Proceedings of Machine Learning Research vol 195:1–39, 2023 36th Annual Conference on Learning Theory

Optimal Extragradient-Based Stochastic Bilinearly Coupled Saddle-Point Optimization

author names withheld

Editor: Under Review for COLT 2023

Abstract

We consider the smooth convex-concave bilinearly coupled saddle-point problem, $\min_{\mathbf{x}} \max_{\mathbf{y}} F(\mathbf{x}) + H(\mathbf{x}, \mathbf{y}) - G(\mathbf{y})$, where one has access to stochastic first-order oracles for F, G as well as the bilinear coupling function H. Building upon standard stochastic extragradient analysis for variational inequalities, we present a stochastic *accelerated gradient-extragradient (AG-EG)* descent-ascent algorithm that combines extragradient and Nesterov's acceleration in general stochastic settings. This algorithm leverages scheduled restarting to admit a fine-grained nonasymptotic convergence rate that matches known lower bounds by both Ibrahim et al. (2020) and Zhang et al. (2022) in their corresponding settings, plus an additional statistical error term for bounded stochastic noise that is optimal up to a constant prefactor.

Keywords: Convex optimization, convex-concave bilinearly coupled saddle-point problem, extragradient-based stochastic optimization, Nesterov's acceleration, scheduled restarting, scaling reduction

1. Introduction

In this work, we focus on a widely studied stochastic convex-concave minimax optimization problem with bilinear coupling, specifically the *convex-concave bilinearly coupled saddle-point problem*:

$$\min_{\mathbf{x}\in\mathbb{R}^n} \max_{\mathbf{y}\in\mathbb{R}^m} \mathscr{F}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_\xi \left[f(\mathbf{x}; \xi) \right] + \mathbb{E}_\zeta \left[h(\mathbf{x}, \mathbf{y}; \zeta) \right] - \mathbb{E}_\xi \left[g(\mathbf{y}; \xi) \right] \equiv F(\mathbf{x}) + H(\mathbf{x}, \mathbf{y}) - G(\mathbf{y}) \quad (1)$$

where $H(\mathbf{x}, \mathbf{y}) \equiv \mathbf{x}^\top \mathbf{B} \mathbf{y} - \mathbf{x}^\top \mathbf{u}_{\mathbf{x}} + \mathbf{u}_{\mathbf{y}}^\top \mathbf{y}$ is the bilinear coupling function with the coupling matrix $\mathbf{B}$ of dimension $n \times m$, and where ξ and ζ are drawn from distributions $\mathcal{D}_\xi$ and $\mathcal{D}_\zeta$, respectively. We aim to solve (1) when either both $F(\mathbf{x})$ and $G(\mathbf{y})$ are smooth and strongly convex, or both are zero. In addition to a wide range of applications in economics, problems of form (1) are becoming increasingly important in machine learning. For instance (1) appears in reinforcement learning, differentiable games, regularized empirical risk minimization, and robust optimization formulations. It also can be seen as a local approximation of the objective of nonconvex-nonconcave minimax games—e.g., a GAN—around a local Nash equilibrium (Mescheder et al., 2017; Nagarajan and Kolter, 2017). Our exposition begins with an overview of some of these applications.

Reinforcement learning. Reinforcement learning problems can be formalized as Markov Decision Processes (MDPs) where, at each step $t = 1, \dots, n$, 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:

$$\min_{\boldsymbol{\theta}} \frac{1}{2} \left\| \mathbf{A}\boldsymbol{\theta} - \mathbf{b} \right\|_{\mathbf{C}^{-1}}^2 + \frac{\rho}{2} \left\| \boldsymbol{\theta} \right\|^2 \quad (2)$$

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 (2):

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

which falls under the umbrella of problem (1) whenever $\mathbf{C}$ is positive definite.

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

$$\mathscr{F}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\mathbf{x}^\top\mathbf{M}_F\mathbf{x} + \mathbf{x}^\top\mathbf{B}\mathbf{y} - \frac{1}{2}\mathbf{y}^\top\mathbf{M}_G\mathbf{y} - \mathbf{x}^\top\mathbf{v}_{\mathbf{x}} + \mathbf{v}_{\mathbf{y}}^\top\mathbf{y} \tag{3}$$

where $\mathbf{M}_F, \mathbf{M}_G$ are real-valued matrices of dimensions $n \times n$ and $m \times m$. This has the form (1) with $F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top\mathbf{M}_F\mathbf{x} - \mathbf{x}^\top\mathbf{v}_{\mathbf{x}}$, $G(\mathbf{y}) = \frac{1}{2}\mathbf{y}^\top\mathbf{M}_G\mathbf{y} - \mathbf{v}_{\mathbf{y}}^\top\mathbf{y}$ and $H(\mathbf{x},\mathbf{y}) \equiv \mathbf{x}^\top\mathbf{B}\mathbf{y}$. A particular case we will be considering in §3 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 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:

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

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

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

See Zhang and Xiao (2017); Wang and Xiao (2017); Xiao et al. (2019) for in-depth studies of solving this problem under such a dual form of representation.

1.1. Main contributions

Despite the range of real-world applications of the bilinearly coupled saddle-point problem in (1), there is a limited nonasymptotic theoretical analysis of the problem. Notable exceptions include Zhang et al. (2022) and Ibrahim et al. (2020), who develop lower bounds in the strongly-convex-strongly-concave and bilinear cases. *Matching these lower bounds in a single algorithm in the general stochastic setting has been an open problem.* In particular, standard acceleration techniques do not achieve the optimal nonasymptotic convergence rate for the bilinear minimax game (Gidel et al., 2019b).

We tackle this problem in a new way, proposing a stochastic *accelerated gradient-extragradient (AG-EG)* descent-ascent algorithm for solving (1), bringing together Nesterov's acceleration method

(Nesterov, 1983)—applied to the individual $F(\mathbf{x})$ and $G(\mathbf{y})$ terms—and the extragradient method (Korpelevich, 1976)—which extrapolates the bilinear coupling term. This combination allows us to arrive at a general algorithmic convergence result that yields optimality in nonasymptotic convergence rates for the strongly-convex-strongly-concave and bilinear cases. This general result subsumes many special cases of interest:

- For the function class of bilinear games where $\nabla f(\mathbf{x};\xi) = \mathbf{0}$ and $\nabla g(\mathbf{y};\xi) = \mathbf{0}$ a.s., Algorithm 1, equipped with scheduled restarting achieves an $O\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}\log\left(\frac{\sqrt[4]{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}{\sigma_{\text{Bil}}}\right) + \frac{\sigma_{\text{Bil}}^2}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\varepsilon^2}\right)$ iteration complexity, 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 the bilinear problem, matching the lower bound of Ibrahim et al. (2020).[1] In other words, our algorithm admits a sharp dependency on $\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$ and matches the Ibrahim et al. (2020) lower bound [§3, Corollary 5].

- For the function class of strongly-convex-strongly-concave objectives, the same stochastic AG-EG descent-ascent Algorithm 1, when equipped with scheduled restarting, achieves an iteration complexity of $O\left(\left(\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}\right)\log\left(\frac{1}{\varepsilon}\right) + \frac{\sigma^2}{\mu_F^2\varepsilon^2}\right)$, 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, and σ^2 is the weighted, uniform variance bound on the stochastic gradient. When the problem is nonrandom, this complexity upper bound matches the Zhang et al. (2022) lower bound [§4, Corollary 8].

- We also present a direct approach for the function class of strongly-convex-strongly-concave objectives, where the lower bound in iteration complexity due to Zhang et al. (2022) is matched as $\left(\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}} + O\left(\frac{\sigma^2}{\mu_F^2\varepsilon^2}\right)\right)\log\left((\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G})\frac{1}{\varepsilon}\right)$, an iteration complexity that admits a sharp near-unity coefficient [§5, Theorem 10].

Throughout our analysis, we frequently make use of a *scheduled-restarting* approach and a *scaling-reduction* argument that allows us to reduce problems to cases that are easier to analyze. This general strategy may be of independent interest.

Organization. The rest of this work is organized as follows. §2 presents the basic settings and assumptions. §3 gives the optimality of convergence for our proposed AcceleratedGradient-Extragradient (AG-EG) descent-ascent algorithm, for the class of bilinear games, and §4 presents the optimality of AG-EG for the class of strongly-convex-strongly-concave objectives. §5 provides an alternative direct approach for the same strongly-convex-strongly-concave function class. §6 discusses future directions. In the Appendix, §A compares our results with related work, §B details the proofs of our main convergence results, and §C supplements the proofs with auxiliary lemmas.

[1]As will be discussed in Assumption 2 of §2, we can assume our coupling matrix $\mathbf{B}$ is tall in the sense that $n \geq m$ without loss of generality. 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 for more on this.

Notations. 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 $a_n = O(b_n)$ or $a_n \lesssim b_n$ (resp. $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$) to denote $a_n \le Cb_n$ (resp. $a_n \ge Cb_n$) for all $n \ge 1$ 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 $a_n = \tilde{O}(b_n)$ denote $a_n \le Cb_n$ where C hides a polylogarithmic factor in problem-dependent constants. We let $[\mathbf{x}; \mathbf{y}] \in \mathbb{R}^{n+m}$ concatenate two vectors, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. 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 $\mathbf{v}^\top(\mathbf{A} - \mathbf{B})\mathbf{v} \le 0$ (resp. $\mathbf{v}^\top(\mathbf{A} - \mathbf{B})\mathbf{v} \ge 0$) holds for all vectors $\mathbf{v}$.

2. Setting and assumptions

In this section, we formally introduce our framework and assumptions. Our development is inspired by the work of Chen et al. (2017) on a (stochastic) *Accelerated MirrorProx (AMP)* algorithm. This work is developed in the general setting of monotone variational inequalities with a $O(1/\sqrt{T})$ convergence rate bound where the prefactor depends on domain size and hence does not accommodate unbounded domains. As a result when translated into minimax optimization, this result does not match the lower bound in Zhang et al. (2022). To achieve the lower bound, we present an alternative approach in Algorithm 1, the stochastic *accelerated gradient-extragradient (AG-EG)* descent-ascent algorithm. Our algorithm applies Nesterov's acceleration method (Nesterov, 1983) to the individual $F(\mathbf{x})$ and $G(\mathbf{y})$ terms and applies the extragradient method (Korpelevich, 1976) to the bilinear coupling term. As we show, a particular combination—with the incorporation of scheduled restarting—succeeds in matching the lower bound provided in Ibrahim et al. (2020) and Zhang et al. (2022) in their corresponding settings.

For simplicity, we consider unconstrained domains $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. For the constrained case with convex domains one can introduce a projection step and proceed analogously with the analysis; we omit this generalization for simplicity. We first state the smoothness and convexity assumptions that we impose on the $F(\mathbf{x})$ and $G(\mathbf{y})$ terms.

