Continuous Dynamical Systems*, 31(4):1383, 2011.

Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, 185(1-2):1–35, 2021.

Brendan O'donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pages 1–35, 2021.

James Renegar and Benjamin Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of Computational Mathematics*, 22(1):211–256, 2022.

Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. *Advances in Neural Information Processing Systems*, 30, 2017.

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.

Othmane Sebbouh, Marco Cuturi, and Gabriel Peyré. Randomized stochastic gradient descent ascent. In *International Conference on Artificial Intelligence and Statistics*, pages 2941–2969. PMLR, 2022.

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.

Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

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

Jialei Wang and Lin Xiao. Exploiting strong convexity from data with primal-dual first-order algorithms. In *International Conference on Machine Learning*, pages 3694–3702, 2017.

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

Lin Xiao, Adams Wei Yu, Qihang Lin, and Weizhu Chen. Dscovr: Randomized primal-dual block coordinate algorithms for asynchronous distributed optimization. *Journal of Machine Learning Research*, 20(1):1634–1691, 2019.

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

Pan Xu, Tianhao Wang, and Quanquan Gu. Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions. In *International Conference on Machine Learning*, pages 5492–5501. PMLR, 2018.

Yan Yan, Yi Xu, Qihang Lin, Lijun Zhang, and Tianbao Yang. Stochastic primal-dual algorithms with faster convergence than $o(1/\sqrt{T})$ for problems without bilinear structure. *arXiv preprint arXiv:1904.10112*, 2019.

Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33:5789–5800, 2020.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR, 2022.

Junyu Zhang, Mingyi Hong, and Shuzhong Zhang. On lower iteration complexity bounds for the convex concave saddle point problems. *Mathematical Programming*, 194(1-2):901–935, 2022.

Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, 18:1–42, 2017.

The supplementary material is organized as follows. Section A provides specific examples in our minimax optimization setting. Section B compares our work with prior related works. Section C discusses the stochastic AG-EG algorithms in detail. Section D proves the main results. Finally, Section E provides proofs of auxiliary lemmas that support the proofs of main results.

A Examples

We conduct an overview of some applications in this section.

Reinforcement learning. Reinforcement learning problems can be formalized as Markov Decision Processes (MDPs) where, at each step $t = 1, \ldots, n$, the learner receives a four-element tuple, $\{s_t, a_t, r_t, s_{t+1}\}$, where (s_t, a_t) is the current state-action pair, r_t is the reward received upon choosing a_t, and s_{t+1} is the next state drawn from a transition distribution. For example, policy evaluation with a linear function approximator can be formalized in terms of the minimization of the *mean squared projected Bellman-Error* (MSPBE) [Dann et al., 2014] based on a set of tuples:

$$\min_{\boldsymbol{\theta}} \ \frac{1}{2} \left\| \mathbf{A}\boldsymbol{\theta} - \mathbf{b} \right\|_{\mathbf{C}^{-1}}^2 + \frac{\rho}{2} \left\| \boldsymbol{\theta} \right\|^2, \tag{22}$$

where $\mathbf{A} = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\phi}(s_t)(\boldsymbol{\phi}(s_t) - \gamma \boldsymbol{\phi}(s_{t+1}))^\top$, $\mathbf{b} = \frac{1}{n} \sum_{t=1}^{n} r_t \boldsymbol{\phi}(s_t)$, and $\mathbf{C} = \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{\phi}(s_t) \boldsymbol{\phi}(s_t)^\top$ for a given feature mapping ϕ. To reduce the computational cost incurred by calculating the inverse of matrix $\mathbf{C}$, Du et al. [2017] propose an alternative minimax form of (22):

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{w}} \ \frac{\rho}{2} \|\boldsymbol{\theta}\|^2 - \mathbf{w}^\top \mathbf{A} \boldsymbol{\theta} - \frac{1}{2} \|\mathbf{w}\|_{\mathbf{C}}^2 + \mathbf{w}^\top \mathbf{b},$$

which falls under the umbrella of problem (3) whenever $\mathbf{C}$ is positive definite.

Quadratic games. Another class of examples arises in the setting of bilinear games, where the minimax objective is:

$$\mathscr{F}(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2} \boldsymbol{x}^\top \mathbf{M}_F \boldsymbol{x} + \boldsymbol{x}^\top \mathbf{B} \boldsymbol{y} - \frac{1}{2} \boldsymbol{y}^\top \mathbf{M}_G \boldsymbol{y} - \boldsymbol{x}^\top \mathbf{v}_{\boldsymbol{x}} + \mathbf{v}_{\boldsymbol{y}}^\top \boldsymbol{y}, \tag{23}$$

where $\mathbf{M}_F, \mathbf{M}_G$ are real-valued matrices of dimensions $n \times n$ and $m \times m$. This has the form (3) with $F(\boldsymbol{x}) = \frac{1}{2} \boldsymbol{x}^\top \mathbf{M}_F \boldsymbol{x} - \boldsymbol{x}^\top \mathbf{v}_{\boldsymbol{x}}$, $G(\boldsymbol{y}) = \frac{1}{2} \boldsymbol{y}^\top \mathbf{M}_G \boldsymbol{y} - \mathbf{v}_{\boldsymbol{y}}^\top \boldsymbol{y}$ and $H(\boldsymbol{x}, \boldsymbol{y}) \equiv \boldsymbol{x}^\top \mathbf{B} \boldsymbol{y}$. A particular case we will be considering in §3.2 is the case of bilinear games; i.e., where there are no quadratic terms. We provide a detailed analysis of the nonasymptotic convergence in this setting in §3.2 and show that the upper bound on the convergence rate given by our algorithm matches the lower bound of Ibrahim et al. [2020, Theorem 3].

Regularized empirical risk minimization. The problem of the minimization of the regularized empirical risk for convex losses and linear predictors is a core problem in classical supervised learning:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \ \mathcal{L}(\mathbf{A}\boldsymbol{x}) + F(\boldsymbol{x}) \equiv \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\mathbf{a}_i^\top \boldsymbol{x}) + F(\boldsymbol{x}),$$

where $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]^\top \in \mathbb{R}^{n \times d}$ consists of feature vectors $\{\mathbf{a}_i\}$, $\mathcal{L}_i(\boldsymbol{y})$ is a univariate convex loss for the ith data point, and $F(\boldsymbol{x})$ is a convex regularizer. A standard construction turns this empirical risk minimization problem into a saddle-point problem as follows:

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \max_{\boldsymbol{y} \in \mathbb{R}^m} \ F(\boldsymbol{x}) + \boldsymbol{x}^\top \mathbf{A} \boldsymbol{y} - \underbrace{\mathcal{L}^\star(\boldsymbol{y})}_{\text{Legendre dual function of } \mathcal{L}(\boldsymbol{y})} \equiv F(\boldsymbol{x}) + \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{y}^\top \mathbf{a}_i - \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}^\star(\boldsymbol{y}_i).$$

See Zhang and Xiao [2017], Wang and Xiao [2017], Xiao et al. [2019] for in-depth discussions of solving this problem under such a dual form of representation.

B Related Work

Here we compare our results with related work on saddle-point (minimax) optimization in the machine learning and optimization literature. In Table 1, we compare our AG-EG algorithm with previous work on solving saddle-point optimization problems, in terms of gradient complexity.

Bilinear games. In the bilinear game setting, where $L_F = \mu_F = L_G = \mu_G = 0$, a lower bound has been established by Ibrahim et al. [2020]: $\Omega\Big(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}\log\left(\frac{1}{\varepsilon}\right)\Big)$. The study of bilinear game has been initiated by Daskalakis et al. [2018] for understanding saddle-point optimization. They proposed the optimistic gradient descent-ascent (OGDA) algorithm and achieved a sublinear convergence rate. Subsequently, the classical methods of ExtraGradient (EG) and Optimistic Gradient Descent Ascent (OGDA) algorithms were proven to have a linear convergence rate for strongly monotone and Lipschitz operator with $\mathcal{O}\Big(\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\log(\frac{1}{\varepsilon})\Big)$ iteration complexity [Gidel et al., 2019b, Mokhtari et al., 2020]. Azizian et al. [2020a] proved that by considering first-order methods with a fixed number of composed gradient evaluations and the last iterate as output (this class of methods is called 1-SCLI and excludes momentum and restarting), the $\mathcal{O}\Big(\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\log(\frac{1}{\varepsilon})\Big)$ iteration complexity for EG is optimal. In the absence of strong monotonicity assumption, Loizou et al. [2020] provided the first set of nonasymptotic last-iterate convergence guarantees for smooth games over a noncompact domain from a Hamiltonian viewpoint. The proposed stochastic Hamiltonian gradient method attains convergence in the finite-sum bilinear game setting as well. In a very recent work, Kovalev et al. [2022] derived an $\mathcal{O}\Big(\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\log(\frac{1}{\varepsilon})\Big)$ iteration complexity for convex-concave saddle-point problems with bilinear coupling. This is comparable to the rates in Daskalakis et al. [2018], Liang and Stokes [2019], Gidel et al. [2019b], Mokhtari et al. [2020], Mishchenko et al. [2020]. To match the $\Omega\Big(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}\log(\frac{1}{\varepsilon})\Big)$ lower bound provided by Ibrahim et al. [2020], Azizian et al. [2020b] considered EG with momentum. They used a perturbed spectral analysis encompassing Polyak momentum. Nonetheless, Azizian et al. [2020b] only provided accelerated rates in the regime where the condition number is large. Li et al. [2022a] is the first to show that a variant of stochastic extragradient method converges at an accelerated rate for bilinear games with unbounded domain and unbounded stochastic noise using restarted iterate averaging, and matches the lower bound [Ibrahim et al., 2020] in the deterministic setting.

Smooth and strongly-convex-strongly-concave saddle point problems. Lower bound has been recently studied by Ouyang and Xu [2021] for smooth convex-concave minimax optimization, and by Zhang et al. [2022] for strongly-convex-strongly-concave saddle-point problems. The latter is of order $\Omega\Big(\Big(\sqrt{\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G}}+\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}\Big)\log\left(\frac{1}{\varepsilon}\right)\Big)$. As for upper bounds, earlier extragradient-based methods [Tseng, 1995] and accelerated dual extrapolation algorithm [Nesterov and Scrimali, 2011] achieve, when restricted to the bilinearly coupled problem, an iteration complexity of $\tilde{\mathcal{O}}\Big(\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G}+\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}\Big)$. The same iteration complexity has also been achieved by Gidel et al. [2019a], Mokhtari et al. [2020], Cohen et al. [2021] from a relative Lipschitz viewpoint.[6] Improving upon this result, Lin et al. [2020b] achieved a complexity of $\tilde{\mathcal{O}}\Big(\sqrt{\frac{L_FL_G}{\mu_F\mu_G}}+\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}\Big)$ using proper acceleration methods. Wang and Li [2020] achieved[7] $\tilde{\mathcal{O}}\Big(\sqrt{\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G}}+\sqrt[4]{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}\cdot\frac{L_FL_G}{\mu_F\mu_G}}+\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}\Big)$ iteration complexity and a Hermitian-skew-based analysis nearly matches Zhang et al. [2022] for the quadratic minimax game. For the same problem, Xie et al. [2021] achieved a complexity of $\tilde{\mathcal{O}}\Big(\sqrt[4]{\frac{L_FL_G}{\mu_F\mu_G}\big(\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G}\big)}+\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}\Big)$. These works improve upon Lin et al. [2020b] in a fine-grained fashion where separate Lipschitz constants on different parts of the objective are allowed. In early 2022, three concurrent works Kovalev et al. [2022], Thekumparampil et al. [2022], Jin et al. [2022] study the deterministic problem and independently match the lower bound by Zhang et al. [2022]. The main novelty of this work is that both lower bounds Ibrahim et al. [2020] and Zhang et al. [2022] are achieved in a single algorithm, plus an optimal statistical error rate up

[6] Mokhtari et al. [2020] report an $\tilde{\mathcal{O}}\Big(\frac{L_F\vee L_G+\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}{\mu_F\wedge\mu_G}\Big)$ complexity, but the mentioned complexity can be obtained via a scaling-reduction argument: consider $\mu_F=\mu_G$ case first, then consider the general case by rescaling the y variable by a factor of $\sqrt{\frac{\mu_G}{\mu_F}}$.

[7] Note the cross term here, $\tilde{\mathcal{O}}\Big(\sqrt[4]{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}\cdot\frac{L_FL_G}{\mu_F\mu_G}}\Big)$, cannot be absorbed into the summation of the remaining terms.

to a constant prefactor in the stochastic setting. Recently, an independent work by Li et al. [2022b] also proposed a single algorithm that can achieve the optimal rates for both settings. However, their algorithm is based on optimistic gradient, and is less general than the variational inequality setting studied in this paper.

Method \ Setting	Bilinearly-coupled SC-SC	Bilinear Game	Stochastic VI
EG / OGDA [Mokhtari et al., 2020] [Cohen et al., 2021]	$\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$	$\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$	✓
Minimax-APPA [Lin et al., 2020b]	$\sqrt{\frac{L_F L_G}{\mu_F \mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$	—	X
Proximal Best Response [Wang and Li, 2020]	$\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt[4]{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G} \cdot \frac{L_F L_G}{\mu_F \mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$	—	X
DIPPA [Xie et al., 2021]	$\sqrt[4]{\frac{L_F L_G}{\mu_F \mu_G}\left(\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}\right)} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$	—	X
LPD [Thekumparampil et al., 2022]	$\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$	$\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$	X
APDG [Kovalev et al., 2022]	$\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$	$\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$	X
PD-EG [Jin et al., 2022]	$\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$	—	X
EG+Momentum [Azizian et al., 2020b]	—	$\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}$	X
SEG with Restarting [Li et al., 2022a]	—	$\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}$	✓
AG-EG-Direct (this work)	$\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$	$\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$	✓
AG-EG with Restarting (this work)	$\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$	$\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}$	✓
Lower Bound [Zhang et al., 2022] [Ibrahim et al., 2020]	$\Omega\left(\left(\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}\right) \log\left(\frac{1}{\varepsilon}\right)\right)$	$\tilde{\Omega}\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}} \log\left(\frac{1}{\varepsilon}\right)\right)$	—

Reference	Stochastic variational inequality	No bounded domain assumption	No bounded noise assumption
[Korpelevich, 1976] [Juditsky et al., 2011] [Hsieh et al., 2020]	$\frac{L \vee M}{\varepsilon}\mathcal{D}^2 + \frac{\sigma^2}{\varepsilon^2}\mathcal{D}^2$	X	X
[Li et al., 2022a] for bilinear games	$\frac{L \vee M}{\varepsilon}\Gamma_0^2 + \frac{\sigma^2}{\varepsilon^2}\Gamma_0^2$	X	✓
[Chen et al., 2017] [Lan and Ouyang, 2021]	$\sqrt{\frac{L}{\varepsilon}}\mathcal{D} + \frac{M}{\varepsilon}\mathcal{D}^2 + \frac{\sigma^2}{\varepsilon^2}\mathcal{D}^2$	X	X
This work	$\sqrt{\frac{L}{\varepsilon}}\Gamma_0 + \frac{M}{\varepsilon}\Gamma_0^2 + \frac{\sigma^2}{\varepsilon^2}\Gamma_0^2$	✓	X
Lower Bound [Zhang et al., 2022] [Ouyang and Xu, 2021]	$\Omega\left(\sqrt{\frac{L}{\varepsilon}}\mathcal{D} + \frac{M}{\varepsilon}\mathcal{D}^2 + \frac{\sigma^2}{\varepsilon^2}\mathcal{D}^2\right)$	X	X

Table 1: A comparison of the first-order gradient complexities of our proposed algorithm with selected prevailing algorithms in terms of gradient complexity for solving a variety of saddle-point problems. Upper tabular: comparison of several cases such as general bilinearly-coupled SC-SC, bilinear games for finding an ε-optimal point, as well as a column indicating whether the stochastic variational inequality (VI) case is discussed. Lower tabular: complexities for stochastic VI for finding a point of ε primal-dual gap, as well as columns of domain/noise assumptions (note that $\Gamma_0 \leq \mathcal{D}$). The row in red background is the convergence result presented in this paper. The "—" indicates that the complexity does not apply to the given case. A polylogarithm factor in each upper bound in the table is ignored.

Stochastic minimax optimization. Stochastic minimax optimization has been studied intensively as a special case of the variational inequalities. It is widely assumed in the classical literature on stochastic variational inequality [Juditsky et al., 2011] that the set of parameters and the variance of the stochastic estimate of the vector field are bounded. Chen et al. [2017] extended the analysis of Juditsky et al. [2011] and achieved an accelerated convergence rates for a class of variational

Algorithm 2 Stochastic AcceleratedGradient-ExtraGradient (AG-EG) Descent-Ascent Algorithm, Direct Approach

Require: Initialization $\boldsymbol{z}_0$, total number of iterates T, step sizes $(\alpha_t, \eta_t : t = 1, 2, \dots)$

Set $\boldsymbol{z}^{\text{ag}}_{-\frac{1}{2}} \leftarrow \boldsymbol{z}_0$, $\boldsymbol{z}^{\text{md}}_0 \leftarrow \boldsymbol{z}_0$

for $t = 1, 2, \dots, T$ **do**

  Draw samples $\xi_{t-\frac{1}{2}} \sim \mathcal{D}_\xi$ from oracle, and also $\zeta_{t-\frac{1}{2}}, \zeta_t \sim \mathcal{D}_\zeta$ independently from oracle

  $\boldsymbol{z}_{t-\frac{1}{2}} \leftarrow \boldsymbol{z}_{t-1} - \eta_t \left(\nabla \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1}; \zeta_{t-\frac{1}{2}}) + \nabla \tilde{\mathcal{F}}(\boldsymbol{z}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) - \mu(\boldsymbol{z}^{\text{md}}_{t-1} - \boldsymbol{z}_{t-1}) \right)$

  $\boldsymbol{z}^{\text{ag}}_{t-\frac{1}{2}} \leftarrow (1 - \alpha_t)\boldsymbol{z}^{\text{ag}}_{t-\frac{3}{2}} + \alpha_t \boldsymbol{z}_{t-\frac{1}{2}}$

  $\boldsymbol{z}_t \leftarrow \boldsymbol{z}_{t-1} - \eta_t \left(\nabla \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t) + \nabla \tilde{\mathcal{F}}(\boldsymbol{z}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) - \mu(\boldsymbol{z}^{\text{md}}_{t-1} - \boldsymbol{z}_{t-\frac{1}{2}}) \right)$

  $\boldsymbol{z}^{\text{md}}_t \leftarrow (1 - \alpha_{t+1})\boldsymbol{z}^{\text{ag}}_{t-\frac{1}{2}} + \alpha_{t+1}\boldsymbol{z}_t$

end for

Output: $\boldsymbol{z}_T$

inequalities. Iusem et al. [2017] proposed an analysis of stochastic extragradient using large batches to reduce the variance. Mertikopoulos et al. [2018] showed almost sure convergence of Stochastic EG to a strictly coherent solution (a.k.a., star-strict monotone variational inequality problem). In a similar vein, Ryu et al. [2019] showed that stochastic gradient descent ascent (SGDA) with anchoring almost surely converges to strictly convex-concave saddle points. Fallah et al. [2020] developed a multistage variant of SGDA and stochastic optimistic gradient descent ascent with constant learning rate decay schedule. We improve upon their rates since their iteration complexity depends on a significantly larger condition number than our method and is infinite in the absence of strong convexity and strong concavity. They achieved the optimal dependency on the noise variance but suboptimal dependency on the condition number. Hsieh et al. [2020] developed a double step size extragradient method and proved the last-iterate convergence rates under an error bound condition similar to star-strong monotonicity. Kotsalis et al. [2020] proposed a simple and optimal scheme for a class of generalized strongly monotone (stochastic) variational inequalities. Due to the unconstrained nature of stochastic bilinear games, these two assumptions do not hold in this case because the noise increases with the value of the parameters. Mishchenko et al. [2020] showed that stochastic extragradients can be computed under a different step size, which removes the bounded domain assumption, while still requiring the bounded noise assumption. They also discussed the advantages of using the same mini-batch for the two stochastic gradients in stochastic extragradient. In another vein, Jelassi et al. [2020] focused on stochastic extragradient in games with a large number of players. They proposed an extragradient algorithm that randomly updates a small subset of the players at each iteration. Yan et al. [2019, 2020], Rafique et al. [2021] studied the nonsmooth setting and obtained fast rates. More recent works consider minimax optimization problems without convexity and/or concavity, where the goal is to find first-order and second-order stationary points [Lin et al., 2020a, Guo et al., 2020, Chen et al., 2021, Yang et al., 2022, Luo et al., 2022, Sebbouh et al., 2022]. One interesting direction is to extend our algorithm to these settings and obtain a fine-grained complexity bound with optimal rates.

C Algorithms

In this section we provide delayed algorithms for the AG-EG (direct approach) and the AG-EG with bounded domain and proximal operator.

C.1 Stochastic AG-EG, Direct Approach

The full algorithm for AG-EG, direct approach is shown in Algorithm 2.

C.2 Stochastic AG-EG, with Restarting and Projection

The full algorithm for AG-EG, with restarting and projection is shown in algorithm 3.

Algorithm 3 Stochastic AcceleratedGradient-ExtraGradient (AG-EG) Descent-Ascent Algorithm, with Scheduled Restarting

Require: Initialization $\boldsymbol{z}_0^{[0]}$, total number of epochs $\mathscr{S} \geq 1$, total number of per-epoch iterates $(T_s : s = 1, \ldots, \mathscr{S})$, step sizes $(\alpha_t, \eta_t : t = 1, 2, \ldots)$, ratio of strong-convexity params. $\mathscr{R} = \frac{\mu_G}{\mu_F}$

for $s = 1, 2, \ldots, \mathscr{S}$ **do**

  Set $\boldsymbol{z}_{-\frac{1}{2}}^{\mathrm{ag}} \leftarrow \boldsymbol{z}_0^{[s-1]}, \boldsymbol{z}_0 \leftarrow \boldsymbol{z}_0^{[s-1]}, \boldsymbol{z}_0^{\mathrm{md}} \leftarrow \boldsymbol{z}_0^{[s-1]}$

  for $t = 1, 2, \ldots, T_s$ **do**

    Draw samples $\xi_{t-\frac{1}{2}} \sim \mathcal{D}_\xi$ from oracle, and also $\zeta_{t-\frac{1}{2}}, \zeta_t \sim \mathcal{D}_\zeta$ independently from oracle

    $\boldsymbol{z}_{t-\frac{1}{2}} \leftarrow \mathrm{prox}_{\boldsymbol{z}_{t-1}}^{\eta_t J}\left(\eta_t \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1}; \zeta_{t-\frac{1}{2}}) + \eta_t \nabla \tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}})\right)$

    $\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} \leftarrow (1-\alpha_t)\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}} + \alpha_t \boldsymbol{z}_{t-\frac{1}{2}}$

    $\boldsymbol{z}_t \leftarrow \mathrm{prox}_{\boldsymbol{z}_{t-1}}^{\eta_t J}\left(\eta_t \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t) + \eta_t \nabla \tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}})\right)$

    $\boldsymbol{z}_t^{\mathrm{md}} \leftarrow (1-\alpha_{t+1})\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} + \alpha_{t+1}\boldsymbol{z}_t$

  end for

  Set $\boldsymbol{z}_0^{[s]} \leftarrow \boldsymbol{z}_{t_s-\frac{1}{2}}^{\mathrm{ag}}$ {//Warm-start using the output of the previous epoch}

end for

Output: $\boldsymbol{z}_0^{[\mathscr{S}]}$

D Proofs of Main Results

In this section we present the proofs of our main results. §D.1 illustrates the scaling reduction argument used in the instance of bilinearly-coupled saddle-point problem. §D.2 provides auxiliary lemmas. With a slight adjustment of their presentation order §D.3 proves Theorem 2.5, §D.4 proves Theorem 2.3, §D.5 proves Corollary 2.9 and finally §D.6 proves Corollary 3.3. Throughout the section, we assume that the Bregman divergence $\mathcal{B}(\cdot, \cdot)$ is $\mu_{\mathcal{B}}$-strongly convex.

D.1 Scaling Reduction Argument

Here we illustrate the scaling reduction argument that reduces our analysis of our AG-EG Algorithm 1 under bilinearly-coupled saddle-point problem to the one with equal strong-convexity parameters of F and G using a reparameterized objective function; the same argument applies to Algorithm 2 and we omit the details. The idea is in fact analogous to mirror descent-ascent with respect to a Bregman divergence, and our goal here is to detail this argument for our analysis.

In lieu to (3) we consider

$$\min_{\hat{\boldsymbol{x}}} \max_{\hat{\boldsymbol{y}}} \hat{\mathscr{F}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = F(\hat{\boldsymbol{x}}) + \hat{H}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) - \hat{G}(\hat{\boldsymbol{y}}),$$

where we have $\hat{\mathscr{F}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \mathscr{F}(\boldsymbol{x}, \boldsymbol{y})$ with the symbolic reparameterization $\hat{\boldsymbol{x}} = \boldsymbol{x}$, $\hat{\boldsymbol{y}} = \sqrt{\frac{\mu_G}{\mu_F}}\boldsymbol{y}$, $\hat{H}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = H(\boldsymbol{x}, \boldsymbol{y})$, $\hat{h}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}; \zeta) = h(\boldsymbol{x}, \boldsymbol{y}; \zeta)$, $\hat{G}(\hat{\boldsymbol{y}}) = G(\boldsymbol{y})$, $\hat{g}(\hat{\boldsymbol{y}}; \xi) = g(\boldsymbol{y}; \xi)$ and also their derivatives

$$\nabla_{\hat{\boldsymbol{y}}} \hat{H}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \sqrt{\frac{\mu_F}{\mu_G}} \nabla_{\boldsymbol{y}} H(\boldsymbol{x}, \boldsymbol{y}), \qquad \nabla_{\hat{\boldsymbol{y}}} \hat{h}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}; \zeta) = \sqrt{\frac{\mu_F}{\mu_G}} \nabla_{\boldsymbol{y}} h(\boldsymbol{x}, \boldsymbol{y}; \zeta),$$

and

$$\nabla \hat{G}(\hat{\boldsymbol{y}}) = \sqrt{\frac{\mu_F}{\mu_G}} \nabla G(\boldsymbol{y}), \qquad \nabla \hat{g}(\hat{\boldsymbol{y}}; \xi) = \sqrt{\frac{\mu_F}{\mu_G}} \nabla g(\boldsymbol{y}; \xi).$$

It is straightforward to verify $\hat{\mathscr{F}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ is arguably μ-strongly-convex-μ-strongly-concave. The essence of our update rules is captured by 8 lines corresponding to Lines 5–8 in Algorithm 1, which becomes:

$$\hat{\boldsymbol{x}}_{t-\frac{1}{2}} = \hat{\boldsymbol{x}}_{t-1} - \eta_t \left(\nabla f(\hat{\boldsymbol{x}}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \nabla_{\hat{\boldsymbol{x}}} h(\hat{\boldsymbol{x}}_{t-1}, \hat{\boldsymbol{y}}_{t-1}; \zeta_{t-\frac{1}{2}})\right), \tag{24a}$$

$$\hat{\boldsymbol{y}}_{t-\frac{1}{2}} = \hat{\boldsymbol{y}}_{t-1} - \eta_t \left(-\nabla_{\hat{\boldsymbol{y}}} h(\hat{\boldsymbol{x}}_{t-1}, \hat{\boldsymbol{y}}_{t-1}; \zeta_{t-\frac{1}{2}}) + \nabla g(\hat{\boldsymbol{y}}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}})\right), \tag{24b}$$

$$\hat{\boldsymbol{x}}^{\text{ag}}_{t-\frac{1}{2}} = (1-\alpha_t)\hat{\boldsymbol{x}}^{\text{ag}}_{t-\frac{3}{2}} + \alpha_t \hat{\boldsymbol{x}}_{t-\frac{1}{2}}, \tag{24c}$$

$$\hat{\boldsymbol{y}}^{\text{ag}}_{t-\frac{1}{2}} = (1-\alpha_t)\hat{\boldsymbol{y}}^{\text{ag}}_{t-\frac{3}{2}} + \alpha_t \hat{\boldsymbol{y}}_{t-\frac{1}{2}}, \tag{24d}$$

$$\hat{\boldsymbol{x}}_t = \hat{\boldsymbol{x}}_{t-1} - \eta_t \left(\nabla f(\hat{\boldsymbol{x}}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\hat{\boldsymbol{x}}} h(\hat{\boldsymbol{x}}_{t-\frac{1}{2}}, \hat{\boldsymbol{y}}_{t-\frac{1}{2}}; \zeta_t) \right), \tag{24e}$$

$$\hat{\boldsymbol{y}}_t = \hat{\boldsymbol{y}}_{t-1} - \eta_t \left(-\nabla_{\hat{\boldsymbol{y}}} h(\hat{\boldsymbol{x}}_{t-\frac{1}{2}}, \hat{\boldsymbol{y}}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\hat{\boldsymbol{y}}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) \right), \tag{24f}$$

$$\hat{\boldsymbol{x}}^{\text{md}}_t = (1-\alpha_{t+1})\hat{\boldsymbol{x}}^{\text{ag}}_{t-\frac{1}{2}} + \alpha_{t+1}\hat{\boldsymbol{x}}_t, \tag{24g}$$

$$\hat{\boldsymbol{y}}^{\text{md}}_t = (1-\alpha_{t+1})\hat{\boldsymbol{y}}^{\text{ag}}_{t-\frac{1}{2}} + \alpha_{t+1}\hat{\boldsymbol{y}}_t. \tag{24h}$$

The rest translations are also straightforward, represented by

$$\begin{aligned}
&\hat{\boldsymbol{x}}_{t-\frac{1}{2}} = \hat{\boldsymbol{x}}_{t-1} - \eta_t \left(\nabla f(\hat{\boldsymbol{x}}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\hat{\boldsymbol{x}}} h(\hat{\boldsymbol{x}}_{t-1}, \hat{\boldsymbol{y}}_{t-1}; \zeta_{t-\frac{1}{2}}) \right) \\
&\Leftrightarrow \ \boldsymbol{x}_{t-\frac{1}{2}} = \boldsymbol{x}_{t-1} - \eta_t \left(\nabla f(\boldsymbol{x}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\boldsymbol{x}} h(\boldsymbol{x}_{t-1}, \boldsymbol{y}_{t-1}; \zeta_{t-\frac{1}{2}}) \right),
\end{aligned}$$

as well as

$$\begin{aligned}
&\hat{\boldsymbol{y}}_t = \hat{\boldsymbol{y}}_{t-1} - \eta_t \left(-\nabla_{\hat{\boldsymbol{y}}} h(\hat{\boldsymbol{x}}_{t-\frac{1}{2}}, \hat{\boldsymbol{y}}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\hat{\boldsymbol{y}}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) \right) \\
&\Leftrightarrow \ \boldsymbol{y}_t = \boldsymbol{y}_{t-1} - \eta_t \cdot \tfrac{\mu_F}{\mu_G} \left(-\nabla_{\boldsymbol{y}} h(\boldsymbol{x}_{t-\frac{1}{2}}, \boldsymbol{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\hat{\boldsymbol{y}}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) \right).
\end{aligned}$$

It is also straightforward to justify that Assumptions 2.1 and 2.2 are rediscovered by reverting the scaling reduction from $\hat{\mathscr{F}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ to $\mathscr{F}(\boldsymbol{x}, \boldsymbol{y})$. Therefore, it suffices to analyze Algorithm 1 for $\hat{\mathscr{F}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ and due to this scaling reduction, we only need to prove all results for the case of $\frac{\mu_F}{\mu_G} = 1$. To keep the notations simple, till the rest of this work we slightly abuse the notations and remove the hats in all symbols.