Assumption 1 (Convexity and smoothness) *We assume that $F(\mathbf{x})$ is L_F-smooth and μ_F-strongly convex, and $G(\mathbf{y})$ is L_G-smooth and μ_G-strongly convex. That is, for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$,*

$$\tfrac{\mu_F}{2}\|\mathbf{x} - \mathbf{x}'\|^2 \le F(\mathbf{x}) - F(\mathbf{x}') - \nabla F(\mathbf{x}')^\top(\mathbf{x} - \mathbf{x}') \le \tfrac{L_F}{2}\|\mathbf{x} - \mathbf{x}'\|^2,$$

and for any $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^m$,

$$\tfrac{\mu_G}{2}\|\mathbf{y} - \mathbf{y}'\|^2 \le G(\mathbf{y}) - G(\mathbf{y}') - \nabla G(\mathbf{y}')^\top(\mathbf{y} - \mathbf{y}') \le \tfrac{L_G}{2}\|\mathbf{y} - \mathbf{y}'\|^2$$

We assume that the coupling matrix $\mathbf{B}$ is a tall matrix, which can otherwise be satisfied by considering the symmetrized problem, $\min_{\mathbf{y}} \max_{\mathbf{x}} -f(\mathbf{x}, \mathbf{y})$ (an equivalence guaranteed by the strong convexity of the functions and Sion's minimax theorem (Sion, 1958).)

Assumption 2 (Coupling matrix) *We assume without loss of generality that $\mathbf{B}$ is tall, i.e., $n \ge m$.*

Assumption 2, which is introduced for the purpose of notational consistency, guarantees that $\lambda_{\max}(\mathbf{B}^\top\mathbf{B}) = \lambda_{\max}(\mathbf{B}\mathbf{B}^\top)$ but $\lambda_{\min}(\mathbf{B}^\top\mathbf{B}) \ge \lambda_{\min}(\mathbf{B}\mathbf{B}^\top)$, where the latter is strictly zero when $\mathbf{B}$ is nonsquare.

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

Require: Initialization $\mathbf{x}_0^{[0]}, \mathbf{y}_0^{[0]}$, total number of epochs $\mathscr{S} \geq 1$, total number of per-epoch iterates $(\mathscr{T}_s : s = 1, \dots, \mathscr{S})$, stepsizes $(\alpha_t, \eta_t : t = 1, 2, \dots)$, ratio of strong-convexity parameters $\mathscr{R} = \frac{\mu_G}{\mu_F}$

for $s = 1, 2, \dots, \mathscr{S}$ **do**
  Set $\mathbf{x}^{\mathrm{ag}}_{-\frac{1}{2}} \leftarrow \mathbf{x}_0^{[s-1]}, \mathbf{y}^{\mathrm{ag}}_{-\frac{1}{2}} \leftarrow \mathbf{y}_0^{[s-1]}, \mathbf{x}_0 \leftarrow \mathbf{x}_0^{[s-1]}, \mathbf{y}_0 \leftarrow \mathbf{y}_0^{[s-1]}, \mathbf{x}_0^{\mathrm{md}} \leftarrow \mathbf{x}_0^{[s-1]}, \mathbf{y}_0^{\mathrm{md}} \leftarrow \mathbf{y}_0^{[s-1]}$
  for $t = 1, 2, \dots, \mathscr{T}_s$ **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
    $\mathbf{x}_{t-\frac{1}{2}} \leftarrow \mathbf{x}_{t-1} - \eta_t \left(\nabla f(\mathbf{x}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}})\right)$
    $\mathbf{y}_{t-\frac{1}{2}} \leftarrow \mathbf{y}_{t-1} - \frac{\eta_t}{\mathscr{R}} \left(-\nabla_{\mathbf{y}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) + \nabla g(\mathbf{y}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}})\right)$
    $\mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}} \leftarrow (1-\alpha_t)\mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}} + \alpha_t \mathbf{x}_{t-\frac{1}{2}}$
    $\mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}} \leftarrow (1-\alpha_t)\mathbf{y}^{\mathrm{ag}}_{t-\frac{3}{2}} + \alpha_t \mathbf{y}_{t-\frac{1}{2}}$
    $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_t \left(\nabla f(\mathbf{x}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t)\right)$
    $\mathbf{y}_t \leftarrow \mathbf{y}_{t-1} - \frac{\eta_t}{\mathscr{R}} \left(-\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\mathbf{y}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}})\right)$
    $\mathbf{x}^{\mathrm{md}}_t \leftarrow (1-\alpha_{t+1})\mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}} + \alpha_{t+1}\mathbf{x}_t$
    $\mathbf{y}^{\mathrm{md}}_t \leftarrow (1-\alpha_{t+1})\mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}} + \alpha_{t+1}\mathbf{y}_t$
  end for
  Set $\mathbf{x}_0^{[s]} \leftarrow \mathbf{x}^{\mathrm{ag}}_{\mathscr{T}_s - \frac{1}{2}}, \mathbf{y}_0^{[s]} \leftarrow \mathbf{y}^{\mathrm{ag}}_{\mathscr{T}_s - \frac{1}{2}}$ //Warm-start using the output of the previous epoch
end for
Output: $[\mathbf{x}_0^{[\mathscr{S}]}; \mathbf{y}_0^{[\mathscr{S}]}]$

It is straightforward to show that (1) admits a unique saddle point (or Nash equilibrium) in the strongly-convex-strongly-concave case [Assumption 1]; i.e., there exists a unique pair $(\boldsymbol{\omega}^\star_{\mathbf{x}}, \boldsymbol{\omega}^\star_{\mathbf{y}})$ such that

$$\mathscr{F}(\boldsymbol{\omega}^\star_{\mathbf{x}}, \mathbf{y}) \leq \mathscr{F}(\boldsymbol{\omega}^\star_{\mathbf{x}}, \boldsymbol{\omega}^\star_{\mathbf{y}}) \leq \mathscr{F}(\mathbf{x}, \boldsymbol{\omega}^\star_{\mathbf{y}}), \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n \text{ and } \mathbf{y} \in \mathbb{R}^d. \tag{SP}$$

For the bilinear game case where $L_F = \mu_F = 0$, $L_G = \mu_G = 0$, this is satisfied for square matrices $\mathbf{B}$ with least singular value being strictly positive.

Third, we impose assumptions on the noise variance. We first introduce the following rescaling parameters:

$$L_{\mathrm{Str}} = L_F \vee \left(\frac{\mu_F}{\mu_G} L_G\right), \qquad L_{\mathrm{Bil}} = \sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B}) \cdot \frac{\mu_F}{\mu_G}}, \qquad \mu_{\mathrm{Str}} = \mu_F, \qquad \mathscr{R} = \frac{\mu_G}{\mu_F} \tag{4}$$

Assumption 3 (Unbiased gradients and variance bounds) *We assume that* $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$, $\xi \sim \mathcal{D}_\xi$ *and* $\zeta \sim \mathcal{D}_\zeta$ *are drawn from distributions such that the following conditions hold:* $\mathbb{E}_\xi[\nabla f(\mathbf{x}; \xi)] = \nabla F(\mathbf{x})$, $\mathbb{E}_\xi[\nabla g(\mathbf{y}; \xi)] = \nabla G(\mathbf{y})$, $\mathbb{E}_\zeta[\nabla_{\mathbf{x}} h(\mathbf{x}, \mathbf{y}; \zeta)] = \nabla_{\mathbf{x}} H(\mathbf{x}, \mathbf{y})$ *and* $\mathbb{E}_\zeta[\nabla_{\mathbf{y}} h(\mathbf{x}, \mathbf{y}; \zeta)] = \nabla_{\mathbf{y}} H(\mathbf{x}, \mathbf{y})$, *with*

$$\mathbb{E}_\xi\left[\|\nabla f(\mathbf{x}; \xi) - \nabla F(\mathbf{x})\|^2 + \frac{1}{\mathscr{R}}\|\nabla g(\mathbf{y}; \xi) - \nabla G(\mathbf{y})\|^2\right] \leq \sigma^2_{\mathrm{Str}} \tag{5}$$

and

$$\mathbb{E}_\zeta\left[\|\nabla_{\mathbf{x}} h(\mathbf{x},\mathbf{y};\zeta)-\nabla_{\mathbf{x}} H(\mathbf{x},\mathbf{y})\|^2+\tfrac{1}{\mathscr{R}}\|-\nabla_{\mathbf{y}} h(\mathbf{x},\mathbf{y};\zeta)+\nabla_{\mathbf{y}} H(\mathbf{x},\mathbf{y})\|^2\right]\le\sigma_{\text{Bil}}^2 \tag{6}$$

For all results in this work, we suppose that Assumptions 1, 2 and 3 hold with appropriate parameter settings. Given a desired accuracy $\varepsilon>0$, our goal is to find an ε-*saddle*, namely a point $[\mathbf{x};\mathbf{y}]$ such that $\|\mathbf{x}-\boldsymbol{\omega}_{\mathbf{x}}^\star\|^2+\mathscr{R}\|\mathbf{y}-\boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\le\varepsilon^2$—for the purposes of our analysis we adopt this slightly different metric that is equivalent to Euclidean norm. The resulting iteration complexities in the Euclidean norm are obtained by replacing the ε-desired accuracy for that metric by $\varepsilon/(\sqrt{\mathscr{R}\vee\mathscr{R}^{-1}})$.[2]

3. Optimality for bilinear games

We first consider the particular case of bilinear games, where we show that Algorithm 1, with proper averaging and scheduled restarting, achieves an optimal statistical rate up to a constant prefactor and with a bias term that matches the lower bound of Ibrahim et al. (2020, Theorem 3) for bilinear games. *We assume in this section that $n=m$ where $\mathbf{B}$ is a nonsingular square matrix*, $\nabla f(\mathbf{x};\xi)=\mathbf{0}$ and $\nabla g(\mathbf{y};\xi)=\mathbf{0}$ a.s., so (1) reduces to

$$\min_{\mathbf{x}}\max_{\mathbf{y}}\ \mathscr{F}(\mathbf{x},\mathbf{y})=\mathbb{E}_\zeta\left[h(\mathbf{x},\mathbf{y};\zeta)\right]=H(\mathbf{x},\mathbf{y})=\mathbf{x}^\top\mathbf{B}\mathbf{y}-\mathbf{x}^\top\mathbf{u}_{\mathbf{x}}+\mathbf{u}_{\mathbf{y}}^\top\mathbf{y} \tag{7}$$

and Algorithm 1 reduces to the independent-sample extragradient descent-ascent algorithm for (7). The saddle point $[\boldsymbol{\omega}_{\mathbf{x}}^\star;\boldsymbol{\omega}_{\mathbf{y}}^\star]$ in this case is the unique solution to the linear equation

$$\begin{bmatrix}\mathbf{0} & \mathbf{B}\\ -\mathbf{B}^\top & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\omega}_{\mathbf{x}}^\star\\ \boldsymbol{\omega}_{\mathbf{y}}^\star\end{bmatrix}=\begin{bmatrix}\mathbf{u}_{\mathbf{x}}\\ \mathbf{u}_{\mathbf{y}}\end{bmatrix},\quad\text{which reduces to }\begin{bmatrix}-(\mathbf{B}^\top)^{-1}\mathbf{u}_{\mathbf{y}}\\ \mathbf{B}^{-1}\mathbf{u}_{\mathbf{x}}\end{bmatrix}.$$