D.2 Auxiliary Lemmas

We first state the following basic lemma to handle the inner-product induced terms for extragradient analysis:

Lemma D.1 . *Given $\boldsymbol{\theta}, \boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \in \mathcal{Z}$, a simple and convex function $J(\cdot)$, and also $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2$ that satisfies*

$$\boldsymbol{\varphi}_1 = \mathit{prox}^J_{\boldsymbol{\theta}}(\boldsymbol{\delta}_1), \qquad \boldsymbol{\varphi}_2 = \mathit{prox}^J_{\boldsymbol{\theta}}(\boldsymbol{\delta}_2), \tag{25}$$

then for any $\mathbf{z} \in \mathcal{Z}$ we have

$$\langle \boldsymbol{\delta}_2, \boldsymbol{\varphi}_1 - \mathbf{z} \rangle + J(\boldsymbol{\varphi}_1) - J(\boldsymbol{z}) \le \frac{1}{2\mu_{\mathcal{B}}} \|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\|^2 + \mathcal{B}(\boldsymbol{\theta}, \mathbf{z}) - \mathcal{B}(\boldsymbol{\varphi}_2, \mathbf{z}) - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\varphi}_1). \tag{26}$$

Furthermore, when taking $J = 0$, $\mathcal{Z} = \mathbb{R}^d$ and $\mathcal{B}(\boldsymbol{z}, \mathbf{u}) = 1/2\|\boldsymbol{z} - \mathbf{u}\|^2$, (26) *reduces to:*

$$\langle \boldsymbol{\delta}_2, \boldsymbol{\varphi}_1 - \mathbf{z} \rangle \le \frac{1}{2}\|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\|^2 + \frac{1}{2}\big[\|\boldsymbol{\theta} - \mathbf{z}\|^2 - \|\boldsymbol{\varphi}_2 - \mathbf{z}\|^2 - \|\boldsymbol{\theta} - \boldsymbol{\varphi}_1\|^2\big]. \tag{27}$$

Proof of Lemma D.1 is provided in §E.1. Lemma D.1 is standard and commonly adopted in extragradient-based analysis; see Lemma 2 of [Chen et al., 2017] for one with similar flavor.

En route to our proofs of Theorems 2.5 and 2.3, we first introduce some notations. Let $\tilde{\boldsymbol{z}} \in \mathcal{Z}$ and let the *pointwise primal-dual gap* function be

$$V(\boldsymbol{z} \mid \tilde{\boldsymbol{z}}) = \mathcal{F}(\boldsymbol{z}) - \mathcal{F}(\tilde{\boldsymbol{z}}) + \langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z} - \tilde{\boldsymbol{z}} \rangle. \tag{28}$$

We prove that this quantity is lower bounded by a positive quadratic function:

Lemma D.2 *For L-smooth and μ-strongly convex $\mathcal{F}(\boldsymbol{z})$, simple and convex J, and for any $\boldsymbol{z} \in \mathcal{Z}$ we have*

$$V(\boldsymbol{z} \mid \boldsymbol{z}^*) = \mathcal{F}(\boldsymbol{z}) - \mathcal{F}(\boldsymbol{z}^*) + \langle \nabla \mathcal{H}(\boldsymbol{z}^*), \boldsymbol{z} - \boldsymbol{z}^* \rangle + J(\boldsymbol{z}) - J(\boldsymbol{z}^*) \ge \frac{\mu}{2} \|\boldsymbol{z} - \boldsymbol{z}^*\|^2. \tag{29}$$

Proof of Lemma D.2 is provided in §E.2. Our final auxiliary lemma on the key properties on step sizes writes as follows:

Lemma D.3 *Our step size choice* (13) *satisfies (i)* $\eta_t \le \frac{t}{B}$*; (ii)* $\left(\frac{t}{\eta_t} : t \ge 1\right)$ *is a nonnegative, nondecreasing arithmetic sequence with common difference* $\sqrt{\frac{1+\beta}{r}}M$*; (iii)* $M\eta_t \le 1$*, and (iv) the step size condition*

$$r - \frac{2L}{t+1}\eta_t - (1+\beta)M^2\eta_t^2 \ge 0. \tag{30}$$

Proof of Lemma D.3 is provided in §E.3.

D.3 Proof of Theorem 2.5

Proof.[Proof of Theorem 2.5]

We first introduce some notations. Denote the incurred stochastic noise terms as

$$\begin{aligned} \boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}} &\equiv \nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}};\xi_{t-\frac{1}{2}}) - \nabla\mathcal{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}}), \qquad \boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}} \equiv \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1};\zeta_{t-\frac{1}{2}}) - \mathcal{H}(\boldsymbol{z}_{t-1}), \\ \boldsymbol{\Delta}_{\mathcal{H}}^{t} &\equiv \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t) - H(\boldsymbol{z}_{t-\frac{1}{2}}). \end{aligned} \tag{31}$$

For our martingale analysis we adopt the filtrations $\mathcal{F}_t^{\xi} \equiv \sigma\left(\xi_s : s = \frac{1}{2}, \frac{3}{2}, \ldots, s \le t\right)$ and $\mathcal{F}_t^{\zeta} \equiv \sigma\left(\zeta_s : s = \frac{1}{2}, 1, \frac{3}{2}, \ldots, s \le t\right)$, and also $\mathcal{F}_t \equiv \sigma(\mathcal{F}_t^{\xi} \cup \mathcal{F}_t^{\zeta})$ be the σ-algebra generated by the union of $\mathcal{F}_t^{\xi}$ and $\mathcal{F}_t^{\zeta}$. We are ready for the proof which proceeds as the following steps:

Step 1. Estimating the primal-dual gap function difference sequence. We provide the following Lemma (D.4), whose proof is in §E.4:

Lemma D.4 *For arbitrary* $\tilde{\boldsymbol{z}} \in \mathcal{Z}$ *the iterates of Algorithm 1 satisfy for* $t = 1, \ldots, T$*, almost surely*

$$\begin{aligned} &V(\boldsymbol{z}_{t-\frac{1}{2}}^{ag} \mid \tilde{\boldsymbol{z}}) - (1-\alpha_t)V(\boldsymbol{z}_{t-\frac{3}{2}}^{ag} \mid \tilde{\boldsymbol{z}}) \\ &\le \alpha_t\langle \nabla\mathcal{F}(\boldsymbol{z}_{t-1}^{md}) + \mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle + \frac{\alpha_t^2 L}{2}\left\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\right\|^2. \end{aligned} \tag{32}$$

Note the proof only relies on the interpolation updates in our algorithm as in Lines 6 and 8, and hence this result holds in a per-trajectory (almost-sure) fashion.

Step 2. We target to prove the following lemma, the complete proof is in §E.5

Lemma D.5 *For our choice of* η_t *that satisfies, for a given* $r \in (0,1)$*,* (30) *of Lemma D.3(iv) that* $r - \frac{2L}{t+1}\eta_t - (1+\beta)M^2\eta_t^2 \ge 0$*, we have for any* $\tilde{\boldsymbol{z}} \in \mathbb{R}^n$*,* $\tilde{\mathbf{y}} \in \mathbb{R}^m$ *and* $t = 1, \ldots, T$ *that*

$$\begin{aligned} &t(t+1)\mathbb{E}[V(\boldsymbol{z}_{t-\frac{1}{2}}^{ag} \mid \tilde{\boldsymbol{z}})] - (t-1)t\mathbb{E}[V(\boldsymbol{z}_{t-\frac{3}{2}}^{ag} \mid \tilde{\boldsymbol{z}})] \\ &\le \frac{t}{\eta_t}\mathbb{E}\left[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2 - \|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2\right] + \left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right)t\eta_t, \end{aligned} \tag{33}$$

Now for a given $1 \le \mathcal{T} \le T$, we finish the proof by telescope the above recursion for $t = 1, \ldots, \mathcal{T}$. We conclude from our choice of step size as in (13) that satisfies (30) so by denoting $\sigma \equiv \frac{1}{\sqrt{3}}\sqrt{\frac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\frac{1}{\beta})\sigma_{\mathrm{Bil}}^2}$, we have by Lemma D.3(i)

$$\begin{aligned} &\left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right)\sum_{t=1}^{\mathcal{T}} t\eta_t = 3\sigma^2\sum_{t=1}^{\mathcal{T}} t\eta_t \le 3\sigma^2 \cdot \frac{1}{B}\sum_{t=1}^{\mathcal{T}} t^2 \\ &= 3\sigma^2 \cdot \frac{C\sqrt{\mathbb{E}[\|\boldsymbol{z}_0 - \tilde{\boldsymbol{z}}\|^2]}}{\sigma[T(T+1)^2]^{1/2}} \cdot \frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{3} = \frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[T(T+1)^2]^{1/2}} \cdot \sigma C\sqrt{\mathbb{E}[\|\boldsymbol{z}_0 - \tilde{\boldsymbol{z}}\|^2]}, \end{aligned}$$

where $B \equiv \frac{\sigma[T(T+1)^2]^{1/2}}{C\sqrt{\mathbb{E}[\|\boldsymbol{z}_0-\tilde{\boldsymbol{z}}\|^2]}}$. Finally by summing over $t=1,\ldots,\mathcal{T}$, we have

$$\begin{aligned}
&\mathcal{T}(\mathcal{T}+1)\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}} \mid \tilde{\boldsymbol{z}})] \\
&\leq \sum_{t=1}^{\mathcal{T}} \frac{t}{\eta_t}\mathbb{E}\left[\|\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\|^2 - \|\boldsymbol{z}_t-\tilde{\boldsymbol{z}}\|^2\right] + \left(\tfrac{1}{1-r}\sigma^2_{\mathrm{Str}} + (2+\tfrac{1}{\beta})\sigma^2_{\mathrm{Bil}}\right)\sum_{t=1}^{\mathcal{T}} t\eta_t \\
&= \frac{1}{\eta_1}\mathbb{E}\|\boldsymbol{z}_0-\tilde{\boldsymbol{z}}\|^2 + \sum_{t=2}^{\mathcal{T}}\left(\frac{t}{\eta_t}-\frac{t-1}{\eta_{t-1}}\right)\mathbb{E}\|\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\|^2 - \frac{\mathcal{T}}{\eta_{\mathcal{T}}}\mathbb{E}\|\boldsymbol{z}_{\mathcal{T}}-\tilde{\boldsymbol{z}}\|^2 \\
&\quad + \frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[T(T+1)^2]^{1/2}}\cdot C\sigma\sqrt{\mathbb{E}\|\boldsymbol{z}_0-\tilde{\boldsymbol{z}}\|^2}.
\end{aligned}$$

Following the above derivations and apply Lemma D.3(ii) we obtain $\frac{t}{\eta_t}-\frac{t-1}{\eta_{t-1}} = \sqrt{\frac{1+\beta}{r}}M$. Rearranging the terms along with Jensen's inequality, and noting that

$$\frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[T(T+1)^2]^{1/2}} \leq \frac{T(T+\frac{1}{2})(T+1)}{[T(T+1)^2]^{1/2}} \leq [T(T+1)^2]^{1/2}$$

proves the following inequality (34).

$$\begin{aligned}
&\mathcal{T}(\mathcal{T}+1)\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}} \mid \tilde{\boldsymbol{z}})] + \frac{\mathcal{T}}{\eta_{\mathcal{T}}}\mathbb{E}\|\boldsymbol{z}_{\mathcal{T}}-\tilde{\boldsymbol{z}}\|^2 \\
&\leq \frac{1}{\eta_1}\mathbb{E}\|\boldsymbol{z}_0-\tilde{\boldsymbol{z}}\|^2 + \sqrt{\tfrac{1+\beta}{r}}M\sum_{t=2}^{\mathcal{T}}\mathbb{E}\|\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\|^2 + [T(T+1)^2]^{1/2}\cdot C\sigma\sqrt{\mathbb{E}\|\boldsymbol{z}_0-\tilde{\boldsymbol{z}}\|^2}.
\end{aligned} \tag{34}$$

Step 3. Bounded Iterates We conduct the following "bootstrapping" argument to arrive at our final theorem. Starting from the recursion (34) we have by setting $\tilde{\boldsymbol{z}} = \boldsymbol{z}^*$, Lemma D.2 implies that its first summand on the left hand $\mathcal{T}(\mathcal{T}+1)\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}} \mid \boldsymbol{z}^*)]$ is nonnegative, and hence we can drop it and have for any $\mathcal{T}=1,\ldots,T$

$$\frac{\mathcal{T}}{\eta_{\mathcal{T}}}\mathbb{E}\|\boldsymbol{z}_{\mathcal{T}}-\boldsymbol{z}^*\|^2 \tag{35}$$

$$\begin{aligned}
&\leq \frac{1}{\eta_1}\mathbb{E}\|\boldsymbol{z}_0-\boldsymbol{z}^*\|^2 + \sqrt{\tfrac{1+\beta}{r}}M\sum_{t=2}^{\mathcal{T}}\mathbb{E}\|\boldsymbol{z}_{t-1}-\boldsymbol{z}^*\|^2 + [T(T+1)^2]^{1/2}\cdot C\sigma\sqrt{\mathbb{E}\left\|\boldsymbol{z}_0-\boldsymbol{z}^*\right\|^2} \\
&= (\tfrac{2}{r}L\vee B)\mathbb{E}\|\boldsymbol{z}_0-\boldsymbol{z}^*\|^2 + \sqrt{\tfrac{1+\beta}{r}}M\underbrace{\sum_{t=1}^{\mathcal{T}}\mathbb{E}\|\boldsymbol{z}_{t-1}-\boldsymbol{z}^*\|^2}_{\equiv\mathcal{Q}_{\mathcal{T}-1}} + \underbrace{[T(T+1)^2]^{1/2}\cdot C\sigma\sqrt{\mathbb{E}\left\|\boldsymbol{z}_0-\boldsymbol{z}^*\right\|^2}}_{\mathcal{R}_0}.
\end{aligned} \tag{36}$$

Converting (36) to a version of partial sum $\mathcal{Q}_{\mathcal{T}-1} \equiv \sum_{t=1}^{\mathcal{T}}\mathbb{E}\|\boldsymbol{z}_{t-1}-\boldsymbol{z}^*\|^2$ that for all $\mathcal{T}=1,\ldots,T$

$$\frac{\mathcal{T}}{\eta_{\mathcal{T}}}\mathbb{E}\|\boldsymbol{z}_{\mathcal{T}}-\boldsymbol{z}^*\|^2 = \frac{\mathcal{T}}{\eta_{\mathcal{T}}}(\mathcal{Q}_{\mathcal{T}}-\mathcal{Q}_{\mathcal{T}-1}) \leq \sqrt{\tfrac{1+\beta}{r}}M\mathcal{Q}_{\mathcal{T}-1} + \underbrace{\mathcal{R}_0 + (\tfrac{2}{r}L\vee B)\mathcal{Q}_0}_{\mathcal{D}_0}. \tag{37}$$

(37) is equivalently written as

$$\frac{\mathcal{T}}{\eta_{\mathcal{T}}}\mathcal{Q}_{\mathcal{T}} \leq \frac{\mathcal{T}+1}{\eta_{\mathcal{T}+1}}\mathcal{Q}_{\mathcal{T}-1} + \mathcal{D}_0.$$

From here and onwards, we denote $\kappa_t \equiv \frac{t}{\eta_t} = \frac{2}{r}L\vee B + \sqrt{\frac{1+\beta}{r}}Mt$ for each $t=1,\ldots,T$. Dividing both sides of the above display by $\kappa_{\mathcal{T}}\kappa_{\mathcal{T}+1} = \frac{\mathcal{T}}{\eta_{\mathcal{T}}}\cdot\frac{\mathcal{T}+1}{\eta_{\mathcal{T}+1}}$ gives

$$\frac{\mathcal{Q}_{\mathcal{T}}}{\kappa_{\mathcal{T}+1}} \leq \frac{\mathcal{Q}_{\mathcal{T}-1}}{\kappa_{\mathcal{T}}} + \frac{\mathcal{D}_0}{\kappa_{\mathcal{T}}\cdot\kappa_{\mathcal{T}+1}}.$$

Telescoping up from $1, \dots, \mathcal{T}-1$ for $1 \le \mathcal{T} \le T$ yields

$$\frac{\mathcal{Q}_{\mathcal{T}-1}}{\kappa_{\mathcal{T}}} \le \frac{\mathcal{Q}_0}{\kappa_1} + \sum_{t=1}^{\mathcal{T}-1} \frac{\mathcal{D}_0}{\kappa_t \cdot \kappa_{t+1}} \le \frac{\mathcal{Q}_0}{\kappa_1} + \mathcal{D}_0 \sum_{t=1}^{\mathcal{T}-1} \frac{1}{\kappa_t \cdot \kappa_{t+1}},$$

where we applied Lemma D.3(ii) that for all $t = 1, \dots, \mathcal{T}-1$ we have $\kappa_{t+1} - \kappa_t = \sqrt{\frac{1+\beta}{r}}M$. This yields

$$\sqrt{\tfrac{1+\beta}{r}}M \sum_{t=1}^{\mathcal{T}-1} \frac{1}{\kappa_t \cdot \kappa_{t+1}} = \sum_{t=1}^{\mathcal{T}-1} \left[\frac{1}{\kappa_t} - \frac{1}{\kappa_{t+1}}\right] = \frac{1}{\kappa_1} - \frac{1}{\kappa_{\mathcal{T}}},$$

and hence

$$\sqrt{\tfrac{1+\beta}{r}}M \frac{\mathcal{Q}_{\mathcal{T}-1}}{\kappa_{\mathcal{T}}} \le \sqrt{\tfrac{1+\beta}{r}}M \frac{\mathcal{Q}_0}{\kappa_1} + \mathcal{D}_0 \sqrt{\tfrac{1+\beta}{r}}M \sum_{t=1}^{\mathcal{T}-1} \frac{1}{\kappa_t \cdot \kappa_{t+1}} = \sqrt{\tfrac{1+\beta}{r}}M \frac{\mathcal{Q}_0}{\kappa_1} + \mathcal{D}_0 \left(\frac{1}{\kappa_1} - \frac{1}{\kappa_{\mathcal{T}}}\right).$$

Next, we rearrange the above quantity and derive

$$\frac{\sqrt{\frac{1+\beta}{r}}M\mathcal{Q}_0 + \mathcal{D}_0}{\kappa_1} - \frac{\mathcal{D}_0}{\kappa_{\mathcal{T}}} = \frac{\sqrt{\frac{1+\beta}{r}}M\mathcal{Q}_0 + \left(\mathcal{R}_0 + (\frac{2}{r}L \vee B)\mathcal{Q}_0\right)}{\kappa_1} - \frac{\mathcal{D}_0}{\kappa_{\mathcal{T}}} = \mathcal{Q}_0 + \frac{\mathcal{R}_0}{\kappa_1} - \frac{\mathcal{D}_0}{\kappa_{\mathcal{T}}}.$$

Plugging this into (37) we have for all iterates $1 \le \mathcal{T} \le T$

$$\begin{aligned} \mathbb{E}\|\boldsymbol{z}_{\mathcal{T}} - \boldsymbol{z}^*\|^2 &\le \sqrt{\tfrac{1+\beta}{r}}M \frac{\mathcal{Q}_{\mathcal{T}-1}}{\kappa_{\mathcal{T}}} + \frac{\mathcal{D}_0}{\kappa_{\mathcal{T}}} \le \mathcal{Q}_0 + \frac{\mathcal{R}_0}{\kappa_1} \le \left(1 + \frac{C\sigma[T(T+1)^2]^{1/2}}{\kappa_1\sqrt{\mathcal{Q}_0}}\right)\mathcal{Q}_0 \\ &= \underbrace{\left(1 + C^2 B\eta_1\right)}_{A} \mathbb{E}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2, \end{aligned} \tag{38}$$

where the prefactor A lies in $[1, 1+C^2]$ and reduces to 1 when the argument is set as 0.

Now we drop the second summand on the left hand of (34) with $\boldsymbol{z}^* = \boldsymbol{z}^*, \mathcal{T} = T$. Combining with (38) gives

$$\begin{aligned} &\mathcal{T}(\mathcal{T}+1)\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}} \mid \boldsymbol{z}^*)] \\ &\le \frac{1}{\eta_1}\mathbb{E}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2 + \sqrt{\tfrac{1+\beta}{r}}M \sum_{t=2}^{\mathcal{T}} \mathbb{E}\|\boldsymbol{z}_{t-1} - \boldsymbol{z}^*\|^2 + [T(T+1)^2]^{1/2} \cdot C\sigma\sqrt{\mathbb{E}\left\|\boldsymbol{z}_0 - \boldsymbol{z}^*\right\|^2} \\ &\le \left(\tfrac{2}{r}L \vee B + \sqrt{\tfrac{1+\beta}{r}}M\right)\mathbb{E}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2 \\ &\quad + \sqrt{\tfrac{1+\beta}{r}}M(T-1) \cdot A \cdot \mathbb{E}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2 + C\sigma[T(T+1)^2]^{1/2}\sqrt{\mathbb{E}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2} \\ &\le \left(\tfrac{2}{r}L + A\sqrt{\tfrac{1+\beta}{r}}MT\right)\mathbb{E}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2 + (\tfrac{1}{C} + C)\sigma[T(T+1)^2]^{1/2}\sqrt{\mathbb{E}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2}. \end{aligned}$$

Using (29) in Lemma D.2 again lower bounds the left hand in the last display as

$$T(T+1)\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{T-\frac{1}{2}} \mid \boldsymbol{z}^*)] \ge \frac{\mu}{2}T(T+1)\mathbb{E}\left\|\boldsymbol{z}^{\mathrm{ag}}_{T-\frac{1}{2}} - \boldsymbol{z}^*\right\|^2 \ge 0.$$

Dividing both sides by $\frac{\mu}{2}T(T+1)$ concludes

$$\mathbb{E}\left\|\boldsymbol{z}^{\mathrm{ag}}_{T-\frac{1}{2}} - \boldsymbol{z}^*\right\|^2 \le \frac{2\left(\frac{2}{r}L + A\sqrt{\frac{1+\beta}{r}}MT\right)}{\mu T(T+1)}\mathbb{E}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2 + \frac{2(\frac{1}{C} + C)\sigma}{\mu T^{1/2}}\sqrt{\mathbb{E}\|\boldsymbol{z}_0 - \boldsymbol{z}^*\|^2},$$

and hence concludes (14) and the whole proof of Theorem 2.5.

D.4 Proof of Theorem 2.3

We overload function notations $\mathcal{F}, \mathcal{H}$ to the new group accordingly where $\mathcal{F}(\boldsymbol{z}) \leftarrow \mathcal{F}(\boldsymbol{z}) - \frac{\mu_\star}{2}\|\boldsymbol{z} - \boldsymbol{z}_0\|^2$ is non-strongly convex and $\mathcal{H}(\boldsymbol{z}) \leftarrow \mathcal{H}(\boldsymbol{z}) + \mu_\star\big(\boldsymbol{z} - \boldsymbol{z}_0\big)$. For convenience we repeat the iterates of Algorithm 2 as

$$\begin{aligned}
\boldsymbol{z}_{t-\frac{1}{2}} &= \boldsymbol{z}_{t-1} - \eta_t \left(\nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1}; \zeta_{t-\frac{1}{2}}) - \mu(\boldsymbol{z}_{t-1}^{\mathrm{md}} - \boldsymbol{z}_{t-1})\right), \\
\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} &= (1-\alpha_t)\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}} + \alpha_t \boldsymbol{z}_{t-\frac{1}{2}}, \\
\boldsymbol{z}_t &= \boldsymbol{z}_{t-1} - \eta_t \left(\nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t) - \mu(\boldsymbol{z}_{t-1}^{\mathrm{md}} - \boldsymbol{z}_{t-\frac{1}{2}})\right), \\
\boldsymbol{z}_t^{\mathrm{md}} &= (1-\alpha_{t+1})\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} + \alpha_{t+1}\boldsymbol{z}_t,
\end{aligned}$$

with the initialization $\boldsymbol{z}_0 = \boldsymbol{z}_0^{\mathrm{md}} = \boldsymbol{z}_{-\frac{1}{2}}^{\mathrm{ag}} \in \mathbb{R}^{n+m}$. We continue to assume the noise-related setting as in (31). Our proof proceeds in the following steps:

Step 1. We prove the following generalization of Lemma D.4, whose proof is in §E.6:

Lemma D.6 *For arbitrary $\tilde{\boldsymbol{z}} \in \mathbb{R}^{n+m}$ and $\alpha_t \in (0,1]$ the iterates of Algorithm 2 satisfy almost surely*

$$\begin{aligned}
&V(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}}) - (1-\alpha_t)V(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}}) \\
&\le \alpha_t\langle \nabla\mathcal{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}}) + \mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle + \tfrac{\alpha_t^2 L}{2}\left\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\right\|^2 - \alpha_t\mu_\star\left\|\boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\right\|^2.
\end{aligned} \tag{39}$$

Step 2. Analogous to Step 2 in the proof of Theorem 2.5 in §D.3 we conclude for all $\boldsymbol{z} \in \mathbb{R}^n$, $\boldsymbol{z} \in \mathbb{R}^m$,

$$\begin{aligned}
&\eta_t\mathbb{E}\langle \nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle \\
&\le \frac{1}{2}\left(\mathbb{E}[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2] - \mathbb{E}[\|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2]\right) - \frac{1-(1+\beta)M^2\eta_t^2}{2}\mathbb{E}\left[\left\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\right\|^2\right] \\
&\quad + \frac{\eta_t^2}{2}(2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2.
\end{aligned}$$

To show this, note that

$$\begin{aligned}
&\eta_t\langle \nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle \\
&\le \frac{1}{2}\left(\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2 - \|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2 - \|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\|^2\right) + \frac{\eta_t^2}{2}\|\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t) - \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1}; \zeta_{t-\frac{1}{2}})\|^2.
\end{aligned}$$

To handle the stochastic terms, Young's inequality combined with the martingale structure, along with the definition of M, indicates

$$\begin{aligned}
&\mathbb{E}\left\|\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t) - \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1}; \zeta_{t-\frac{1}{2}})\right\|^2 = \mathbb{E}\left\|\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}) - \mathcal{H}(\boldsymbol{z}_{t-1}) - \boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2 + \mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2 \\
&\le (1+\beta)M^2\mathbb{E}\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\|^2 + (1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2 + \mathbb{E}\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\|^2.
\end{aligned}$$

Combining the last three displays gives

$$\begin{aligned}
&\eta_t\mathbb{E}\langle \nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle \\
&\le \frac{1}{2}\left(\mathbb{E}[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2] - \mathbb{E}[\|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2]\right) - \frac{1-(1+\beta)M^2\eta_t^2}{2}\mathbb{E}\left[\left\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\right\|^2\right] \\
&\quad + \frac{\eta_t^2}{2}\left((1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2 + \mathbb{E}\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\|^2\right).
\end{aligned} \tag{40}$$

Combining this with Lemma D.6, we have

$$
\begin{aligned}
&\mathbb{E}[V(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}})] - (1-\alpha_t)\mathbb{E}[V(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}})] \\
&\le \alpha_t \mathbb{E}\langle \nabla \mathcal{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}}) + \mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \rangle + \tfrac{\alpha_t^2 L}{2} \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right] - \alpha_t \mu_\star \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \right\|^2 \right] \\
&= \alpha_t \mathbb{E}\langle \nabla \tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \rangle \\
&\quad - \alpha_t \mathbb{E}\langle \boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}} + \boldsymbol{\Delta}_{\mathcal{H}}^{t}, \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \rangle + \tfrac{\alpha_t^2 L}{2} \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right] - \alpha_t \mu_\star \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \right\|^2 \right] \\
&\le \frac{\alpha_t}{\eta_t} \left(\frac{1}{2} \left(\mathbb{E}[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2] - \mathbb{E}[\|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2] \right) - \frac{1 - (1+\beta) M^2 \eta_t^2}{2} \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right] \right. \\
&\quad \left. + \frac{\eta_t^2}{2} (2 + \tfrac{1}{\beta}) \sigma_{\mathrm{Bil}}^2 \right) - \alpha_t \mathbb{E}\langle \boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}} + \boldsymbol{\Delta}_{\mathcal{H}}^{t}, \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \rangle + \tfrac{\alpha_t^2 L}{2} \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right] \\
&\quad - \alpha_t \mu_\star \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \right\|^2 \right].
\end{aligned}
$$

Continuing this estimation gives (note Young's inequality applies, and $\mathbb{E}\langle \boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}} + \boldsymbol{\Delta}_{\mathcal{H}}^{t}, \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \rangle = \mathbb{E}\langle \boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}, \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \rangle$)

$$
\begin{aligned}
&\mathbb{E}[V(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}})] - (1-\alpha_t)\mathbb{E}[V(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}})] \\
&\le \frac{\alpha_t}{2\eta_t} \left(\mathbb{E}[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2] - \mathbb{E}[\|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2] \right) - \frac{\alpha_t}{2\eta_t} \left(r - \alpha_t L \eta_t - (1+\beta) M^2 \eta_t^2 \right) \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right] \\
&\quad + \frac{\alpha_t \eta_t}{2} (2 + \tfrac{1}{\beta}) \sigma_{\mathrm{Bil}}^2 - \frac{\alpha_t (1-r)}{2\eta_t} \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right] \\
&\quad - \alpha_t \mathbb{E}\langle \boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}, \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \rangle - \alpha_t \mu_\star \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \right\|^2 \right] \\
&\le \frac{\alpha_t}{2\eta_t} \left(\mathbb{E}[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2] - \mathbb{E}[\|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2] \right) - \frac{\alpha_t}{2\eta_t} \left(r - \alpha_t L \eta_t - (1+\beta) M^2 \eta_t^2 \right) \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right] \\
&\quad + \frac{\alpha_t \eta_t}{2} (2 + \tfrac{1}{\beta}) \sigma_{\mathrm{Bil}}^2 + \frac{\alpha_t \eta_t}{2(1-r)} \mathbb{E}\left\| \boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}} \right\|^2 - \alpha_t \mu_\star \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \right\|^2 \right] \\
&\le \frac{\alpha_t}{2\eta_t} \left(\mathbb{E}\left[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2 \right] - \mathbb{E}\left[\left\| \boldsymbol{z}_t - \tilde{\boldsymbol{z}} \right\|^2 \right] \right) - \frac{\alpha_t}{2\eta_t} \left(r - \alpha_t L \eta_t - (1+\beta) M^2 \eta_t^2 \right) \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right] \\
&\quad - \alpha_t \mu_\star \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \right\|^2 \right] + \frac{\alpha_t \eta_t}{2} \left(\tfrac{1}{1-r} \sigma_{\mathrm{Str}}^2 + (2 + \tfrac{1}{\beta}) \sigma_{\mathrm{Bil}}^2 \right).
\end{aligned}
$$

By applying Young's inequality, it yields

$$
\begin{aligned}
&\mathbb{E}[V(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}})] - (1-\alpha_t)\mathbb{E}[V(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}})] - \frac{\alpha_t \eta_t}{2} \left(\tfrac{1}{1-r} \sigma_{\mathrm{Str}}^2 + (2 + \tfrac{1}{\beta}) \sigma_{\mathrm{Bil}}^2 \right) \\
&\quad + \frac{\alpha_t}{2\eta_t} \left(r - \alpha_t L \eta_t - (1+\beta) M^2 \eta_t^2 \right) \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right] \\
&\le \frac{\alpha_t}{2\eta_t} \left(\mathbb{E}[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2] - \mathbb{E}[\|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2] \right) - \alpha_t \mu_\star \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \right\|^2 \right] \\
&\le \frac{\alpha_t}{2\eta_t} \left((1-\alpha_t) \mathbb{E}[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2] - \mathbb{E}[\|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2] \right) + \frac{\alpha_t^2}{2\eta_t} \mathbb{E}[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2] - \alpha_t \mu_\star \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} \right\|^2 \right] \\
&\le \frac{\alpha_t}{2\eta_t} \left((1-\alpha_t) \mathbb{E}[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2] - \mathbb{E}[\|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2] \right) + \eta_t \mu_\star^2 \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right].
\end{aligned}
$$