In earlier work, Azizian et al. (2020b, Proposition 7) achieve an upper bound that matches the lower bound of Ibrahim et al. (2020). Our algorithm is in the independent-sample setting with bounded noise variance, that is, it consumes two independent samples, one for each in the extrapolation and update steps. This is different from the version of Li et al. (2021) that consumes a shared sample in both steps. We allow the algorithm to be initialized randomly, which generalized the case of nonrandom initialization where the point mass is put on $[\mathbf{x}_0;\mathbf{y}_0]$. Due to the special stepsize selection in the averaging, our analysis of stochastic bilinear games yields the following:

Theorem 4 (Convergence of stochastic AG-EG, bilinear case) *Setting parameters as in* (4) *with $\mathscr{R}$ being arbitrary, $L_{\text{Bil}}=\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})\cdot\frac{1}{\mathscr{R}}}$, $L_{\text{Str}}=\mu_{\text{Str}}=0$, $\mathscr{R}=1$,and also choosing the stepsizes $\alpha_t=\frac{2}{t+1}$ and $\eta_t\equiv\frac{1}{L_{\text{Bil}}}=\sqrt{\frac{\mathscr{R}}{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$, we have*

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{x}_{\mathscr{T}-\frac12}^{\text{ag}}-\boldsymbol{\omega}_{\mathbf{x}}^\star\|^2+\mathscr{R}\|\mathbf{y}_{\mathscr{T}-\frac12}^{\text{ag}}-\boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right]\\ &\le\frac{\mathscr{R}}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\left(\frac{4\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})\cdot\frac{1}{\mathscr{R}}}}{\mathscr{T}}\sqrt{\mathbb{E}\left[\|\mathbf{x}_0-\boldsymbol{\omega}_{\mathbf{x}}^\star\|^2+\mathscr{R}\|\mathbf{y}_0-\boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right]}+\frac{7\sigma_{\text{Bil}}}{\sqrt{\mathscr{T}}}\right)^2\end{aligned} \tag{8}$$

[2]The metric conversion from duality gap to Euclidean distance or weighted Euclidean distance to saddle in our case is straightforward: they are equivalent up to a multiplicative factor. In other words, although our algorithmic convergence is characterized by the metric of weighted Euclidean distance, it completely matches both lower bounds by Zhang et al. (2022) and Ibrahim et al. (2020) in the nonstochastic setting.

The proof of Theorem 4 is provided in §B.5. For regularity purposes we set the ratio of strong-convexity parameters (both being zero) as $\mathscr{R} = 1$ in the rest of this section.[3] Note that our choice of the stepsize is maximal and is independent of the noise. Let us now consider a scheduled restarting version of the algorithm, with a constant epoch length $\asymp \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}$ steps using with the same constant stepsize, until the iteration reaches the stationary noise level in the sense that $\mathbb{E}\left[\|\mathbf{x}_0 - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right] \asymp \frac{\sigma_{\text{Bil}}^2}{\sqrt{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}$. The convergence rate for this restarting variant is linear in bias term plus an optimal statistical error term, as follows:

Corollary 5 (Convergence of stochastic AG-EG with scheduled restarting, bilinear case) *Equipped with scheduled restarting, the iteration complexity is bounded by*

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

In the setting where there is no stochasticity, and letting $\sigma_{\text{Bil}} \asymp \varepsilon\sqrt[4]{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}$, the complexity bound in Corollary 5 reduces to $O\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}\log\left(\frac{1}{\varepsilon}\right)\right)$ and hence matches the lower bound of Ibrahim et al. (2020). The $\frac{\sigma_{\text{Bil}}^2}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)\varepsilon^2}$ term corresponds to the optimal statistical rate for the current problem.

4. Optimality for strongly-convex-strongly-concave objectives

In this section, we proceed to solve (1) using Algorithm 1 in the general strongly-convex-strongly-concave case. We assume throughout this section that $H(\mathbf{x}, \mathbf{y})$ is of bilinear form $\mathbf{x}^\top\mathbf{B}\mathbf{y} - \mathbf{x}^\top\mathbf{u}_{\mathbf{x}} + \mathbf{u}_{\mathbf{y}}^\top\mathbf{y}$, without assuming $n = m$ or the nonsingularity of $\mathbf{B}$. Recall that Algorithm 1 conducts acceleration on the terms $F(\mathbf{x})$ and $G(\mathbf{y})$ and extrapolates on the bilinear term $H(\mathbf{x}, \mathbf{y})$. We continue to allow $[\mathbf{x}_0; \mathbf{y}_0]$ be randomly initialized and denote

$$\bar{\eta}_t(\tilde{\sigma}; \mathscr{T}, \mathscr{C}, r, \beta) \equiv \frac{t}{\frac{2}{r}L_{\text{Str}} \vee \frac{\tilde{\sigma}[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}{\mathscr{C}\sqrt{\mathbb{E}\left[\|\mathbf{x}_0 - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right]}} + \sqrt{\frac{1+\beta}{r}}L_{\text{Bil}}t} \tag{9}$$

where $\mathscr{C} \in (0, \infty)$ is an input parameter that allows flexibility in our stepsize selection. We state our general result as follows:

Theorem 6 (Convergence of stochastic AG-EG) *Let the epoch length $\mathscr{T} \geq 1$ be known in advance, fix $r \in (0, 1)$ and $\beta \in (0, \infty)$ arbitrarily, set the rescaling parameters L_{Str}, L_{Bil}, μ_{Str}, $\mathscr{R}$ as in (4), set $\sigma \equiv \frac{1}{\sqrt{3}}\sqrt{\frac{1}{1-r}\sigma_{\text{Str}}^2 + (2 + \frac{1}{\beta})\sigma_{\text{Bil}}^2}$ and choose the stepsizes $\alpha_t = \frac{2}{t+1}$ and $\eta_t = \bar{\eta}_t(\sigma; \mathscr{T}, \mathscr{C}, r, \beta)$ to be defined as in (9) with $\mathscr{C} \in (0, \infty)$ being an input parameter. We have that*

[3] Since it is of an $\frac{0}{0}$-indefinite form, the result also holds for an arbitrary choice of $\mathscr{R} \in (0, \infty)$, providing flexibility on the parameter choices.

the output of single-epoch ($\mathscr{S} = 1$) Algorithm 1 $[\mathbf{x}^{\text{ag}}_{\mathscr{T}-\frac{1}{2}}; \mathbf{y}^{\text{ag}}_{\mathscr{T}-\frac{1}{2}}]$ *satisfies*

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}^{\text{ag}}_{\mathscr{T}-\frac{1}{2}} - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}^{\text{ag}}_{\mathscr{T}-\frac{1}{2}} - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right] \\
&\le \frac{2}{\mu_{\text{Str}}(\mathscr{T}+1)}\left(\frac{\frac{2}{r}L_{\text{Str}}}{\mathscr{T}} + \mathcal{A}(\sigma;\mathscr{T},\mathscr{C},r,\beta)\sqrt{\frac{1+\beta}{r}}L_{\text{Bil}}\right)\mathbb{E}\left[\|\mathbf{x}_0 - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right] \\
&\quad + \frac{2(\frac{1}{\mathscr{C}}+\mathscr{C})\sigma}{\mu_{\text{Str}}\mathscr{T}^{1/2}}\sqrt{\mathbb{E}\left[\|\mathbf{x}_0 - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right]},
\end{aligned} \tag{10}
$$

where the prefactor

$$
\mathcal{A}(\tilde{\sigma};\mathscr{T},\mathscr{C},r,\beta) \equiv 1 + \frac{\mathscr{C}\tilde{\sigma}[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}{\frac{1}{\eta_1(\tilde{\sigma};\mathscr{T},\mathscr{C},r,\beta)}\sqrt{\mathbb{E}\left[\|\mathbf{x}_0 - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right]}} \tag{11}
$$

lies in $[1, 1+\mathscr{C}^2]$ *and reduces to 1 when* $\tilde{\sigma} = 0$.

The proof of Theorem 6 is provided in §B.3. In the case that there is no stochasticity, by taking $r \to 1^-$, $\beta \to 0^+$ in our analysis we obtain the following result:

Theorem 7 (Convergence of AG-EG) *Setting the rescaling parameters* L_{Str}, L_{Bil}, μ_{Str}, $\mathscr{R}$ *as in (4), we have that by choosing* $\eta_t = \frac{t}{2L_{\text{Str}}+L_{\text{Bil}}t}$ *the output of Algorithm 1 with* $\mathscr{S} = 1$ *satisfies*

$$
\|\mathbf{x}^{\text{ag}}_{\mathscr{T}-\frac{1}{2}} - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}^{\text{ag}}_{\mathscr{T}-\frac{1}{2}} - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2 \le \frac{2}{\mu_{\text{Str}}(\mathscr{T}+1)}\left(\frac{2L_{\text{Str}}}{\mathscr{T}} + L_{\text{Bil}}\right)\left[\|\mathbf{x}_0 - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right], \tag{12}
$$

where $\mathbf{x}^{\text{ag}}_{\mathscr{T}-\frac{1}{2}}$ *and* $\mathbf{y}^{\text{ag}}_{\mathscr{T}-\frac{1}{2}}$ *are defined in Algorithm 1.*

We make a few remarks on Theorems 6 and 7 as follows:

(i) When L_{Bil} is set as zero the problem is decoupled, and our algorithm for a single variate reduces to the standard three-line formulation of stochastic Nesterov's accelerated gradient descent, where the choice of $\alpha_t = \frac{2}{t+1}$ is essential to achieve desirable convergence behavior (Nesterov, 1983). The stepsize choice $\eta_t = \bar{\eta}_t(\sigma;\mathscr{T},\mathscr{C},r,\beta)$ as in (9) is directly generalized from the optimal choice in stochastic Nesterov's method by incorporating the bilinear coupling term in its dominator; we refer interested readers to (Lan, 2020, Chap. 4) for a careful treatment. Our hyperparameter dependency is in a fine-grained fashion; often, the convergence rate coefficients are not a concern, and the coarse choices of $r = \frac{1}{2}$ and $\beta = 1$ should suffice. In words, how r deviating from 1 and β deviating from 0 should be a tradeoff between the noise variance and the convergence rate coefficients.