Setting $\eta_t = \frac{\alpha_t}{\mu_\star}$ we have

$$
\begin{aligned}
&\mathbb{E}[V(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}})] + \frac{\mu_\star}{2} \mathbb{E}[\|\boldsymbol{z}_t - \tilde{\boldsymbol{z}}\|^2] - (1-\alpha_t) \left(\mathbb{E}[V(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}})] + \frac{\mu_\star}{2} \mathbb{E}[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2] \right) \\
&\le -\frac{\mu_\star}{2} \left(r - 2\alpha_t - \left(\tfrac{L}{\mu_\star} + \tfrac{(1+\beta) M^2}{\mu_\star^2} \right) \alpha_t^2 \right) \mathbb{E}\left[\left\| \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1} \right\|^2 \right]
\end{aligned}
$$

$$+\frac{\alpha_t^2}{2\mu_\star}\left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2+(2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right).$$

Step 3. By the definition α_t we have $r-2\alpha_t-\left(\frac{L}{\mu_\star}+\frac{(1+\beta)M^2}{\mu_\star^2}\right)\alpha_t^2\geq 0$, so we obtain regularity condition $\alpha_t\leq\bar{\alpha}=\frac{r}{1+\sqrt{1+r\left(\frac{L}{\mu_\star}+\frac{(1+\beta)M^2}{\mu_\star^2}\right)}}$ of Theorem 2.3. Since we assumed both F and G are nonstrongly convex and H is a $\mu_\star$-strongly-convex-$\mu_\star$-strongly-concave isotropic quadratic, this implies

$$\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}\mid\tilde{\boldsymbol{z}})]+\frac{\mu_\star}{2}\mathbb{E}[\|\boldsymbol{z}_t-\tilde{\boldsymbol{z}}\|^2]\leq(1-\alpha_t)\left(\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}}\mid\tilde{\boldsymbol{z}})]+\frac{\mu_\star}{2}\mathbb{E}[\|\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\|^2]\right)+\frac{3\alpha_t^2}{2\mu_\star}\sigma^2.$$

Plugging in $\tilde{\boldsymbol{z}}=\boldsymbol{z}^*$ gives

$$\mathbb{E}[V(\tilde{\boldsymbol{z}}\mid\boldsymbol{z}^*)]=\mathcal{F}(\tilde{\boldsymbol{z}})-\mathcal{F}(\boldsymbol{z}^*)+\langle\mathcal{H}(\boldsymbol{z}^*),\tilde{\boldsymbol{z}}-\boldsymbol{z}^*\rangle\geq\langle\nabla\mathcal{F}(\boldsymbol{z}^*)+\mathcal{H}(\boldsymbol{z}^*),\tilde{\boldsymbol{z}}-\boldsymbol{z}^*\rangle=0,$$

and also

$$\mathbb{E}[V(\tilde{\boldsymbol{z}}\mid\boldsymbol{z}^*)]\leq\langle\nabla\mathcal{F}(\boldsymbol{z}^*)+\mathcal{H}(\boldsymbol{z}^*),\tilde{\boldsymbol{z}}-\boldsymbol{z}^*\rangle+\tfrac{L}{2}\|\tilde{\boldsymbol{z}}-\boldsymbol{z}^*\|^2=\tfrac{L}{2}\|\tilde{\boldsymbol{z}}-\boldsymbol{z}^*\|^2,$$

so (by the fact that $\boldsymbol{z}^{\mathrm{ag}}_{-\frac{1}{2}}=\boldsymbol{z}_0$ and $\boldsymbol{z}^{\mathrm{ag}}_{-\frac{1}{2}}=\boldsymbol{z}_0$)

$$\begin{aligned}
&\frac{\mu_\star}{2}\mathbb{E}[\|\boldsymbol{z}_t-\boldsymbol{z}^*\|^2]\leq\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}\mid\boldsymbol{z}^*)+\frac{\mu_\star}{2}\mathbb{E}[\|\boldsymbol{z}_t-\boldsymbol{z}^*\|^2]\\
&\leq\left(V(\boldsymbol{z}^{\mathrm{ag}}_{-\frac{1}{2}}\mid\boldsymbol{z}^*)+\frac{\mu_\star}{2}\|\boldsymbol{z}_0-\boldsymbol{z}^*\|^2\right)\prod_{\tau=1}^{t}(1-\alpha_\tau)+\sum_{\tau=1}^{t}\frac{3\alpha_\tau^2}{2\mu_\star}\left[\prod_{\tau'=\tau+1}^{t}(1-\alpha_{\tau'})\right]\sigma^2\\
&\leq\|\boldsymbol{z}_0-\boldsymbol{z}^*\|^2\frac{L+\mu_\star}{2}\prod_{\tau=1}^{t}(1-\alpha_\tau)+\frac{3\sigma^2}{2\mu_\star}\sum_{\tau=1}^{t}\alpha_\tau^2\prod_{\tau'=\tau+1}^{t}(1-\alpha_{\tau'}).
\end{aligned}$$

Dividing both sides by $\frac{\mu_\star}{2}$ gives (11) and our theorem.

D.5 Proof of Corollary 2.9

The proof of Corollary 2.9 mostly follows the proof of Theorem 2.5 and Corollary 2.8, except that we modify some steps to adapt to the proximal operator. The proof is as follows:

Step 1. Estimating the primal-dual gap function difference sequence. We have the following Lemma (D.7), whose proof is in §E.7:

Lemma D.7 *For arbitrary $\tilde{\boldsymbol{z}}\in\mathcal{Z}$ the iterates of Algorithm 1 satisfy for $t=1,\ldots,T$, almost surely*

$$\begin{aligned}
&V(\boldsymbol{z}^{ag}_{t-\frac{1}{2}}\mid\tilde{\boldsymbol{z}})-(1-\alpha_t)V(\boldsymbol{z}^{ag}_{t-\frac{3}{2}}\mid\tilde{\boldsymbol{z}})\\
&\leq\alpha_t\langle\nabla\mathcal{F}(\boldsymbol{z}^{md}_{t-1})+\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\rangle+\tfrac{\alpha_t^2L}{2}\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2+\alpha_t\left(J(\boldsymbol{z}_{t-\frac{1}{2}})-J(\tilde{\boldsymbol{z}})\right).
\end{aligned}\tag{41}$$

Note the proof only relies on the interpolation updates in our algorithm as in Lines 6 and 8, and hence this result holds in a per-trajectory (almost-sure) fashion.

Step 2. We target to prove the following lemma, the complete proof is in §E.8:

Lemma D.8 *For our choice of η_t that satisfies, for a given $r\in(0,1)$, that*

$$r\mu_{\mathcal{B}}-\frac{2L}{t+1}\eta_t-\frac{(1+\beta)M^2\eta_t^2}{\mu_{\mathcal{B}}}\geq 0,\tag{42}$$

we have for any $\tilde{\boldsymbol{z}}\in\mathbb{R}^n$, $\tilde{\mathbf{y}}\in\mathbb{R}^m$ and $t=1,\ldots,T$ that

$$\begin{aligned}
&t(t+1)\mathbb{E}[V(\boldsymbol{z}^{ag}_{t-\frac{1}{2}}\mid\tilde{\boldsymbol{z}})]-(t-1)t\mathbb{E}[V(\boldsymbol{z}^{ag}_{t-\frac{3}{2}}\mid\tilde{\boldsymbol{z}})]\\
&\leq\frac{2t}{\eta_t}\left(\mathbb{E}[\mathcal{B}(\boldsymbol{z}_{t-1},\tilde{\boldsymbol{z}})]-\mathbb{E}[\mathcal{B}(\boldsymbol{z}_t,\tilde{\boldsymbol{z}})]\right)+\left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2+(2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right)\frac{t\eta_t}{\mu_{\mathcal{B}}},
\end{aligned}\tag{43}$$

We note that (43) in Lemma D.8 only differs with (33) in Lemma D.5 by the use of Bregman distance $\mathcal{B}$ and a factor of $1/\mu_{\mathcal{B}}$ on the variance term. Following similar derivations as in the proof of Theorem 2.5, we telescope the above recursion for $t = 1, \ldots, \mathcal{T}$ and choose the step size as

$$\mu_{\mathcal{B}} \cdot \tfrac{t}{\eta_t} = \tfrac{2}{r} L \vee B + \sqrt{\tfrac{1+\beta}{r}} M t, \tag{44}$$

with $B = \frac{\sigma\sqrt{T}(T+1)}{C\sqrt{\frac{2}{\mu_{\mathcal{B}}}\mathcal{B}(\boldsymbol{z}_0, \tilde{\boldsymbol{z}})}}$ that satisfies (42). By denoting $\sigma \equiv \frac{1}{\sqrt{3}}\sqrt{\frac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\frac{1}{\beta})\sigma_{\mathrm{Bil}}^2}$, we have by Lemma D.3(i) and the same derivative as in §D.3

$$\begin{aligned}
&\Big(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\Big)\sum_{t=1}^{\mathcal{T}} t\eta_t \le 3\sigma^2 \cdot \frac{\mu_{\mathcal{B}}}{B}\sum_{t=1}^{\mathcal{T}} t^2 \\
&= \frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[T(T+1)^2]^{1/2}} \cdot \sigma\mu_{\mathcal{B}} C\sqrt{\frac{2}{\mu_{\mathcal{B}}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_0, \tilde{\boldsymbol{z}})},
\end{aligned}$$

Finally by summing over $t = 1, \ldots, \mathcal{T}$, we have

$$\begin{aligned}
&\mathcal{T}(\mathcal{T}+1)\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}} \mid \tilde{\boldsymbol{z}})] \\
&\le \sum_{t=1}^{\mathcal{T}} \frac{2t}{\eta_t}\left(\mathbb{E}[\mathcal{B}(\boldsymbol{z}_{t-1}, \tilde{\boldsymbol{z}})] - \mathbb{E}[\mathcal{B}(\boldsymbol{z}_t, \tilde{\boldsymbol{z}})]\right) + \Big(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\Big)\sum_{t=1}^{\mathcal{T}} \frac{t\eta_t}{\mu_{\mathcal{B}}} \\
&= \frac{2}{\eta_1}\mathcal{B}(\boldsymbol{z}_0, \tilde{\boldsymbol{z}}) + 2\sum_{t=2}^{\mathcal{T}}\left(\frac{t}{\eta_t} - \frac{t-1}{\eta_{t-1}}\right)\mathbb{E}\mathcal{B}(\boldsymbol{z}_{t-1}, \tilde{\boldsymbol{z}}) - \frac{2\mathcal{T}}{\eta_{\mathcal{T}}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_{\mathcal{T}}, \tilde{\boldsymbol{z}}) \\
&\quad + \frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[T(T+1)^2]^{1/2}} \cdot C\sigma\sqrt{\frac{2}{\mu_{\mathcal{B}}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_0, \tilde{\boldsymbol{z}})} \\
&= \frac{2}{\eta_1}\mathcal{B}(\boldsymbol{z}_0, \tilde{\boldsymbol{z}}) + 2\sqrt{\tfrac{1+\beta}{r}}\frac{M}{\mu_{\mathcal{B}}}\sum_{t=2}^{\mathcal{T}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_{t-1}, \tilde{\boldsymbol{z}}) - \frac{2\mathcal{T}}{\eta_{\mathcal{T}}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_{\mathcal{T}}, \tilde{\boldsymbol{z}}) \\
&\quad + [T(T+1)^2]^{1/2} \cdot C\sigma\sqrt{\frac{2}{\mu_{\mathcal{B}}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_0, \tilde{\boldsymbol{z}})}.
\end{aligned}$$

Rearranging the terms proves the following inequality (45).

$$\begin{aligned}
&\mathcal{T}(\mathcal{T}+1)\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}} \mid \tilde{\boldsymbol{z}})] + \frac{2\mathcal{T}}{\eta_{\mathcal{T}}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_{\mathcal{T}}, \tilde{\boldsymbol{z}}) \\
&\le \frac{2}{\eta_1}\mathcal{B}(\boldsymbol{z}_0, \tilde{\boldsymbol{z}}) + 2\sqrt{\tfrac{1+\beta}{r}}\frac{M}{\mu_{\mathcal{B}}}\sum_{t=2}^{\mathcal{T}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_{t-1}, \tilde{\boldsymbol{z}}) + [T(T+1)^2]^{1/2} \cdot C\sigma\sqrt{\frac{2}{\mu_{\mathcal{B}}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_0, \tilde{\boldsymbol{z}})}.
\end{aligned} \tag{45}$$

The same bootstrapping argument gives

$$\mathbb{E}\mathcal{B}(\boldsymbol{z}_{t-1}, \boldsymbol{z}^*) \le \big(1 + C^2 B\eta_1\big)\,\mathbb{E}\mathcal{B}(\boldsymbol{z}_0, \boldsymbol{z}^*),$$

which further derives

$$\begin{aligned}
\mathcal{T}(\mathcal{T}+1)\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}} \mid \boldsymbol{z}^*)] &\le \frac{2}{\mu_{\mathcal{B}}}\left(\tfrac{2}{r}L + A\sqrt{\tfrac{1+\beta}{r}}MT\right)\mathbb{E}\mathcal{B}(\boldsymbol{z}_0, \boldsymbol{z}^*) \\
&\quad + (\tfrac{1}{C} + C)\sigma[T(T+1)^2]^{1/2}\sqrt{\frac{2}{\mu_{\mathcal{B}}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_0, \boldsymbol{z}^*)}
\end{aligned}$$

Again we can lower bounds the left hand in the last display as

$$T(T+1)\mathbb{E}[V(\boldsymbol{z}^{\mathrm{ag}}_{T-\frac{1}{2}} \mid \boldsymbol{z}^*)] \ge \frac{\mu}{2} \cdot T(T+1)\mathbb{E}\mathcal{B}\Big(\boldsymbol{z}^{\mathrm{ag}}_{T-\frac{1}{2}}, \boldsymbol{z}^*\Big) \ge 0.$$

Dividing both sides by $\frac{\mu}{2}T(T+1)$ concludes

$$\mathbb{E}\mathcal{B}\Big(\boldsymbol{z}^{\mathrm{ag}}_{T-\frac{1}{2}}, \boldsymbol{z}^*\Big) \le \frac{4\left(\frac{2}{r}L + A\sqrt{\frac{1+\beta}{r}}MT\right)}{\mu\mu_{\mathcal{B}}T(T+1)}\mathbb{E}\mathcal{B}(\boldsymbol{z}_0, \boldsymbol{z}^*) + \frac{2(\frac{1}{C}+C)\sigma}{\mu T^{1/2}}\sqrt{\frac{2}{\mu_{\mathcal{B}}}\mathbb{E}\mathcal{B}(\boldsymbol{z}_0, \boldsymbol{z}^*)},$$

The rest of the proof follows the same bounded iterates argument and the restarting argument exactly as in the previous proof of Theorem 2.5 with only a difference in a factor of $\mu_{\mathcal{B}}$. Similar derivatives gives us a total iteration complexity of

$$\mathcal{O}\left(\left(\sqrt{\tfrac{L}{\mu\mu_{\mathcal{B}}}}+\tfrac{M}{\mu\mu_{\mathcal{B}}}\right)\log\left(\tfrac{1}{\varepsilon}\right)+\tfrac{\sigma^2\mathcal{B}(\boldsymbol{z}_0,\boldsymbol{z}^*)}{\mu^2\mu_{\mathcal{B}}\varepsilon^2}\right)$$

with epoch length $T_s \asymp \sqrt{\frac{L}{\mu\mu_{\mathcal{B}}}}+\frac{M}{\mu\mu_{\mathcal{B}}}+\frac{\sigma^2\mathcal{B}(\boldsymbol{z}_0,\boldsymbol{z}^*)}{\mu^2\mu_{\mathcal{B}}\Gamma_0^2 e^{1-s}}$.

D.6 Proof of Corollary 3.3

Before the proof we first adopt the scaling reduction argument as in §D.1, to argue that we only need to prove the result for the case of bilinear games centered at zero, i.e., $F(\boldsymbol{x})=0=G(\boldsymbol{y})$ we have $L=\mu=\mu_F=0$. We set the iteration symbol $\mathbf{z}\equiv\begin{bmatrix}\hat{\boldsymbol{x}}\\ \hat{\boldsymbol{y}}\end{bmatrix}=\begin{bmatrix}\boldsymbol{x}-\boldsymbol{z}^*\\ \boldsymbol{y}-\boldsymbol{\omega}_{\boldsymbol{y}}^\star\end{bmatrix}$ and also $\hat{\mathscr{F}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})=\hat{\boldsymbol{x}}^\top\mathbf{B}\hat{\boldsymbol{y}}$, with $\hat{\mathscr{F}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ being equal to $\mathscr{F}(\boldsymbol{x},\boldsymbol{y})$ defined as in (20) up to an additive constant. Our scaling-reduction argument hence applies.

Proof.[Proof of Corollary 3.3] From the update rule we have

$$\mathbf{z}_{t-\frac{1}{2}}=\mathbf{z}_{t-1}-\eta\mathbf{J}\mathbf{z}_{t-1}+\eta\boldsymbol{\varepsilon}_{t-\frac{1}{2}}, \tag{46a}$$

$$\mathbf{z}^{\mathrm{ag}}_{t-\frac{1}{2}}=\tfrac{t-1}{t+1}\mathbf{z}^{\mathrm{ag}}_{t-\frac{3}{2}}+\tfrac{2}{t+1}\mathbf{z}_{t-\frac{1}{2}}, \tag{46b}$$

$$\mathbf{z}_t=\mathbf{z}_{t-1}-\eta\mathbf{J}\mathbf{z}_{t-\frac{1}{2}}+\eta\boldsymbol{\varepsilon}_t. \tag{46c}$$

Note the $[\boldsymbol{x}_t^{\mathrm{md}};\boldsymbol{y}_t^{\mathrm{md}}]$ sequence becomes irrelevant in this update; $\mathbf{J}\equiv\begin{bmatrix}\mathbf{0}&\mathbf{B}\\ -\mathbf{B}^\top&\mathbf{0}\end{bmatrix}$ is skew-symmetric with $\mathbf{J}^\top=-\mathbf{J}$, so $\mathbf{J}^2=-\mathbf{J}^\top\mathbf{J}$ is symmetric and negative semidefinite. We proceed with the proof in steps:

Step 1. We target to show the last-iterate bound

$$\mathbb{E}\|\mathbf{z}_t\|^2\le\mathbb{E}\|\mathbf{z}_0\|^2+2t\eta^2\sigma_{\mathrm{Bil}}^2 \tag{47}$$

Note (46a) and (46c) together gives

$$\mathbf{z}_t=\left(\mathbf{I}-\eta\mathbf{J}+\eta^2\mathbf{J}^2\right)\mathbf{z}_{t-1}-\eta^2\mathbf{J}\boldsymbol{\varepsilon}_{t-\frac{1}{2}}+\eta\boldsymbol{\varepsilon}_t \tag{48}$$

Taking squared norm on both sides of (48), we have when $\eta\le\frac{1}{\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$, $\mathbf{z}_t$ does not expand in Euclidean norm (noiseless), so

$$\begin{aligned}\mathbb{E}\|\mathbf{z}_t\|^2&=\mathbb{E}\left[(\mathbf{z}_{t-1})^\top\left(\mathbf{I}+\eta^2\mathbf{J}^2+\eta^4\mathbf{J}^4\right)\mathbf{z}_{t-1}\right]+\mathbb{E}\left\|-\eta^2\mathbf{J}\boldsymbol{\varepsilon}_{t-\frac{1}{2}}+\eta\boldsymbol{\varepsilon}_t\right\|^2\\&\le\mathbb{E}\|\mathbf{z}_{t-1}\|^2+\mathbb{E}\left\|\eta^2\mathbf{J}\boldsymbol{\varepsilon}_{t-\frac{1}{2}}\right\|^2+\mathbb{E}\left\|\eta\boldsymbol{\varepsilon}_t\right\|^2\\&\le\mathbb{E}\|\mathbf{z}_{t-1}\|^2+\eta^2\left(1+\eta^2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})\right)\sigma_{\mathrm{Bil}}^2\le\mathbb{E}\|\mathbf{z}_{t-1}\|^2+2\eta^2\sigma_{\mathrm{Bil}}^2.\end{aligned} \tag{49}$$

Recursively applying the above concludes (47).

Step 2. We start from the update rule (46b) which implies $(t+1)t\mathbf{z}^{\mathrm{ag}}_{t-\frac{1}{2}}=t(t-1)\mathbf{z}^{\mathrm{ag}}_{t-\frac{3}{2}}+2t\mathbf{z}_{t-\frac{1}{2}}$ holds for $t=1,\dots,T$, so

$$(T+1)T\mathbf{z}^{\mathrm{ag}}_{T-\frac{1}{2}}=2\sum_{t=1}^{T}t\mathbf{z}_{t-\frac{1}{2}}\quad\Rightarrow\quad\mathbf{z}^{\mathrm{ag}}_{T-\frac{1}{2}}=\frac{2}{(T+1)T}\sum_{t=1}^{T}t\mathbf{z}_{t-\frac{1}{2}}.$$

Using this to analyze our algorithm:

$$t\mathbf{z}_t-(t-1)\mathbf{z}_{t-1}-\mathbf{z}_{t-1}=t(\mathbf{z}_t-\mathbf{z}_{t-1})=-\eta\mathbf{J}\left[t\mathbf{z}_{t-\frac{1}{2}}\right]+\eta t\boldsymbol{\varepsilon}_t,$$

so telescoping gives

$$T\mathbf{z}_T-\sum_{t=1}^{T}\mathbf{z}_{t-1}=-\eta\mathbf{J}\sum_{t=1}^{T}t\mathbf{z}_{t-\frac{1}{2}}+\eta\sum_{t=1}^{T}t\boldsymbol{\varepsilon}_t,$$

which yields

$$\mathbf{z}_{T-\frac{1}{2}}^{\mathrm{ag}} = \frac{2}{(T+1)T}\sum_{t=1}^{T} t\mathbf{z}_{t-\frac{1}{2}} = \frac{2}{-\eta(T+1)T}\mathbf{J}^{-1}\left(T\mathbf{z}_T - \sum_{t=1}^{T}\mathbf{z}_{t-1} - \eta\sum_{t=1}^{T} t\boldsymbol{\varepsilon}_t\right). \tag{50}$$

Obviously the least singular value of the matrix $\mathbf{J}$ can be lower bounded as $\sigma_{\min}(\mathbf{J}) \geq \sqrt{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$. We conclude from (50) along with Young's inequality that

$$\begin{aligned}
&\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\mathbb{E}\left\|\mathbf{z}_{T-\frac{1}{2}}^{\mathrm{ag}}\right\|^2 \leq \mathbb{E}\left\|\mathbf{J}\mathbf{z}_{T-\frac{1}{2}}^{\mathrm{ag}}\right\|^2 \\
&= (1+\gamma)\frac{4}{\eta^2(T+1)^2T^2}\mathbb{E}\left\|\sum_{t=1}^{T}(\mathbf{z}_T - \mathbf{z}_{t-1})\right\|^2 + (1+\tfrac{1}{\gamma})\frac{4}{\eta^2(T+1)^2T^2}\mathbb{E}\left\|\eta\sum_{t=1}^{T} t\boldsymbol{\varepsilon}_t\right\|^2 \\
&\equiv (1+\gamma)\mathrm{I} + (1+\tfrac{1}{\gamma})\mathrm{II},
\end{aligned}$$

where applying the last-iterate bound (47) together with some elementary estimates leads to

$$\begin{aligned}
\mathrm{I} &\leq \frac{4}{\eta^2(T+1)^2T^2}\cdot T\sum_{t=1}^{T}\left[2\mathbb{E}\left\|\mathbf{z}_T\right\|^2 + 2\mathbb{E}\left\|\mathbf{z}_{t-1}\right\|^2\right] \\
&\leq \frac{4}{\eta^2(T+1)^2T^2}\cdot T\sum_{t=1}^{T}\left[4\mathbb{E}\|\mathbf{z}_0\|^2 + 4(T+t-1)\eta^2\sigma_{\mathrm{Bil}}^2\right] \\
&\leq \frac{16\mathbb{E}\|\mathbf{z}_0\|^2 + 24\eta^2\sigma_{\mathrm{Bil}}^2 T}{\eta^2(T+1)^2} \leq \frac{16\lambda_{\max}(\mathbf{B}^\top\mathbf{B})\mathbb{E}\|\mathbf{z}_0\|^2}{(T+1)^2} + \frac{24\sigma_{\mathrm{Bil}}^2}{T+1},
\end{aligned}$$

and, using the property of square-integrable martingales,

$$\begin{aligned}
\mathrm{II} &\leq \frac{4}{\eta^2(T+1)^2T^2}\mathbb{E}\left\|\eta\sum_{t=1}^{T} t\boldsymbol{\varepsilon}_t\right\|^2 = \frac{4}{\eta^2(T+1)^2T^2}\cdot\eta^2\sum_{t=1}^{T} t^2\mathbb{E}\left\|\boldsymbol{\varepsilon}_t\right\|^2 \\
&\leq \frac{4\sigma_{\mathrm{Bil}}^2}{\eta^2(T+1)^2T^2}\cdot\eta^2\frac{T(T+\frac{1}{2})(T+1)}{3} \leq \frac{4\sigma_{\mathrm{Bil}}^2}{3T}.
\end{aligned}$$

To summarize we have for arbitrary $\gamma \in (0,\infty)$

$$\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\mathbb{E}\left\|\mathbf{z}_{T-\frac{1}{2}}^{\mathrm{ag}}\right\|^2 \leq (1+\gamma)\left(\frac{16\lambda_{\max}(\mathbf{B}^\top\mathbf{B})\mathbb{E}\|\mathbf{z}_0\|^2}{(T+1)^2} + \frac{24\sigma_{\mathrm{Bil}}^2}{T+1}\right) + (1+\tfrac{1}{\gamma})\frac{4\sigma_{\mathrm{Bil}}^2}{3T}.$$

Optimizing γ gives along with $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for nonnegatives a and b:

$$\begin{aligned}
&\sqrt{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\sqrt{\mathbb{E}\left\|\mathbf{z}_{T-\frac{1}{2}}^{\mathrm{ag}}\right\|^2} \leq \sqrt{\frac{16\lambda_{\max}(\mathbf{B}^\top\mathbf{B})\mathbb{E}\|\mathbf{z}_0\|^2}{(T+1)^2} + \frac{24\sigma_{\mathrm{Bil}}^2}{T+1}} + \sqrt{\frac{4\sigma_{\mathrm{Bil}}^2}{3T}} \\
&\leq \sqrt{\frac{16\lambda_{\max}(\mathbf{B}^\top\mathbf{B})\mathbb{E}\|\mathbf{z}_0\|^2}{(T+1)^2}} + \sqrt{\frac{24\sigma_{\mathrm{Bil}}^2}{T+1}} + \sqrt{\frac{4\sigma_{\mathrm{Bil}}^2}{3T}} \leq \frac{4\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}{T+1}\sqrt{\mathbb{E}\|\mathbf{z}_0\|^2} + \frac{7\sigma_{\mathrm{Bil}}}{\sqrt{T}}.
\end{aligned}$$

Dividing both sides by $\sqrt{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$ and taking squares conclude the result.

E Proof of Auxiliary Lemmas

E.1 Proof of Lemma D.1

The analysis in this subsection is partially motivated by Lemma 2 of Chen et al. [2017].

Proof.[Proof of Lemma D.1] We first introduce the following lemma on the operator $\mathcal{P}$:

Lemma E.1 (Lemma 2 in Ghadimi and Lan 2012 and Lemma 1 in Chen et al. 2017) *If $\boldsymbol{\phi} = \mathcal{P}_{\boldsymbol{\theta}}(\boldsymbol{\delta})$ for arbitrarily chosen $\boldsymbol{\theta}, \boldsymbol{\delta} \in \mathbb{R}^d$, then for $\forall \boldsymbol{z} \in \mathcal{Z}$, we have the following inequality*

$$\langle \boldsymbol{\delta}, \boldsymbol{\phi} - \boldsymbol{z}\rangle + J(\boldsymbol{\phi}) - J(\boldsymbol{z}) \leq \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{z}) - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\phi}) - V(\boldsymbol{\phi}, \boldsymbol{z})$$

By applying Lemma E.1 to (25), we have for any $\mathbf{z} \in \mathcal{Z}$

$$\langle \boldsymbol{\delta}_1, \boldsymbol{\varphi}_1 - \mathbf{z} \rangle + J(\boldsymbol{\varphi}_1) - J(\boldsymbol{z}) \leq \mathcal{B}(\boldsymbol{\theta}, \mathbf{z}) - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\varphi}_1) - \mathcal{B}(\boldsymbol{\varphi}_1, \mathbf{z}), \tag{51}$$

$$\langle \boldsymbol{\delta}_2, \boldsymbol{\varphi}_2 - \mathbf{z} \rangle + J(\boldsymbol{\varphi}_2) - J(\boldsymbol{z}) \leq \mathcal{B}(\boldsymbol{\theta}, \mathbf{z}) - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\varphi}_2) - \mathcal{B}(\boldsymbol{\varphi}_2, \mathbf{z}). \tag{52}$$

Specifically, letting $\mathbf{z} = \boldsymbol{\varphi}_2$ in (51) we have

$$\langle \boldsymbol{\delta}_1, \boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2 \rangle + J(\boldsymbol{\varphi}_1) - J(\boldsymbol{\varphi}_2) = \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\varphi}_2) - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\varphi}_1) - \mathcal{B}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2). \tag{53}$$

Now, combining inequalities (52) and (53) we have

$$\langle \boldsymbol{\delta}_2, \boldsymbol{\varphi}_2 - \mathbf{z} \rangle + \langle \boldsymbol{\delta}_1, \boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2 \rangle + J(\boldsymbol{\varphi}_1) - J(\boldsymbol{z}) \leq \mathcal{B}(\boldsymbol{\theta}, \mathbf{z}) - \mathcal{B}(\boldsymbol{\varphi}_2, \mathbf{z}) - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\varphi}_1) - \mathcal{B}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2),$$

which in turn gives

$$\begin{aligned}
&\langle \boldsymbol{\delta}_2, \boldsymbol{\varphi}_1 - \mathbf{z} \rangle + J(\boldsymbol{\varphi}_1) - J(\boldsymbol{z}) \\
&\leq \langle \boldsymbol{\delta}_2 - \boldsymbol{\delta}_1, \boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2 \rangle + \mathcal{B}(\boldsymbol{\theta}, \mathbf{z}) - \mathcal{B}(\boldsymbol{\varphi}_2, \mathbf{z}) - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\varphi}_1) - \mathcal{B}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2).
\end{aligned}$$

An application of the Young and Cauchy-Schwartz inequalities gives

$$\begin{aligned}
&\langle \boldsymbol{\delta}_2, \boldsymbol{\varphi}_1 - \mathbf{z} \rangle + J(\boldsymbol{\varphi}_1) - J(\boldsymbol{z}) \\
&\leq \|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\| \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\| + \mathcal{B}(\boldsymbol{\theta}, \mathbf{z}) - \mathcal{B}(\boldsymbol{\varphi}_2, \mathbf{z}) - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\varphi}_1) - \mathcal{B}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \\
&\leq \frac{1}{2\mu_{\mathcal{B}}} \|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\|^2 + \frac{\mu_{\mathcal{B}}}{2} \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|^2 \\
&\quad + \mathcal{B}(\boldsymbol{\theta}, \mathbf{z}) - \mathcal{B}(\boldsymbol{\varphi}_2, \mathbf{z}) - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\varphi}_1) - \mathcal{B}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2) \\
&\leq \frac{1}{2\mu_{\mathcal{B}}} \|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\|^2 + \mathcal{B}(\boldsymbol{\theta}, \mathbf{z}) - \mathcal{B}(\boldsymbol{\varphi}_2, \mathbf{z}) - \mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\varphi}_1).
\end{aligned} \tag{54}$$

In the last inequality, we uses the fact that

$$\frac{\mu_{\mathcal{B}}}{2} \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|^2 \leq \mathcal{B}(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2).$$

This establishes (27) and hence Lemma D.1.