(ii) Compared with Theorem 4, the nonasymptotic convergence rate in Theorem 6 is slowed down from $O(\frac{1}{\mathscr{T}})$ to $O(\frac{1}{\sqrt{\mathscr{T}}})$ in squared metric due to the nonlinear nature of our system. As we will see immediately afterwards, with the use of scheduled restarting the bias term converges to zero at a linear, accelerated rate. The stepsize choice in Theorem 7 and Theorem 4 is consistent, although they apply in different cases.

(iii) The choice of $\mathscr{C}$ reflects a tradeoff between the terms in our convergence rate bounds. In the nonrandom setting the algorithm does not require any knowledge or estimate of the initial distance to a saddle to achieve the desirable rate.

(iv) In the general case that the randomness exists, the leading-order term on the right hand of (10) admits a *hyperbolic dependency* on $\mathscr{C}$; that is, the $\mathscr{C}$-dependent prefactor in the bound is precisely $\frac{1}{\mathscr{C}} + \mathscr{C}$. In the case that we are given the full knowledge of the initial distance to a saddle we can optimally choose $\mathscr{C} = 1$ on the right-hand side of (10). In the alternative case where only an upper estimate Γ_0 of $\sqrt{\mathbb{E}\left[\|\mathbf{x}_0 - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right]}$ is provided, we simply set $\mathscr{C} = \frac{\Gamma_0}{\sqrt{\mathbb{E}\left[\|\mathbf{x}_0 - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right]}} \ge 1$ which yields the following bound to (10)

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}^{\mathrm{ag}}_{\mathscr{T}-\frac{1}{2}} - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \mathscr{R}\|\mathbf{y}^{\mathrm{ag}}_{\mathscr{T}-\frac{1}{2}} - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right] \\
&\le \tfrac{2}{\mu_{\mathrm{Str}}(\mathscr{T}+1)}\left(\frac{\frac{2}{r}L_{\mathrm{Str}}}{\mathscr{T}}\mathbb{E}\left[\|\mathbf{x}_0 - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right] + 2\sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}}\Gamma_0^2\right) + \tfrac{4\sigma}{\mu_{\mathrm{Str}}\mathscr{T}^{1/2}}\Gamma_0 \\
&\le \tfrac{2}{\mu_{\mathrm{Str}}(\mathscr{T}+1)}\left(\frac{\frac{2}{r}L_{\mathrm{Str}}}{\mathscr{T}} + 2\sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}}\right)\Gamma_0^2 + \tfrac{4\sigma}{\mu_{\mathrm{Str}}\mathscr{T}^{1/2}}\Gamma_0
\end{aligned} \tag{13}$$

Note we used $\mathcal{A}(\sigma; \mathscr{T}, \mathscr{C}, r, \beta) \le 2\mathscr{C}^2$. The analysis for our upcoming algorithm with scheduled restarting is highly reliant on the aforementioned bound.

To prepare for our multi-epoch result with the help of scheduled restarting, in an argument analogous to Corollary 5 we perform an induction based on (13): suppose $\mathbb{E}\left[\|\mathbf{x}_0^{[s-1]} - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \mathscr{R}\|\mathbf{y}_0^{[s-1]} - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right] \le \Gamma_0^2 e^{1-s}$ hold, and we obtain (by taking $r = \frac{1}{2}$ and $\beta = 1$ for simplicity)

$$\mathbb{E}\left[\|\mathbf{x}_0^{[s]} - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \mathscr{R}\|\mathbf{y}_0^{[s]} - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right] \lesssim \tfrac{L_{\mathrm{Str}}}{\mu_{\mathrm{Str}}\mathscr{T}_s^2}\Gamma_0^2 e^{1-s} + \tfrac{L_{\mathrm{Bil}}}{\mu_{\mathrm{Str}}\mathscr{T}_s}\Gamma_0^2 e^{1-s} + \tfrac{\sigma}{\mu_{\mathrm{Str}}\mathscr{T}_s^{1/2}}\Gamma_0 e^{\frac{1-s}{2}}$$

Setting the above display as $\le \Gamma_0^2 e^{-s}$, and setting the length of epoch s as $\mathscr{T}_s \asymp \sqrt{\frac{L_{\mathrm{Str}}}{\mu_{\mathrm{Str}}}} + \frac{L_{\mathrm{Bil}}}{\mu_{\mathrm{Str}}} + \frac{\sigma^2}{\mu_{\mathrm{Str}}^2\Gamma_0^2 e^{1-s}}$ we arrive at a total complexity of

$$\lesssim \sum_{s=1}^{LOG}\left[\sqrt{\tfrac{L_{\mathrm{Str}}}{\mu_{\mathrm{Str}}}} + \tfrac{L_{\mathrm{Bil}}}{\mu_{\mathrm{Str}}} + \tfrac{\sigma^2}{\mu_{\mathrm{Str}}^2\Gamma_0^2 e^{1-s}}\right] = \left(\sqrt{\tfrac{L_{\mathrm{Str}}}{\mu_{\mathrm{Str}}}} + \tfrac{L_{\mathrm{Bil}}}{\mu_{\mathrm{Str}}}\right)LOG + \tfrac{\sigma^2}{\mu_{\mathrm{Str}}^2\Gamma_0^2}\cdot\tfrac{e^{LOG}-1}{e-1}$$

where $LOG \equiv \left\lceil \log\frac{\Gamma_0^2}{\varepsilon^2}\right\rceil$, so it is bounded by a constant multiple of

$$\left(\sqrt{\tfrac{L_{\mathrm{Str}}}{\mu_{\mathrm{Str}}}} + \tfrac{L_{\mathrm{Bil}}}{\mu_{\mathrm{Str}}}\right)\left\lceil \log\tfrac{\Gamma_0^2}{\varepsilon^2}\right\rceil + \tfrac{\sigma^2}{\mu_{\mathrm{Str}}^2\Gamma_0^2}e^{\left\lceil \log\frac{\Gamma_0^2}{\varepsilon^2}\right\rceil} \asymp \left(\sqrt{\tfrac{L_{\mathrm{Str}}}{\mu_{\mathrm{Str}}}} + \tfrac{L_{\mathrm{Bil}}}{\mu_{\mathrm{Str}}}\right)\left\lceil \log\tfrac{\Gamma_0^2}{\varepsilon^2}\right\rceil + \tfrac{\sigma^2}{\mu_{\mathrm{Str}}^2\varepsilon^2}$$

This yields the following multi-epoch iteration complexity bound result:

Corollary 8 (Convergence of stochastic AG-EG with scheduled restarting) *When a scheduled restarting argument is employed on top of Algorithm 1, with an epoch length* $\mathscr{T}_s \asymp \sqrt{\frac{L_{\mathrm{Str}}}{\mu_{\mathrm{Str}}}} + \frac{L_{\mathrm{Bil}}}{\mu_{\mathrm{Str}}} + \frac{\sigma^2}{\mu_{\mathrm{Str}}^2 \Gamma_0^2 e^{1-s}}$ *we obtain the iteration complexity of*

$$O\left(\left(\sqrt{\frac{L_{\mathrm{Str}}}{\mu_{\mathrm{Str}}}} + \frac{L_{\mathrm{Bil}}}{\mu_{\mathrm{Str}}}\right)\log\left(\frac{1}{\varepsilon}\right) + \frac{\sigma^2}{\mu_{\mathrm{Str}}^2 \varepsilon^2}\right) = O\left(\left(\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}\right)\log\left(\frac{1}{\varepsilon}\right) + \frac{\sigma^2}{\mu_F^2 \varepsilon^2}\right)$$

In the nonrandom setting, the iteration complexity upper bound in Theorem 8 matches the lower bound of Zhang et al. (2022), which is $\Omega\left(\left(\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$, and we achieve the optimal statistical rate $\frac{\sigma^2}{\mu_F^2 \varepsilon^2}$ up to a constant prefactor. Note that the hard instance constructed by Zhang et al. (2022) has the form of a quadratic minimax game, and hence it is in a special case of a bilinearly coupled saddle-point problem, and the same lower bound holds for problem (1) in this case.

5. A direct approach for strongly-convex-strongly-concave objectives

For solving (1) we turn to our (AMP-inspired) stochastic AG-EG algorithm that targets strongly-convex-strongly-concave problems. For $F(\mathbf{x})$ being μ_F-strongly-convex and $G(\mathbf{y})$ being μ_G-strongly-convex, and letting the algorithm be initialized at a fixed $[\mathbf{x}_0; \mathbf{y}_0]$, we group the objective in (1) as

$$\begin{aligned}\mathscr{F}(\mathbf{x}, \mathbf{y}) = \left(F(\mathbf{x}) - \tfrac{\mu_F}{2}\|\mathbf{x} - \mathbf{x}_0\|^2\right) + \left(\tfrac{\mu_F}{2}\|\mathbf{x} - \mathbf{x}_0\|^2 + H(\mathbf{x}, \mathbf{y}) - \tfrac{\mu_G}{2}\|\mathbf{y} - \mathbf{y}_0\|^2\right) \\ - \left(G(\mathbf{y}) - \tfrac{\mu_G}{2}\|\mathbf{y} - \mathbf{y}_0\|^2\right)\end{aligned} \tag{14}$$

where $\frac{\mu_F}{2}\|\mathbf{x} - \mathbf{x}_0\|^2 + H(\mathbf{x}, \mathbf{y}) - \frac{\mu_G}{2}\|\mathbf{y} - \mathbf{y}_0\|^2$ is a μ_F-strongly-convex-μ_G-strongly-concave isotropic quadratic function. Applying the updates in Lines 5 to 12 in Algorithm 1 to the new grouping yields Algorithm 2, which resembles the algorithmic design of Thekumparampil et al. (2022), except we are employing an extragradient method instead of the Chambolle-Pock-style primal-dual method as an approximation of proximal point methods. We also redefine our rescaling parameters in this section as

$$L_{\mathrm{Str}} = L_F \vee \left(\frac{\mu_F}{\mu_G} L_G\right) - \mu_F, \quad L_{\mathrm{Bil}} = \sqrt{\lambda_{\max}(\mathbf{B}^\top \mathbf{B}) \cdot \frac{\mu_F}{\mu_G} + \mu_F^2}, \quad \mu_\star = \mu_F, \quad \mathscr{R} = \frac{\mu_G}{\mu_F} \tag{15}$$

Our new result is as follows:

Theorem 9 (Convergence of stochastic AG-EG, direct approach) *For solving problem* (14)*, assume for each* $\mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^m$ *and* $\xi \sim \mathcal{D}_\xi, \zeta \sim \mathcal{D}_\zeta$ *that* (5) *and* (6) *are satisfied. Fix arbitrarily* $r \in (0, 1)$, $\beta \in (0, \infty)$, *set the rescaling parameters* L_{Str}, L_{Bil}, μ_{Str}, $\mathscr{R}$ *as in* (15)*, and choose the stepsizes* $\alpha_t \in (0, \bar{\alpha}(r, \beta)]$ *with*

$$\bar{\alpha}(r, \beta) \equiv \frac{r}{1 + \sqrt{1 + r\left(\frac{L_{\mathrm{Str}}}{\mu_\star} + \frac{(1+\beta)L_{\mathrm{Bil}}^2}{\mu_\star^2}\right)}} \tag{16}$$

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

Require: Initialization $\mathbf{x}_0, \mathbf{y}_0$, total number of iterates $\mathscr{T}$, stepsizes $(\alpha_t, \eta_t : t = 1, 2, \dots)$, ratio of strong-convexity parameters $\mathscr{R} = \frac{\mu_G}{\mu_F}$

1: Set $\mathbf{x}^{\text{ag}}_{-\frac{1}{2}} \leftarrow \mathbf{x}_0$, $\mathbf{y}^{\text{ag}}_{-\frac{1}{2}} \leftarrow \mathbf{y}_0$, $\mathbf{x}^{\text{md}}_0 \leftarrow \mathbf{x}_0$, $\mathbf{y}^{\text{md}}_0 \leftarrow \mathbf{y}_0$

2: **for** $t = 1, 2, \dots, \mathscr{T}$ **do**

3: 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

4: $\mathbf{x}_{t-\frac{1}{2}} \leftarrow \mathbf{x}_{t-1} - \eta_t \left(\nabla f(\mathbf{x}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) - \mu_F (\mathbf{x}^{\text{md}}_{t-1} - \mathbf{x}_{t-1}) \right)$

5: $\mathbf{y}_{t-\frac{1}{2}} \leftarrow \mathbf{y}_{t-1} - \frac{\eta_t}{\mathscr{R}} \left(-\nabla_{\mathbf{y}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) + \nabla g(\mathbf{y}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) - \mu_G (\mathbf{y}^{\text{md}}_{t-1} - \mathbf{y}_{t-1}) \right)$

6: $\mathbf{x}^{\text{ag}}_{t-\frac{1}{2}} \leftarrow (1 - \alpha_t) \mathbf{x}^{\text{ag}}_{t-\frac{3}{2}} + \alpha_t \mathbf{x}_{t-\frac{1}{2}}$

7: $\mathbf{y}^{\text{ag}}_{t-\frac{1}{2}} \leftarrow (1 - \alpha_t) \mathbf{y}^{\text{ag}}_{t-\frac{3}{2}} + \alpha_t \mathbf{y}_{t-\frac{1}{2}}$

8: $\mathbf{x}_t \leftarrow \mathbf{x}_{t-1} - \eta_t \left(\nabla f(\mathbf{x}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) - \mu_F (\mathbf{x}^{\text{md}}_{t-1} - \mathbf{x}_{t-\frac{1}{2}}) \right)$

9: $\mathbf{y}_t \leftarrow \mathbf{y}_{t-1} - \frac{\eta_t}{\mathscr{R}} \left(-\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\mathbf{y}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) - \mu_G (\mathbf{y}^{\text{md}}_{t-1} - \mathbf{y}_{t-\frac{1}{2}}) \right)$

10: $\mathbf{x}^{\text{md}}_t \leftarrow (1 - \alpha_{t+1}) \mathbf{x}^{\text{ag}}_{t-\frac{1}{2}} + \alpha_{t+1} \mathbf{x}_t$

11: $\mathbf{y}^{\text{md}}_t \leftarrow (1 - \alpha_{t+1}) \mathbf{y}^{\text{ag}}_{t-\frac{1}{2}} + \alpha_{t+1} \mathbf{y}_t$

12: **end for**

13: **Output:** $[\mathbf{x}_{\mathscr{T}}; \mathbf{y}_{\mathscr{T}}]$

as well as $\eta_t = \frac{\alpha_t}{\mu_\star}$. *Then the iterates of Algorithm 2 satisfy, for all* $t = 1, \dots, \mathscr{T}$,

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}_t - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_t - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right] \\
&\le \left[\|\mathbf{x}_0 - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right] \left(\frac{L_{\text{Str}}}{\mu_\star} + 1\right) \prod_{\tau=1}^{t} (1 - \alpha_\tau) + \frac{3\sigma^2}{\mu_\star^2} \sum_{\tau=1}^{t} \alpha_\tau^2 \prod_{\tau'=\tau+1}^{t} (1 - \alpha_{\tau'})
\end{aligned}
\tag{17}
$$

where we employ the notation $\sigma = \frac{1}{\sqrt{3}} \sqrt{\frac{1}{1-r}\sigma^2_{\text{Str}} + (2 + \frac{1}{\beta})\sigma^2_{\text{Bil}}}$ *from Theorem 6.*

The proof of Theorem 9 is provided in §B.4. We highlight that our result applies to the output in Line 13 as $\mathbf{x}_{\mathscr{T}}, \mathbf{y}_{\mathscr{T}}$ instead of $\mathbf{x}^{\text{ag}}_{\mathscr{T}-\frac{1}{2}}, \mathbf{y}^{\text{ag}}_{\mathscr{T}-\frac{1}{2}}$ as in Line 14 of Algorithm 1. Additionally, let the total number of iterates $\mathscr{T} \ge 1$ be known in advance, and consider a constant stepsize $\alpha_t \equiv \alpha$. Optimizing the error bound over α gives

$$
\alpha = \alpha^\star = \frac{1}{\mathscr{T}} \left(1 + \log\left(\left[\|\mathbf{x}_0 - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right] \left(\frac{L_{\text{Str}}}{\mu_\star} + 1\right) \cdot \frac{\mu_\star^2 \mathscr{T}}{3\sigma^2}\right)\right) \wedge \bar{\alpha}(r, \beta)
$$

and hence (17) leads to the following:

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{x}_{\mathscr{T}} - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_{\mathscr{T}} - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right] &\le \left[\|\mathbf{x}_0 - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right] \left(\frac{L_{\text{Str}}}{\mu_\star} + 1\right) e^{-\bar{\alpha}(r,\beta)\mathscr{T}} \\
&\quad + \frac{3\sigma^2}{\mu_\star^2 \mathscr{T}} \left(1 + \log\left(\left[\|\mathbf{x}_0 - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2\right] \left(\frac{L_{\text{Str}}}{\mu_\star} + 1\right) \cdot \frac{\mu_\star^2 \mathscr{T}}{3\sigma^2}\right)\right)
\end{aligned}
$$

Prescribing the desired accuracy $\varepsilon > 0$, Theorem 9 shows that the iteration complexity to output an iterate $\mathbf{x}_{\mathscr{T}} \in \mathbb{R}^n$, $\mathbf{y}_{\mathscr{T}} \in \mathbb{R}^m$ that satisfies $\mathbb{E}[\|\mathbf{x}_{\mathscr{T}} - \boldsymbol{\omega}^\star_{\mathbf{x}}\|^2 + \mathscr{R}\|\mathbf{y}_{\mathscr{T}} - \boldsymbol{\omega}^\star_{\mathbf{y}}\|^2] \le \varepsilon^2$ is upper bounded by a constant multiple of[4]

$$\left(\sqrt{\frac{L_{\text{Str}}}{\mu_\star}}+\frac{L_{\text{Bil}}}{\mu_\star}+\frac{\sigma^2}{\mu_\star^2\varepsilon^2}\right)\log\left((\frac{L_{\text{Str}}}{\mu_\star}+1)\frac{1}{\varepsilon}\right)=\left(\sqrt{\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G}}+\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}+\frac{\sigma^2}{\mu_F^2\varepsilon^2}\right)\log\left((\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G})\frac{1}{\varepsilon}\right)$$

Compared to the stochastic AG-EG algorithm with restarting in Theorem 8, we see that there is a multiplicative $\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G}$ term inside the logarithmic factor. We believe that the extra logarithmic factor on the optimal statistical rate $\frac{\sigma^2}{\mu_\star^2\varepsilon^2}$ is removable using a proper diminishing stepsize strategy, a possibility that we reserve for future study.

Analogous to Theorem 7 in the case of no stochasticity, setting $r\to 1^-$, $\beta\to 0^+$ gives us the following convergence rate which matches the Zhang et al. (2022) lower bound:

Theorem 10 (Convergence of AG-EG, direct approach) *Suppose we are in the setting of Theorem 9 with no stochasticity. We have by choosing $\alpha_t\equiv\bar{\alpha}(1,0)$ defined as in (16) as well as $\eta_t\equiv\frac{\bar{\alpha}(1,0)}{\mu_\star}$, the output of Algorithm 2 satisfies*

$$\begin{aligned}&\|\mathbf{x}_{\mathscr{T}}-\boldsymbol{\omega}_{\mathbf{x}}^\star\|^2+\mathscr{R}\|\mathbf{y}_{\mathscr{T}}-\boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\\&\le\left[\|\mathbf{x}_0-\boldsymbol{\omega}_{\mathbf{x}}^\star\|^2+\mathscr{R}\|\mathbf{y}_0-\boldsymbol{\omega}_{\mathbf{y}}^\star\|^2\right]\left(\frac{L_{\text{Str}}}{\mu_\star}+1\right)\exp\left(-\frac{\mathscr{T}}{1+\sqrt{1+\frac{L_{\text{Str}}}{\mu_\star}+\frac{L_{\text{Bil}}^2}{\mu_\star^2}}}\right)\end{aligned}\tag{18}$$

We end this section by remarking that in the nonrandom Theorem 10, this upper bound on convergence rate indicates a near-unity coefficient on its condition-number exponent, yielding an iteration complexity that is asymptotically

$$\left(1+\sqrt{1+\frac{L_{\text{Str}}}{\mu_\star}+\frac{L_{\text{Bil}}^2}{\mu_\star^2}}\right)\log\left((\frac{L_{\text{Str}}}{\mu_\star}+1)\frac{1}{\varepsilon}\right)\sim\left(\sqrt{\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G}+\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}\right)\log\left((\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G})\frac{1}{\varepsilon}\right)$$

which is sharper (by at least $\exp(1)$) in its coefficient than the restarting iteration complexity result in Corollary 8 in §4. It should be noted that our argument is *essentially different from* Chen et al. (2017), and hence should *not* be viewed as a straightforward generalization of the argument therein, to the case of a strongly monotone operator. Indeed, adopting the arguments of Chen et al. (2017) necessarily requires a projection step, without which a scheduled restarting argument leads to an extra multiplicative logarithmic factor in condition number in its iteration complexity.