E.2 Proof of Lemma D.2

Proof.[Proof of Lemma D.2]

Since $\mathcal{F}(\boldsymbol{z})$ is L-smooth and μ-strongly convex. For the rest of this proof, we observe that the saddle definition of $\boldsymbol{z}^*$ satisfies the first-order stationary condition for problem (3):

$$\nabla \mathcal{F}(\boldsymbol{z}^*) + \mathcal{H}(\boldsymbol{z}^*) + J'(\boldsymbol{z}^*) = 0. \tag{55}$$

Furthermore, we have

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

where in both of the two displays, the inequality holds due to the μ-strong convexity of $\mathcal{F}$, and the equality holds due to the first-order stationary condition (55). This completes the proof.

E.3 Proof of Lemma D.3

Proof.[Proof of Lemma D.3] Items (i)—(iii) are straightforward. For the proof of (30) in item (iv), we note that $\eta_t = \bar{\eta}_t(\sigma; T, C, r, \beta) \leq \frac{t}{\frac{2}{r}L + \sqrt{\frac{1+\beta}{r}} Mt} \leq \frac{1}{\sqrt{\frac{1+\beta}{r}} M}$ which gives

$$r - \frac{2L}{t+1}\eta_t - (1+\beta)M^2\eta_t^2 \geq \frac{r}{t}\left(t - \left(\frac{2}{r}L + \sqrt{\tfrac{1+\beta}{r}}Mt\right)\eta_t\right) \geq 0,$$

and hence completes the proof.

E.4 Proof of Lemma D.4

Proof.[Proof of Lemma D.4] From the convexity and L-smoothness of F as in Assumption 2.1, we know that for arbitrary $\tilde{\boldsymbol{z}}$:

$$\begin{aligned}
&\mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}) - \mathcal{F}(\tilde{\boldsymbol{z}}) = \mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}) - \mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}) - \big(\mathcal{F}(\tilde{\boldsymbol{z}}) - \mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1})\big) \\
&\leq \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\rangle + \tfrac{L}{2}\left\|\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\right\|^2 - \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \tilde{\boldsymbol{z}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\rangle.
\end{aligned}$$

Taking $\tilde{\boldsymbol{z}} = \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}}$ in the above inequality, we have

$$\begin{aligned}
&\mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}) - \mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}}) = \mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}) - \mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}) - \Big(\mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}}) - \mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1})\Big) \\
&\leq \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\rangle + \tfrac{L}{2}\left\|\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\right\|^2 - \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\rangle.
\end{aligned}$$

Multiplying the first display by α_t and the second display by $(1-\alpha_t)$ and adding them up, we have

$$\begin{aligned}
&\mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}) - (1-\alpha_t)\mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}}) - \alpha_t\mathcal{F}(\tilde{\boldsymbol{z}}) \\
&\leq \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\rangle + \tfrac{L}{2}\left\|\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\right\|^2 - \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), (1-\alpha_t)\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} + \alpha_t\tilde{\boldsymbol{z}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\rangle \\
&\leq \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \alpha_t(\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1})\rangle + \tfrac{L}{2}\|\alpha_t(\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1})\|^2 - \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \alpha_t(\tilde{\boldsymbol{z}} - \boldsymbol{z}_{t-1})\rangle \\
&= \alpha_t\langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle + \tfrac{\alpha_t^2 L}{2}\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\|^2,
\end{aligned} \tag{56}$$

where we applied the fact from our update rules that $\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1} = \alpha_t(\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1})$.

On the other hand, due to Line (6) in Algorithm 1 we have

$$\begin{aligned}
\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle - (1-\alpha_t)\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\boldsymbol{z}}\rangle &= \langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} - (1-\alpha_t)(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\boldsymbol{z}})\rangle \\
&= \alpha_t\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle.
\end{aligned}$$

Further, due to our monotonicity assumption on $\mathcal{H}$ we have

$$\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle \leq \langle \mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle.$$

Combining the above two displays together yields

$$\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle - (1-\alpha_t)\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\boldsymbol{z}}\rangle \leq \alpha_t\langle \mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle. \tag{57}$$

Now, summing up Eqs. (56), (57) and recalling the definition of V in (28), we conclude that

$$\begin{aligned}
&V(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} \mid \tilde{\boldsymbol{z}}) - (1-\alpha_t)V(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} \mid \tilde{\boldsymbol{z}}) \\
&= \mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}) - (1-\alpha_t)\mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}}) - \alpha_t\mathcal{F}(\tilde{\boldsymbol{z}}) + \langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle - (1-\alpha_t)\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\boldsymbol{z}}\rangle \\
&\leq \alpha_t\langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}) + \mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle + \tfrac{\alpha_t^2 L}{2}\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\|^2,
\end{aligned}$$

and hence conclude (32) and Lemma D.4.

E.5 Proof of Lemma D.5

Proof.[Proof of Lemma D.5] To bound the inner-product terms in (32), by setting $\boldsymbol{\varphi}_1 = \boldsymbol{z}_{t-\frac{1}{2}}$, $\boldsymbol{\theta} = \boldsymbol{z}_{t-1}$, $\boldsymbol{\varphi}_2 = \boldsymbol{z}_t$, $\boldsymbol{\delta}_1 = \eta_t\Big(\nabla\tilde{\mathcal{F}}(\boldsymbol{z}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1}; \zeta_{t-\frac{1}{2}})\Big)$, $\boldsymbol{\delta}_2 = \eta_t\Big(\nabla\tilde{\mathcal{F}}(\boldsymbol{z}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t)\Big)$ as in Lemma D.1 (with $\mathbf{z} = \tilde{\boldsymbol{z}}$), we have

$$\begin{aligned}
&\eta_t\langle \nabla\tilde{\mathcal{F}}(\boldsymbol{z}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle \\
&\leq \frac{1}{2}\left[\|\boldsymbol{z}_{t-1} - \tilde{\boldsymbol{z}}\|^2 - \|\mathbf{x}_t - \tilde{\boldsymbol{z}}\|^2 - \left\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\right\|^2\right] + \frac{\eta_t^2}{2}\|\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t) - \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1}; \zeta_{t-\frac{1}{2}})\|^2,
\end{aligned}$$

where Young's inequality combined with the martingale structure yields (also noting (31))

$$\begin{aligned}
&\mathbb{E}\|\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t)-\tilde{\mathcal{H}}(\boldsymbol{z}_{t-1};\zeta_{t-\frac{1}{2}})\|^2\\
&=\mathbb{E}\|\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}})-\mathcal{H}(\boldsymbol{z}_{t-1})-\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\|^2+\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2\\
&\leq(1+\beta)M^2\mathbb{E}\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\|^2+(1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2+\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2.
\end{aligned}$$

Combining the above two displays with expectation taken gives

$$\begin{aligned}
&\mathbb{E}\left[\eta_t\langle\nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}};\xi_{t-\frac{1}{2}})+\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\rangle\right]\\
&\leq\frac{1}{2}\mathbb{E}\left[\|\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\|^2-\|\mathbf{x}_t-\tilde{\boldsymbol{z}}\|^2-\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2\right]\\
&\quad+\frac{\eta_t^2}{2}\left((1+\beta)M^2\mathbb{E}\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2+(1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2+\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2\right).
\end{aligned}\tag{58}$$

Further, by definition of the primal-dual gap function and the definition of the noisy terms (31), by taking $\alpha_t=\frac{2}{t+1}$ in (32) of Lemma D.4 and taking expectations on both sides, we have

$$\begin{aligned}
&\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]-\frac{t-1}{t+1}\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]\\
&\leq\frac{2}{t+1}\mathbb{E}\Big\langle\nabla\mathcal{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}})+\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\Big\rangle+\frac{2L}{(t+1)^2}\mathbb{E}\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2\\
&=\frac{2}{t+1}\mathbb{E}\Big\langle\nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}};\xi_{t-\frac{1}{2}})+\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\Big\rangle+\frac{2L}{(t+1)^2}\mathbb{E}\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2\\
&\quad-\frac{2}{t+1}\mathbb{E}\Big\langle\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}+\boldsymbol{\Delta}_{\mathcal{H}}^{t},\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\Big\rangle
\end{aligned}$$

Bringing in (58) into the above derivation, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]-\frac{t-1}{t+1}\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]\\
&\leq\frac{1}{(t+1)\eta_t}\mathbb{E}\big[\|\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\|^2-\|\boldsymbol{z}_t-\tilde{\boldsymbol{z}}\|^2\big]\\
&\quad-\frac{1}{(t+1)\eta_t}\left(1-\frac{2L}{t+1}\eta_t-(1+\beta)M^2\eta_t^2\right)\mathbb{E}\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2\\
&\quad+\frac{\eta_t}{t+1}\left((1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2+\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2\right)-\frac{2}{t+1}\mathbb{E}\Big\langle\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}+\boldsymbol{\Delta}_{\mathcal{H}}^{t},\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\Big\rangle.
\end{aligned}$$

Recalling that we use the choice of η_t that satisfies for a given $r\in(0,1)$ that $r-\frac{2L}{t+1}\eta_t-(1+\beta)M^2\eta_t^2\geq 0$. With some manipulations we obtain

$$\begin{aligned}
&\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]-\frac{t-1}{t+1}\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]\\
&\leq\frac{1}{(t+1)\eta_t}\mathbb{E}\big[\|\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\|^2-\|\boldsymbol{z}_t-\tilde{\boldsymbol{z}}\|^2\big]\\
&\quad+\frac{\eta_t}{(t+1)}\left((1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2+\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2\right)-\frac{(1-r)}{(t+1)\eta_t}\mathbb{E}\left[\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2\right]\\
&\quad-\frac{2}{t+1}\mathbb{E}\Big\langle\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}},\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\Big\rangle\underbrace{-\frac{2}{t+1}\mathbb{E}\Big\langle\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}},\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\Big\rangle-\frac{2}{t+1}\mathbb{E}\Big\langle\boldsymbol{\Delta}_{\mathcal{H}}^{t},\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\Big\rangle}_{\mathrm{I}}.
\end{aligned}\tag{59}$$

Due to the law of iterated expectation applied to martingale difference conditions $\mathbb{E}\Big[\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}\mid\mathcal{F}_{t-1}\Big]=\mathbf{0}$ and $\mathbb{E}\Big[\boldsymbol{\Delta}_{\mathcal{H}}^{t}\mid\mathcal{F}_{t-\frac{1}{2}}\Big]=\mathbf{0}$, $i=1,2$, we have

$$I=0.$$

Moreover, for the rest of the terms in (59), we note that there is a basic quadratic inequality that $-\frac{1-r}{\eta_t}\left\|\boldsymbol{z}_{t-1}-\mathbf{x}_{t-\frac{1}{2}}\right\|^2-2\left\langle\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}},\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\rangle\le\frac{\eta_t}{1-r}\left\|\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}\right\|^2$. (59) reduces to

$$\begin{aligned}\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]-\frac{t-1}{t+1}\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]&\le\frac{1}{(t+1)\eta_t}\mathbb{E}\big[\|\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\|^2-\|\boldsymbol{z}_t-\tilde{\boldsymbol{z}}\|^2\big]\\&+\frac{\eta_t}{t+1}\left((1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2+\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2\right)+\frac{\eta_t}{(1-r)(t+1)}\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}\right\|^2.\end{aligned}\tag{60}$$

Multiplying both sides of (60) by $t(t+1)$, we obtain for all $t=1,\ldots,T$

$$\begin{aligned}&t(t+1)\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]-(t-1)t\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]\\&\le\frac{t}{\eta_t}\mathbb{E}\big[\|\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\|-\|\boldsymbol{z}_t-\tilde{\boldsymbol{z}}\|^2\big]+t\eta_t\left(\tfrac{1}{1-r}\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}\right\|^2+(1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2+\mathbb{E}\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\|^2\right)\\&\le\frac{t}{\eta_t}\mathbb{E}\big[\|\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\|^2-\|\boldsymbol{z}_t-\tilde{\boldsymbol{z}}\|^2\big]+\left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2+(2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right)t\eta_t,\end{aligned}$$

where in the last line above we applied Assumption 2.2, so by law of iterated expectations

$$\begin{aligned}\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}\right\|^2&=\mathbb{E}\left[\|\nabla\tilde{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}};\xi_{t-\frac{1}{2}})-\nabla F(\boldsymbol{z}_{t-1}^{\mathrm{md}})\|^2\right]\le\sigma_{\mathrm{Str}}^2,\\\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2&=\mathbb{E}\left[\|\tilde{H}(\boldsymbol{z}_{t-1};\zeta_{t-\frac{1}{2}})-H(\boldsymbol{z}_{t-1})\|^2\right]\le\sigma_{\mathrm{Bil}}^2,\\\mathbb{E}\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\|^2&=\mathbb{E}\left[\|\tilde{H}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t)-H(\boldsymbol{z}_{t-\frac{1}{2}})\|^2\right]\le\sigma_{\mathrm{Bil}}^2.\end{aligned}\tag{61}$$

E.6 Proof of Lemma D.6

Proof.[Proof of Lemma D.6] The proof of Lemma D.6 goes in an analogous fashion as the proof of Lemma D.4, except that the display above (57) is replaced by

$$\langle\mathcal{H}(\tilde{\boldsymbol{z}}),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\rangle\le\langle\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\rangle-\mu_\star\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\right\|^2,$$

due to that H is a $\mu_\star$-strongly-convex-$\mu_\star$-strongly-concave isotropic quadratic function after scaling reduction. Hence (57) becomes

$$\langle\mathcal{H}(\tilde{\boldsymbol{z}}),\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}-\tilde{\boldsymbol{z}}\rangle-(1-\alpha_t)\langle\mathcal{H}(\tilde{\boldsymbol{z}}),\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}-\tilde{\boldsymbol{z}}\rangle\le\alpha_t\left[\langle\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\rangle-\mu_\star\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\right\|^2\right].\tag{62}$$

Therefore, we omit its detailed proof.

E.7 Proof of Lemma D.7

Proof.[Proof of Lemma D.7] From the convexity and L-smoothness of F as in Assumption 2.1, we know that for arbitrary $\tilde{\boldsymbol{z}}$:

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

Taking $\tilde{\boldsymbol{z}}=\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}$ in the above inequality, we have

$$\begin{aligned}&\mathcal{F}(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}})-\mathcal{F}(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}})=\mathcal{F}(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}})-\mathcal{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}})-\Big(\mathcal{F}(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}})-\mathcal{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}})\Big)\\&\le\langle\nabla\mathcal{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}}),\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}-\boldsymbol{z}_{t-1}^{\mathrm{md}}\rangle+\tfrac{L}{2}\left\|\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}-\boldsymbol{z}_{t-1}^{\mathrm{md}}\right\|^2-\langle\nabla\mathcal{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}}),\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}-\boldsymbol{z}_{t-1}^{\mathrm{md}}\rangle.\end{aligned}$$

Multiplying the first display by α_t and the second display by $(1-\alpha_t)$ and adding them up, we have

$$\begin{aligned}
&\mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}) - (1-\alpha_t)\mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}}) - \alpha_t\mathcal{F}(\tilde{\boldsymbol{z}}) \\
&\le \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\rangle + \tfrac{L}{2}\left\|\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\right\|^2 - \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), (1-\alpha_t)\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} + \alpha_t\tilde{\boldsymbol{z}} - \boldsymbol{z}^{\mathrm{md}}_{t-1}\rangle \\
&\le \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \alpha_t(\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1})\rangle + \tfrac{L}{2}\|\alpha_t(\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1})\|^2 - \langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \alpha_t(\tilde{\boldsymbol{z}} - \boldsymbol{z}_{t-1})\rangle \\
&= \alpha_t\langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle + \tfrac{\alpha_t^2 L}{2}\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\|^2,
\end{aligned} \tag{63}$$

where we applied the fact from our update rules that $\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \boldsymbol{z}^{\mathrm{md}}_{t-1} = \alpha_t(\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1})$.