Implications for bilinear game case. It is worth noting that while the coefficient in complexity is superior in the strongly-convex-strongly-concave case, in the *bilinear game case* our direct approach in Algorithm 2 reduces to a *last-iterate* independent-sample stochastic extragradient (SEG) algorithm instead of *iteration averaged* SEG as in Algorithm 1. Consequently, the bias term cannot match the lower bound of Ibrahim et al. (2020). Moreover, in the stochastic case with noise variance bounded away from zero, Algorithm 2 can remain a constant distance from the saddle in expectation, resulting in *non-convergence behavior* (Hsieh et al., 2020, §3).

[4]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 (one query to $[\nabla f(\mathbf{x};\xi);\nabla g(\mathbf{y};\xi)]$ and two queries to $\nabla h(\mathbf{x},\mathbf{y};\zeta)$).

6. Discussion

We have presented a stochastic extragradient-based acceleration algorithm, AG-EG, for solving the bilinearly coupled saddle-point problem (1) that 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 both lower bounds have been met by a single algorithm. There are some remaining issues to be addressed, however, including the case of one-sided non-strong 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.

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. 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.

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.

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.

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.

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, Waïss 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. *arXiv preprint arXiv:2202.04640*, 2022.

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. *arXiv preprint arXiv:2112.15199*, 2021.

Guanghui Lan. *First-Order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

Chris Junchi Li, Yaodong Yu, Nicolas Loizou, Gauthier Gidel, Yi Ma, Nicolas Le Roux, and Michael I Jordan. On the convergence of stochastic extragradient for bilinear games using restarted iteration averaging. *arXiv preprint arXiv:2107.00464*, 2021.

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.

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 and Cheng Chen. Finding second-order stationary point for nonconvex-strongly-concave minimax problem. *arXiv preprint arXiv:2110.04814*, 2021.

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 *ICLR*, 2018.

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

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 descentGAN optimization is locally stable. *Advances in Neural Information Processing Systems*, 30, 2017.

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. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. In *Doklady Akademii Nauk USSR*, volume 269, pages 543–547, 1983.

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.

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.

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.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1), 1958.

Kiran Koshy Thekumparampil, Niao He, and Sewoong Oh. Lifted primal-dual method for bilinearly coupled smooth minimax optimization. *arXiv preprint arXiv:2201.07427*, 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.

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. *arXiv preprint arXiv:2006.06359*, 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.

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.

Appendix A. Related work

Here we compare our results with related work on saddle-point (minimax) optimization in machine learning and optimization literature.

Bilinear game case, nonstochastic setting. In the bilinear game case where $L_F = \mu_F = L_G = \mu_G = 0$, a lower bound has been established by Ibrahim et al. (2020): $\Omega\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}\log\left(\frac{1}{\varepsilon}\right)\right)$. The study of bilinear example has been initiated by Daskalakis et al. (2018) for understanding saddle-point optimization. They proposed the gradient descent-ascent (OGDA) algorithm and achieved sublinear convergence. Subsequently, the classical methods of ExtraGradient (EG) and Optimistic Gradient Descent Ascent (OGDA) algorithms were proven to have linear convergence rate for strongly monotone and Lipschitz operator with $O\left(\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\log(\frac{1}{\varepsilon})\right)$ iteration complexity (Gidel et al., 2019b; Mokhtari et al., 2020). Azizian et al. (2020a) proved in another study that by considering first order methods using a fixed number of composed gradient evaluations and only the last iteration (this class of methods is called 1-SCLI and excludes momentum and restarting), the $O\left(\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\log(\frac{1}{\varepsilon})\right)$ iteration complexity for EG is optimal. In the absence of strong monotonicity assumptions, Loizou et al. (2020) generated the first set of global non-asymptotic last-iterate convergence guarantees for a stochastic game over a non-compact domain using a Hamiltonian viewpoint. In particular, the proposed stochastic Hamiltonian gradient method ensures convergence in the finite-sum stochastic bilinear game as well. In very recent work, when restricted to the bilinear minimax optimization, Kovalev et al. (2021) derive an iteration complexity that is essentially $\mathcal{O}\left(\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}\log(\frac{1}{\varepsilon})\right)$. 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). For matching the $\mathcal{O}\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}\log(\frac{1}{\varepsilon})\right)$ lower bound provided by Ibrahim et al. (2020), the work of Azizian et al. (2020b) considered EG with momentum. They used a perturbed spectral analysis encompassing Polyak momentum. Nonetheless, Azizian et al. (2020b) only provide accelerated rates in the regime where the condition number is large. Li et al. (2021) was the first to show that a version of stochastic extragradient method converges at an accelerated convergence rates for bilinear games with unbounded domain and unbounded stochastic noise using restarted iteration averaging, and when focusing on the nonstochastic setting, matches the lower bound of (Ibrahim et al., 2020).

Smooth strongly convex-concave case, nonstochastic setting. Lower bound studies has been recently conducted, represented by Ouyang and Xu (2021) for smooth convex-concave minimax optimization, and analogously by Zhang et al. (2022) for strongly-convex-strongly-concave objectives. The latter lower bound is of the form $\Omega\left(\left(\sqrt{\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G}}+\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$. As for upper bounds, earlier extragradient-based methods Tseng (1995) and accelerated dual extrapolation algorithm Nesterov and Scrimali (2011) achieve, when translated to our bilinearly coupled problem, an iteration complexity of $\tilde{\mathcal{O}}\left(\frac{L_F}{\mu_F}\vee\frac{L_G}{\mu_G}+\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}\right)$. The same complexity has also been matched by Gidel et al. (2019a), Mokhtari et al. (2020), Cohen et al. (2021) from a relative Lipschitz viewpoint.[5] Improving upon this result, Lin et al. (2020b) achieve a complexity of $\tilde{\mathcal{O}}\left(\sqrt{\frac{L_F L_G}{\mu_F\mu_G}}+\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}{\mu_F\mu_G}}\right)$

[5] Mokhtari et al. (2020) report a $\tilde{\mathcal{O}}\left(\frac{L_F\vee L_G+\sqrt{\lambda_{\max}(\mathbf{B}^\top\mathbf{B})}}{\mu_F\wedge\mu_G}\right)$ 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 $\sqrt{\frac{\mu_G}{\mu_F}}$.

References	Iteration Complexity
Mokhtari et al. (2020); Cohen et al. (2021)	$\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$
Lin et al. (2020b)	$\sqrt{\frac{L_F L_G}{\mu_F \mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$
Wang and Li (2020)	$\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt[4]{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G} \cdot \frac{L_F L_G}{\mu_F \mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$
Xie et al. (2021)	$\sqrt[4]{\frac{L_F L_G}{\mu_F \mu_G} \left(\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}\right)} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$
Kovalev et al. (2021) and concurrently Thekumparampil et al. (2022); Jin et al. (2022)	$\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$
AG-EG (this work), Theorem 7	$\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$
AG-EG-Direct (this work), Theorem 10	$\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}$
Zhang et al. (2022) (lower bound)	$\Omega\left(\left(\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}\right) \log\left(\frac{1}{\varepsilon}\right)\right)$
Gidel et al. (2019b) among other work	$\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$
Azizian et al. (2020b); Li et al. (2021)	$\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}$
AG-EG (this work), Corollary 5	$\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}}$
Ibrahim et al. (2020) (lower bound)	$\Omega\left(\sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}} \log\left(\frac{1}{\varepsilon}\right)\right)$

Table 1: Table of comparison with related work for both strongly case and bilinear case, concentrating on the nonstochastic setting. For upper bounds, a polylogarithmic prefactor is ignored.

using proper acceleration methods, when restricted to the bilinearly coupled problem. Wang and Li (2020) achieves[6] $\tilde{\mathcal{O}}\left(\sqrt{\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}} + \sqrt[4]{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G} \cdot \frac{L_F L_G}{\mu_F \mu_G}} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}\right)$ and a Hermitian-skew-based analysis nearly matches Zhang et al. (2022) for the quadratic minimax game case. For the same problem, Xie et al. (2021) achieves a complexity of $\tilde{\mathcal{O}}\left(\sqrt[4]{\frac{L_F L_G}{\mu_F \mu_G} \left(\frac{L_F}{\mu_F} \vee \frac{L_G}{\mu_G}\right)} + \sqrt{\frac{\lambda_{\max}(\mathbf{B}^\top \mathbf{B})}{\mu_F \mu_G}}\right)$. 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. (2021); Thekumparampil et al. (2022); Jin et al. (2022) studies the nonstochastic 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 term up to a constant prefactor in the stochastic setting.

Stochastic setting. Stochastic minimax optimization has been studied intensively as a special case of the variational inequalities. It is widely accepted in classical literature on stochastic variational inequality (Nemirovski et al., 2009; Juditsky et al., 2011) that the set of parameters and the variance

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

of the stochastic estimate of the vector field are bounded. Chen et al. (2017) extended the analysis of Juditsky et al. (2011) that accelerates the convergence rates for a class of variational 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 SEG to a strictly coherent solution (a.k.a. star-strict monotone variational inequality problem). In a similar vein, Ryu et al. (2019) showed that SGDA with anchoring almost surely converge to strictlyconvex-concave saddle points. Fallah et al. (2020) developed a multistage variant of stochastic gradient descent ascent 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 absence of strong convex-concavity. They achieved the optimal dependency on the noise variance but suboptimal dependency on the condition number. Hsieh et al. (2020) developed a double stepsize 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 models, these two assumptions do not hold in this case because the noise increases with the value of the parameters. In recent work, Mishchenko et al. (2020) has shown that stochastic extragradients can be computed under a different stepsize, which removes the bounded domain assumption, while still requiring the bounded noise assumption. The work also discussed the advantages of using the same mini-batch for the two gradients in stochastic extragradients. In another vein, Jelassi et al. (2020) focuses on stochastic extragradient in games with a large number of players. In that case they propose an extragradient algorithm that randomly update a small subset of the players at each iterations. Yan et al. (2019, 2020); Rafique et al. (2021) studied the non-smooth settings and obtained fast rates. More recent works consider minimax optimization problems without covexity and/or concavity where the goal is to find first-order and second-order stationary points (Sebbouh et al., 2022; Lin et al., 2020a; Yang et al., 2022; Guo et al., 2020; Chen et al., 2021; Luo and Chen, 2021). One interesting direction is to extend our algorithm to their settings and obtain a fine-grained complexity bound with optimal rates.

Appendix B. Proofs of main results

In this section we present the proofs of our main results. §B.1 illustrates the scaling reduction argument. §B.2 provides auxiliary lemmas. With a slight adjustment of their presentation order §B.3 proves Theorem 6, §B.4 proves Theorem 9 and finally §B.5 proves Theorem 4.

B.1. Scaling reduction argument

Here we illustrate the scaling reduction argument that reduces our analysis of our AG-EG Algorithm 1 to the one with equal strong-convexity parameters of F and G using a reparametrized 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 (1) we consider

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

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

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

and

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

It is straightforward to verify $\hat{\mathscr{F}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ is arguably μ_{Str}-strongly-convex-μ_{Str}-strongly-concave. The essence of our update rules is captured by 8 lines corresponding to Lines 5–12 in Algorithm 1, which becomes:

$$\begin{aligned}
\hat{\mathbf{x}}_{t-\frac{1}{2}} &= \hat{\mathbf{x}}_{t-1} - \eta_t \left(\nabla f(\hat{\mathbf{x}}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\hat{\mathbf{x}}} h(\hat{\mathbf{x}}_{t-1}, \hat{\mathbf{y}}_{t-1}; \zeta_{t-\frac{1}{2}}) \right), && \text{(19a)} \\
\hat{\mathbf{y}}_{t-\frac{1}{2}} &= \hat{\mathbf{y}}_{t-1} - \eta_t \left(-\nabla_{\hat{\mathbf{y}}} h(\hat{\mathbf{x}}_{t-1}, \hat{\mathbf{y}}_{t-1}; \zeta_{t-\frac{1}{2}}) + \nabla g(\hat{\mathbf{y}}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) \right), && \text{(19b)} \\
\hat{\mathbf{x}}^{\mathrm{ag}}_{t-\frac{1}{2}} &= (1-\alpha_t) \hat{\mathbf{x}}^{\mathrm{ag}}_{t-\frac{3}{2}} + \alpha_t \hat{\mathbf{x}}_{t-\frac{1}{2}}, && \text{(19c)} \\
\hat{\mathbf{y}}^{\mathrm{ag}}_{t-\frac{1}{2}} &= (1-\alpha_t) \hat{\mathbf{y}}^{\mathrm{ag}}_{t-\frac{3}{2}} + \alpha_t \hat{\mathbf{y}}_{t-\frac{1}{2}}, && \text{(19d)} \\
\hat{\mathbf{x}}_t &= \hat{\mathbf{x}}_{t-1} - \eta_t \left(\nabla f(\hat{\mathbf{x}}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\hat{\mathbf{x}}} h(\hat{\mathbf{x}}_{t-\frac{1}{2}}, \hat{\mathbf{y}}_{t-\frac{1}{2}}; \zeta_t) \right), && \text{(19e)} \\
\hat{\mathbf{y}}_t &= \hat{\mathbf{y}}_{t-1} - \eta_t \left(-\nabla_{\hat{\mathbf{y}}} h(\hat{\mathbf{x}}_{t-\frac{1}{2}}, \hat{\mathbf{y}}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\hat{\mathbf{y}}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) \right), && \text{(19f)} \\
\hat{\mathbf{x}}^{\mathrm{md}}_t &= (1-\alpha_{t+1}) \hat{\mathbf{x}}^{\mathrm{ag}}_{t-\frac{1}{2}} + \alpha_{t+1} \hat{\mathbf{x}}_t, && \text{(19g)} \\
\hat{\mathbf{y}}^{\mathrm{md}}_t &= (1-\alpha_{t+1}) \hat{\mathbf{y}}^{\mathrm{ag}}_{t-\frac{1}{2}} + \alpha_{t+1} \hat{\mathbf{y}}_t. && \text{(19h)}
\end{aligned}$$

It is obvious to translate Eqs. (19c), (19d), (19g), (19h) into Lines 7, 8, 11, 12, separately. The rest translations are also straightforward, represented by Eqs. (19a) into Line 5

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

as well as Eqs. (19f) into Line 10

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

It is also straightforward to justify that Assumptions 1, 2 and 3 are rediscovered by reverting the scaling reduction from $\hat{\mathscr{F}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ to $\mathscr{F}(\mathbf{x}, \mathbf{y})$. Therefore, it suffices to analyze Algorithm 1 for $\hat{\mathscr{F}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ and due to this scaling reduction, we only need to prove all results for the case of $\mathscr{R} = 1$. To keep the notations simple, till the rest of this work we slightly abuse the notations and remove the hats in all symbols.

B.2. Auxiliary lemmas

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

Lemma 11 *Given $\boldsymbol{\theta}, \boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2 \in \mathbb{R}^d$ and also $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2$ that satisfies*

$$\boldsymbol{\varphi}_1 = \boldsymbol{\theta} - \boldsymbol{\delta}_1, \qquad \boldsymbol{\varphi}_2 = \boldsymbol{\theta} - \boldsymbol{\delta}_2 \tag{20}$$

then for any $\mathbf{z} \in \mathbb{R}^d$ we have

$$\langle \boldsymbol{\delta}_2, \boldsymbol{\varphi}_1 - \mathbf{z} \rangle \le \frac{1}{2}\|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\|^2 + \frac{1}{2}\left[\|\boldsymbol{\theta} - \mathbf{z}\|^2 - \|\boldsymbol{\varphi}_2 - \mathbf{z}\|^2 - \|\boldsymbol{\theta} - \boldsymbol{\varphi}_1\|^2\right] \tag{21}$$

Proof of Lemma 11 is provided in §C.1. Lemma 11 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 6 and 9 we first introduce some notations. Let $\tilde{\mathbf{x}} \in \mathbb{R}^n, \tilde{\mathbf{y}} \in \mathbb{R}^m$ and let the *pointwise primal-dual gap function* be

$$V(\mathbf{x}, \mathbf{y} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = F(\mathbf{x}) - F(\tilde{\mathbf{x}}) + G(\mathbf{y}) - G(\tilde{\mathbf{y}}) + \langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x} - \tilde{\mathbf{x}} \rangle + \langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y} - \tilde{\mathbf{y}} \rangle \tag{22}$$

and they can be separated $V(\mathbf{x}, \mathbf{y} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = V_F(\mathbf{x} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) + V_G(\mathbf{y} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ defined as

$$\begin{aligned} V_F(\mathbf{x} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) &= F(\mathbf{x}) - F(\tilde{\mathbf{x}}) + \langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x} - \tilde{\mathbf{x}} \rangle \\ V_G(\mathbf{y} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) &= G(\mathbf{y}) - G(\tilde{\mathbf{y}}) + \langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y} - \tilde{\mathbf{y}} \rangle \end{aligned} \tag{23}$$

We prove that either of these two quantities is lower-bounded by a positive quadratic:

Lemma 12 *We have both $F(\mathbf{x})$ and $G(\mathbf{y})$ are L_{Str}-smooth and μ_{Str}-strongly convex. Furthermore, for any $\mathbf{x} \in \mathbb{R}^n$ we have*

$$V_F(\mathbf{x} \mid \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star) = F(\mathbf{x}) - F(\boldsymbol{\omega}_{\mathbf{x}}^\star) + \left\langle \nabla_{\mathbf{x}} H(\boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star), \mathbf{x} - \boldsymbol{\omega}_{\mathbf{x}}^\star \right\rangle \ge \frac{\mu_{\mathrm{Str}}}{2}\left\|\mathbf{x} - \boldsymbol{\omega}_{\mathbf{x}}^\star\right\|^2 \tag{24}$$

and for any $\mathbf{y} \in \mathbb{R}^m$

$$V_G(\mathbf{y} \mid \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star) = G(\mathbf{y}) - G(\boldsymbol{\omega}_{\mathbf{y}}^\star) - \left\langle \nabla_{\mathbf{y}} H(\boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star), \mathbf{y} - \boldsymbol{\omega}_{\mathbf{y}}^\star \right\rangle \ge \frac{\mu_{\mathrm{Str}}}{2}\left\|\mathbf{y} - \boldsymbol{\omega}_{\mathbf{y}}^\star\right\|^2 \tag{25}$$

where these two V-quantities are defined as in (23).

Proof of Lemma 12 is provided in §C.2. Our final auxiliary lemma on the key properties on stepsizes spells as follows:

Lemma 13 *Set $\square \equiv \frac{\tilde{\sigma}[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}{\mathscr{C}\sqrt{\mathbb{E}[\|\mathbf{x}_0 - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \mathscr{R}\|\mathbf{y}_0 - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2]}}$. Our stepsize choice (9) satisfies (i) $\eta_t \le \frac{t}{\square}$; (ii) $\left(\frac{t}{\eta_t} : t \ge 1\right)$ is a nonnegative, nondecreasing arithmetic sequence with common difference $\sqrt{\frac{1+\beta}{r}}L_{\mathrm{Bil}}$; (iii) $L_{\mathrm{Bil}}\eta_t \le 1$, and (iv) the stepsize condition*

$$r - \frac{2L_{\mathrm{Str}}}{t+1}\eta_t - (1+\beta)L_{\mathrm{Bil}}^2\eta_t^2 \ge 0 \tag{26}$$

Proof of Lemma 13 is provided in §C.3.

B.3. Proof of Theorem 6

Throughout the proof we assume $L_{\text{Str}} + L_{\text{Bil}} > 0$ without loss of generality, since otherwise the result holds trivially. Due to the scaling reduction argument in §B.1, we assume without loss of generality that $\mathscr{R} = 1$.

We introduce some notations. Denote the (squared) metric by $\mathcal{S}(\mathbf{x}, \mathbf{y}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \equiv \|\mathbf{x} - \tilde{\mathbf{x}}\|^2 + \|\mathbf{y} - \tilde{\mathbf{y}}\|^2$, and denote the incurred stochastic noise terms

$$\boldsymbol{\Delta}_{\text{Str}}^{t-\frac{1}{2}} \equiv \begin{bmatrix} \nabla f(\mathbf{x}_{t-1}^{\text{md}}; \xi_{t-\frac{1}{2}}) - \nabla F(\mathbf{x}_{t-1}^{\text{md}}) \\ \nabla g(\mathbf{y}_{t-1}^{\text{md}}; \xi_{t-\frac{1}{2}}) - \nabla G(\mathbf{y}_{t-1}^{\text{md}}) \end{bmatrix}, \quad \boldsymbol{\Delta}_{\text{Bil}}^{t-\frac{1}{2}} \equiv \begin{bmatrix} \nabla_{\mathbf{x}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) - \nabla_{\mathbf{x}} H(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \\ -\nabla_{\mathbf{y}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) + \nabla_{\mathbf{y}} H(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \end{bmatrix}$$
$$\boldsymbol{\Delta}_{\text{Bil}}^{t} \equiv \begin{bmatrix} \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) - \nabla_{\mathbf{x}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}) \\ -\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla_{\mathbf{y}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}) \end{bmatrix} \tag{27}$$

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. We prove the following lemma:

Lemma 14 *For arbitrary $\tilde{\mathbf{x}} \in \mathbb{R}^n, \tilde{\mathbf{y}} \in \mathbb{R}^m$ and $\alpha_t \in (0, 1]$ the iterates of Algorithm 1 ($\mathscr{S} = 1$) satisfy for $t = 1, \ldots, \mathscr{T}$, almost surely*

$$\begin{aligned} &V(\mathbf{x}_{t-\frac{1}{2}}^{\text{ag}}, \mathbf{y}_{t-\frac{1}{2}}^{\text{ag}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - (1 - \alpha_t) V(\mathbf{x}_{t-\frac{3}{2}}^{\text{ag}}, \mathbf{y}_{t-\frac{3}{2}}^{\text{ag}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \\ &\le \alpha_t \langle \nabla F(\mathbf{x}_{t-1}^{\text{md}}) + \nabla_{\mathbf{x}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle + \alpha_t \langle -\nabla_{\mathbf{y}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}) + \nabla G(\mathbf{y}_{t-1}^{\text{md}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} \rangle \\ &\quad + \frac{\alpha_t^2 L_{\text{Str}}}{2} \mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1}) \end{aligned} \tag{28}$$

Note the proof only relies on the interpolation updates in our algorithm as in Lines 7, 8, 11 and 12, and hence this result holds in a per-trajectory (almost-sure) fashion.