On the other hand, due to Line (6) in Algorithm 3 we have

$$\begin{aligned}
\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle - (1-\alpha_t)\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\boldsymbol{z}}\rangle &= \langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}} - (1-\alpha_t)(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\boldsymbol{z}})\rangle \\
&= \alpha_t\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle.
\end{aligned}$$

Further, due to our monotonicity assumption on $\mathcal{H}$ we have

$$\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle \le \langle \mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle.$$

Combining the above two displays together yields

$$\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle - (1-\alpha_t)\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\boldsymbol{z}}\rangle \le \alpha_t\langle \mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle. \tag{64}$$

Moreover, we have

$$\begin{aligned}
J(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}) - J(\tilde{\boldsymbol{z}}) - (1-\alpha_t)\left(J(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}}) - J(\tilde{\boldsymbol{z}})\right) &= J(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}) - (1-\alpha_t)J(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}}) - \alpha_t J(\tilde{\boldsymbol{z}}) \\
&\le \alpha_t J(\boldsymbol{z}_{t-\frac{1}{2}}) - \alpha_t J(\tilde{\boldsymbol{z}})
\end{aligned} \tag{65}$$

Now, summing up Eqs. (63), (64) and recalling the definition of V, we conclude that

$$\begin{aligned}
&V(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} \mid \tilde{\boldsymbol{z}}) - (1-\alpha_t)V(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} \mid \tilde{\boldsymbol{z}}) \\
&= \mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}}) - (1-\alpha_t)\mathcal{F}(\boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}}) - \alpha_t\mathcal{F}(\tilde{\boldsymbol{z}}) + \langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle - (1-\alpha_t)\langle \mathcal{H}(\tilde{\boldsymbol{z}}), \boldsymbol{z}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\boldsymbol{z}}\rangle \\
&\le \alpha_t\langle \nabla\mathcal{F}(\boldsymbol{z}^{\mathrm{md}}_{t-1}) + \mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle + \tfrac{\alpha_t^2 L}{2}\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\|^2,
\end{aligned}$$

and hence conclude (41) and Lemma D.7.

E.8 Proof of Lemma D.8

Proof.[Proof of Lemma D.8] To bound the inner-product terms in (41), by setting $\boldsymbol{\varphi}_1 = \boldsymbol{z}_{t-\frac{1}{2}}$, $\boldsymbol{\theta} = \boldsymbol{z}_{t-1}$, $\boldsymbol{\varphi}_2 = \boldsymbol{z}_t$, $\boldsymbol{\delta}_1 = \eta_t\left(\nabla\tilde{\mathcal{F}}(\boldsymbol{z}^{\mathrm{md}}_{t-1};\xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1};\zeta_{t-\frac{1}{2}})\right)$, $\boldsymbol{\delta}_2 = \eta_t\Big(\nabla\tilde{\mathcal{F}}(\boldsymbol{z}^{\mathrm{md}}_{t-1};\xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t)\Big)$ and $J = \eta_t J$ as in Lemma D.1 (with $\mathbf{z} = \tilde{\boldsymbol{z}}$), we have

$$\begin{aligned}
&\eta_t\langle \nabla\tilde{\mathcal{F}}(\boldsymbol{z}^{\mathrm{md}}_{t-1};\xi_{t-\frac{1}{2}}) + \tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t), \boldsymbol{z}_{t-\frac{1}{2}} - \tilde{\boldsymbol{z}}\rangle + \eta_t\left(J(\boldsymbol{z}_{t-\frac{1}{2}}) - J(\tilde{\boldsymbol{z}})\right) \\
&\le \mathcal{B}(\boldsymbol{z}_{t-1}, \tilde{\boldsymbol{z}}) - \mathcal{B}(\mathbf{x}_t, \tilde{\boldsymbol{z}}) - \mathcal{B}(\boldsymbol{z}_{t-\frac{1}{2}}, \boldsymbol{z}_{t-1}) + \frac{\eta_t^2}{2\mu_{\mathcal{B}}}\|\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t) - \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1};\zeta_{t-\frac{1}{2}})\|^2,
\end{aligned}$$

where Young's inequality combined with the martingale structure yields (also noting (31))

$$\begin{aligned}
&\mathbb{E}\|\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t) - \tilde{\mathcal{H}}(\boldsymbol{z}_{t-1};\zeta_{t-\frac{1}{2}})\|^2 \\
&= \mathbb{E}\|\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}) - \mathcal{H}(\boldsymbol{z}_{t-1}) - \boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\|^2 + \mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2 \\
&\le (1+\beta)M^2\mathbb{E}\|\boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\|^2 + (1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2 + \mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2.
\end{aligned}$$

Combining the above two displays with expectation taken gives

$$\begin{aligned}
&\mathbb{E}\left[\eta_t\langle\nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}};\xi_{t-\frac{1}{2}})+\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\rangle+\eta_t\left(J(\boldsymbol{z}_{t-\frac{1}{2}})-J(\tilde{\boldsymbol{z}})\right)\right]\\
&\leq\mathbb{E}\left[\mathcal{B}(\boldsymbol{z}_{t-1},\tilde{\boldsymbol{z}})-\mathcal{B}(\mathbf{x}_t,\tilde{\boldsymbol{z}})-\mathcal{B}(\boldsymbol{z}_{t-\frac{1}{2}},\boldsymbol{z}_{t-1})\right]\\
&\quad+\frac{\eta_t^2}{2\mu_{\mathcal{B}}}\left((1+\beta)M^2\mathbb{E}\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\|^2+(1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2+\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2\right).
\end{aligned}\tag{66}$$

Applying the inequality $\mathcal{B}(\boldsymbol{z}_{t-\frac{1}{2}},\boldsymbol{z}_{t-1})\geq\frac{\mu_{\mathcal{B}}}{2}\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2$ again gives

$$\begin{aligned}
&\mathbb{E}\left[\eta_t\langle\nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}};\xi_{t-\frac{1}{2}})+\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\rangle+\eta_t\left(J(\boldsymbol{z}_{t-\frac{1}{2}})-J(\tilde{\boldsymbol{z}})\right)\right]\\
&\leq\mathbb{E}\left[\mathcal{B}(\boldsymbol{z}_{t-1},\tilde{\boldsymbol{z}})-\mathcal{B}(\mathbf{x}_t,\tilde{\boldsymbol{z}})\right]\\
&\quad-\left(\frac{\mu_{\mathcal{B}}}{2}-\frac{\eta_t^2}{2\mu_{\mathcal{B}}}(1+\beta)M^2\right)\mathbb{E}\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\|^2+\frac{\eta_t^2}{2\mu_{\mathcal{B}}}\left((1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2+\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2\right).
\end{aligned}$$

Further, by definition of the primal-dual gap function and the definition of the noisy terms (31), by taking $\alpha_t=\frac{2}{t+1}$ in (41) of Lemma D.7 and taking expectations on both sides, we have

$$\begin{aligned}
&\mathbb{E}[V(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}})]-\frac{t-1}{t+1}\mathbb{E}[V(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}})]\\
&\leq\frac{2}{t+1}\mathbb{E}\langle\nabla\mathcal{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}})+\mathcal{H}(\boldsymbol{z}_{t-\frac{1}{2}}),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\rangle+\frac{2L}{(t+1)^2}\mathbb{E}\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2\\
&\quad+\frac{2}{t+1}\mathbb{E}\left(J(\boldsymbol{z}_{t-\frac{1}{2}})-J(\tilde{\boldsymbol{z}})\right)\\
&=\frac{2}{t+1}\mathbb{E}\langle\nabla\tilde{\mathcal{F}}(\boldsymbol{z}_{t-1}^{\mathrm{md}};\xi_{t-\frac{1}{2}})+\tilde{\mathcal{H}}(\boldsymbol{z}_{t-\frac{1}{2}};\zeta_t),\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\rangle+\frac{2L}{(t+1)^2}\mathbb{E}\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2\\
&\quad-\frac{2}{t+1}\mathbb{E}\langle\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}+\boldsymbol{\Delta}_{\mathcal{H}}^{t},\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\rangle+\frac{2}{t+1}\mathbb{E}\left(J(\boldsymbol{z}_{t-\frac{1}{2}})-J(\tilde{\boldsymbol{z}})\right)
\end{aligned}$$

Bringing in (66) into the above derivation, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]-\frac{t-1}{t+1}\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]\\
&\leq\frac{2}{(t+1)\eta_t}\left(\mathbb{E}\Big[\mathcal{B}(\boldsymbol{z}_{t-1},\tilde{\boldsymbol{z}})\Big]-\mathbb{E}\Big[\mathcal{B}(\boldsymbol{z}_t,\tilde{\boldsymbol{z}})\Big]\right)\\
&\quad-\frac{1}{(t+1)\eta_t}\left(\mu_{\mathcal{B}}-\frac{2L}{t+1}\eta_t-\frac{(1+\beta)M^2\eta_t^2}{\mu_{\mathcal{B}}}\right)\mathbb{E}\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2\\
&\quad+\frac{\eta_t}{(t+1)\mu_{\mathcal{B}}}\left((1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2+\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2\right)-\frac{2}{t+1}\mathbb{E}\Big\langle\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}+\boldsymbol{\Delta}_{\mathcal{H}}^{t},\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\Big\rangle.
\end{aligned}$$

Recalling that we use the choice of η_t that satisfies for a given $r\in(0,1)$ that $r\mu_{\mathcal{B}}-\frac{2L}{t+1}\eta_t-\frac{(1+\beta)M^2\eta_t^2}{\mu_{\mathcal{B}}}\geq 0$. With some manipulations we obtain

$$\begin{aligned}
&\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]-\frac{t-1}{t+1}\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}}\mid\tilde{\boldsymbol{z}}\big)\Big]\\
&\leq\frac{2}{(t+1)\eta_t}\left(\mathbb{E}\Big[\mathcal{B}\big(\boldsymbol{z}_{t-1},\tilde{\boldsymbol{z}}\big)\Big]-\mathbb{E}\Big[\mathcal{B}\big(\boldsymbol{z}_t,\tilde{\boldsymbol{z}}\big)\Big]\right)\\
&\quad+\frac{\eta_t}{(t+1)\mu_{\mathcal{B}}}\left((1+\tfrac{1}{\beta})\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\right\|^2+\mathbb{E}\left\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\right\|^2\right)-\frac{(1-r)\mu_{\mathcal{B}}}{(t+1)\eta_t}\mathbb{E}\left[\left\|\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\right\|^2\right]\\
&\quad-\frac{2}{t+1}\mathbb{E}\Big\langle\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}},\boldsymbol{z}_{t-\frac{1}{2}}-\boldsymbol{z}_{t-1}\Big\rangle\underbrace{-\frac{2}{t+1}\mathbb{E}\Big\langle\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}},\boldsymbol{z}_{t-1}-\tilde{\boldsymbol{z}}\Big\rangle-\frac{2}{t+1}\mathbb{E}\Big\langle\boldsymbol{\Delta}_{\mathcal{H}}^{t},\boldsymbol{z}_{t-\frac{1}{2}}-\tilde{\boldsymbol{z}}\Big\rangle}_{\mathrm{I}}.
\end{aligned}\tag{67}$$

Due to the law of iterated expectation applied to martingale difference conditions $\mathbb{E}\Big[\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}} \mid \mathcal{F}_{t-1}\Big] = \mathbf{0}$ and $\mathbb{E}\Big[\boldsymbol{\Delta}_{\mathcal{H}}^{t} \mid \mathcal{F}_{t-\frac{1}{2}}\Big] = \mathbf{0}$, $i = 1, 2$, we have

$$I = 0.$$

Moreover, for the rest of the terms in (67), we note that there is a basic quadratic inequality that $-\frac{(1-r)\mu_{\mathcal{B}}}{\eta_t}\Big\|\boldsymbol{z}_{t-1} - \mathbf{x}_{t-\frac{1}{2}}\Big\|^2 - 2\Big\langle \boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}, \boldsymbol{z}_{t-\frac{1}{2}} - \boldsymbol{z}_{t-1}\Big\rangle \leq \frac{\eta_t}{(1-r)\mu_{\mathcal{B}}}\Big\|\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}\Big\|^2$. (67) reduces to

$$\begin{aligned}
&\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}}\big)\Big] - \frac{t-1}{t+1}\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}}\big)\Big] \leq \frac{2}{(t+1)\eta_t}\Big(\mathbb{E}\Big[\mathcal{B}(\boldsymbol{z}_{t-1}, \tilde{\boldsymbol{z}})\Big] - \mathbb{E}\Big[\mathcal{B}(\boldsymbol{z}_t, \tilde{\boldsymbol{z}})\Big]\Big) \\
&\quad + \frac{\eta_t}{(t+1)\mu_{\mathcal{B}}}\left((1+\tfrac{1}{\beta})\mathbb{E}\Big\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\Big\|^2 + \mathbb{E}\Big\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\Big\|^2\right) + \frac{\eta_t}{(1-r)(t+1)\mu_{\mathcal{B}}}\mathbb{E}\Big\|\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}\Big\|^2. \qquad (68)
\end{aligned}$$

Multiplying both sides of (68) by $t(t+1)$, we obtain for all $t = 1, \ldots, T$

$$\begin{aligned}
&t(t+1)\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{1}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}}\big)\Big] - (t-1)t\mathbb{E}\Big[V\big(\boldsymbol{z}_{t-\frac{3}{2}}^{\mathrm{ag}} \mid \tilde{\boldsymbol{z}}\big)\Big] \\
&\leq \frac{2t}{\eta_t}\Big(\mathbb{E}\Big[\mathcal{B}(\boldsymbol{z}_{t-1}, \tilde{\boldsymbol{z}})\Big] - \mathbb{E}\Big[\mathcal{B}(\boldsymbol{z}_t, \tilde{\boldsymbol{z}})\Big]\Big) + \frac{t\eta_t}{\mu_{\mathcal{B}}}\left(\tfrac{1}{1-r}\mathbb{E}\Big\|\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}\Big\|^2 + (1+\tfrac{1}{\beta})\mathbb{E}\Big\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\Big\|^2 + \mathbb{E}\Big\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\Big\|^2\right) \\
&\leq \frac{2t}{\eta_t}\Big(\mathbb{E}\Big[\mathcal{B}(\boldsymbol{z}_{t-1}, \tilde{\boldsymbol{z}})\Big] - \mathbb{E}\Big[\mathcal{B}(\boldsymbol{z}_t, \tilde{\boldsymbol{z}})\Big]\Big) + \left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right)\frac{t\eta_t}{\mu_{\mathcal{B}}},
\end{aligned}$$

where in the last line above we applied Assumption 2.2, so by law of iterated expectations

$$\begin{aligned}
\mathbb{E}\Big\|\boldsymbol{\Delta}_{\mathcal{F}}^{t-\frac{1}{2}}\Big\|^2 &= \mathbb{E}\left[\Big\|\nabla\tilde{F}(\boldsymbol{z}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) - \nabla F(\boldsymbol{z}_{t-1}^{\mathrm{md}})\Big\|^2\right] \leq \sigma_{\mathrm{Str}}^2, \\
\mathbb{E}\Big\|\boldsymbol{\Delta}_{\mathcal{H}}^{t-\frac{1}{2}}\Big\|^2 &= \mathbb{E}\left[\Big\|\tilde{H}(\boldsymbol{z}_{t-1}; \zeta_{t-\frac{1}{2}}) - H(\boldsymbol{z}_{t-1})\Big\|^2\right] \leq \sigma_{\mathrm{Bil}}^2, \qquad (69) \\
\mathbb{E}\Big\|\boldsymbol{\Delta}_{\mathcal{H}}^{t}\Big\|^2 &= \mathbb{E}\left[\Big\|\tilde{H}(\boldsymbol{z}_{t-\frac{1}{2}}; \zeta_t) - H(\boldsymbol{z}_{t-\frac{1}{2}})\Big\|^2\right] \leq \sigma_{\mathrm{Bil}}^2.
\end{aligned}$$