Proof [Proof of Lemma 14] From the convexity and L_{Str}-smoothness of F, we know that for arbitrary $\tilde{\mathbf{x}}, \tilde{\mathbf{y}}$

$$\begin{aligned} F(\mathbf{x}_{t-\frac{1}{2}}^{\text{ag}}) - F(\tilde{\mathbf{x}}) &= F(\mathbf{x}_{t-\frac{1}{2}}^{\text{ag}}) - F(\mathbf{x}_{t-1}^{\text{md}}) - \left(F(\tilde{\mathbf{x}}) - F(\mathbf{x}_{t-1}^{\text{md}})\right) \\ &\le \langle \nabla F(\mathbf{x}_{t-1}^{\text{md}}), \mathbf{x}_{t-\frac{1}{2}}^{\text{ag}} - \mathbf{x}_{t-1}^{\text{md}} \rangle + \frac{L_{\text{Str}}}{2} \left\| \mathbf{x}_{t-\frac{1}{2}}^{\text{ag}} - \mathbf{x}_{t-1}^{\text{md}} \right\|^2 - \langle \nabla F(\mathbf{x}_{t-1}^{\text{md}}), \tilde{\mathbf{x}} - \mathbf{x}_{t-1}^{\text{md}} \rangle \end{aligned}$$

Taking $\tilde{\mathbf{x}} = \mathbf{x}_{t-\frac{3}{2}}^{\text{ag}}$ in the above inequality, we have

$$\begin{aligned} F(\mathbf{x}_{t-\frac{1}{2}}^{\text{ag}}) - F(\mathbf{x}_{t-\frac{3}{2}}^{\text{ag}}) &= F(\mathbf{x}_{t-\frac{1}{2}}^{\text{ag}}) - F(\mathbf{x}_{t-1}^{\text{md}}) - \left(F(\mathbf{x}_{t-\frac{3}{2}}^{\text{ag}}) - F(\mathbf{x}_{t-1}^{\text{md}})\right) \\ &\le \langle \nabla F(\mathbf{x}_{t-1}^{\text{md}}), \mathbf{x}_{t-\frac{1}{2}}^{\text{ag}} - \mathbf{x}_{t-1}^{\text{md}} \rangle + \frac{L_{\text{Str}}}{2} \left\| \mathbf{x}_{t-\frac{1}{2}}^{\text{ag}} - \mathbf{x}_{t-1}^{\text{md}} \right\|^2 - \langle \nabla F(\mathbf{x}_{t-1}^{\text{md}}), \mathbf{x}_{t-\frac{3}{2}}^{\text{ag}} - \mathbf{x}_{t-1}^{\text{md}} \rangle \end{aligned}$$

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

$$
\begin{aligned}
&F(\mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}}) - (1-\alpha_t)F(\mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}}) - \alpha_t F(\tilde{\mathbf{x}}) \\
&\le \langle \nabla F(\mathbf{x}^{\mathrm{md}}_{t-1}), \mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}} - \mathbf{x}^{\mathrm{md}}_{t-1}\rangle + \tfrac{L_{\mathrm{Str}}}{2}\left\|\mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}} - \mathbf{x}^{\mathrm{md}}_{t-1}\right\|^2 - \langle \nabla F(\mathbf{x}^{\mathrm{md}}_{t-1}), (1-\alpha_t)\mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}} + \alpha_t\tilde{\mathbf{x}} - \mathbf{x}^{\mathrm{md}}_{t-1}\rangle \\
&\le \langle \nabla F(\mathbf{x}^{\mathrm{md}}_{t-1}), \alpha_t(\mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1})\rangle + \tfrac{L_{\mathrm{Str}}}{2}\|\alpha_t(\mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1})\|^2 - \langle \nabla F(\mathbf{x}^{\mathrm{md}}_{t-1}), \alpha_t(\tilde{\mathbf{x}} - \mathbf{x}_{t-1})\rangle \\
&= \alpha_t\langle \nabla F(\mathbf{x}^{\mathrm{md}}_{t-1}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle + \tfrac{\alpha_t^2 L_{\mathrm{Str}}}{2}\|\mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\|^2
\end{aligned}
\tag{29}
$$

where we applied the fact from our update rules that $\mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}} - \mathbf{x}^{\mathrm{md}}_{t-1} = \alpha_t(\mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1})$. Following an analogous argument for G we obtain

$$
G(\mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}}) - (1-\alpha_t)G(\mathbf{y}^{\mathrm{ag}}_{t-\frac{3}{2}}) - \alpha_t G(\tilde{\mathbf{y}}) \le \alpha_t\langle \nabla G(\mathbf{y}^{\mathrm{md}}_{t-1}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle + \tfrac{\alpha_t^2 L_{\mathrm{Str}}}{2}\|\mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\|^2
\tag{30}
$$

On the other hand, due to Lines 7 and 8 we have

$$
\begin{aligned}
&\langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle - (1-\alpha_t)\langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\mathbf{x}}\rangle \\
&\quad = \langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} - (1-\alpha_t)(\mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}} - \mathbf{x})\rangle = \alpha_t\langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle
\end{aligned}
$$

and analogously

$$
\begin{aligned}
&\langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle - (1-\alpha_t)\langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\mathbf{y}}\rangle \\
&\quad = \langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} - (1-\alpha_t)(\mathbf{y}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\mathbf{y}})\rangle = \alpha_t\langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle
\end{aligned}
$$

Due to our assumption on H we have

$$
\begin{aligned}
&\langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle + \langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle \\
&\quad \le \langle \nabla_{\mathbf{x}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle + \langle -\nabla_{\mathbf{y}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle
\end{aligned}
$$

Combining the above three displays together yields

$$
\begin{aligned}
&\langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle - (1-\alpha_t)\langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\mathbf{x}}\rangle \\
&\qquad\qquad + \langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle - (1-\alpha_t)\langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}^{\mathrm{ag}}_{t-\frac{3}{2}} - \tilde{\mathbf{y}}\rangle \\
&\le \alpha_t\left[\langle \nabla_{\mathbf{x}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle + \langle -\nabla_{\mathbf{y}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle\right]
\end{aligned}
\tag{31}
$$

Now, summing up Eqs. (29), (30) and (31) and noting the definition of V in (22), we have

$$
\begin{aligned}
&V(\mathbf{x}^{\text{ag}}_{t-\frac{1}{2}}, \mathbf{y}^{\text{ag}}_{t-\frac{1}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - (1-\alpha_t) V(\mathbf{x}^{\text{ag}}_{t-\frac{3}{2}}, \mathbf{y}^{\text{ag}}_{t-\frac{3}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \\
&= F(\mathbf{x}^{\text{ag}}_{t-\frac{1}{2}}) - (1-\alpha_t) F(\mathbf{x}^{\text{ag}}_{t-\frac{3}{2}}) - \alpha_t F(\tilde{\mathbf{x}}) + G(\mathbf{y}^{\text{ag}}_{t-\frac{1}{2}}) - (1-\alpha_t) G(\mathbf{y}^{\text{ag}}_{t-\frac{3}{2}}) - \alpha_t G(\tilde{\mathbf{y}}) \\
&\quad + \langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}^{\text{ag}}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle - (1-\alpha_t) \langle \nabla_{\mathbf{x}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}^{\text{ag}}_{t-\frac{3}{2}} - \tilde{\mathbf{x}} \rangle \\
&\qquad\qquad + \langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}^{\text{ag}}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} \rangle - (1-\alpha_t) \langle -\nabla_{\mathbf{y}} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}^{\text{ag}}_{t-\frac{3}{2}} - \tilde{\mathbf{y}} \rangle \\
&\le \alpha_t \left[\langle \nabla F(\mathbf{x}^{\text{md}}_{t-1}) + \nabla_{\mathbf{x}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle + \langle -\nabla_{\mathbf{y}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}) + \nabla G(\mathbf{y}^{\text{md}}_{t-1}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} \rangle \right] \\
&\quad + \frac{\alpha_t^2 L_{\text{Str}}}{2} \left[\|\mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\|^2 + \|\mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\|^2 \right]
\end{aligned}
$$

and hence conclude (28) and Lemma 14. ■

Step 2. We target to prove, for our choice of η_t that satisfies, for a given $r \in (0,1)$, (26) of Lemma 13(iv) that $r - \frac{2L_{\text{Str}}}{t+1}\eta_t - (1+\beta)L_{\text{Bil}}^2\eta_t^2 \ge 0$ we have for any $\tilde{\mathbf{x}} \in \mathbb{R}^n$, $\tilde{\mathbf{y}} \in \mathbb{R}^m$ and $\mathcal{T} = 1, \ldots, \mathscr{T}$ that

$$
\begin{aligned}
&\mathcal{T}(\mathcal{T}+1)\mathbb{E}[V(\mathbf{x}^{\text{ag}}_{\mathcal{T}-\frac{1}{2}}, \mathbf{y}^{\text{ag}}_{\mathcal{T}-\frac{1}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] + \frac{\mathcal{T}}{\eta_{\mathcal{T}}} \mathbb{E}[\mathcal{S}(\mathbf{x}_{\mathcal{T}}, \mathbf{y}_{\mathcal{T}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le \frac{1}{\eta_1} \mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] + \sqrt{\tfrac{1+\beta}{r}} L_{\text{Bil}} \sum_{t=2}^{\mathcal{T}} \mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\quad + \frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma \sqrt{\mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]}
\end{aligned}
\tag{32}
$$

To bound the inner-product terms in (28), by setting $\boldsymbol{\varphi}_1 = \mathbf{x}_{t-\frac{1}{2}}$, $\boldsymbol{\theta} = \mathbf{x}_{t-1}$, $\boldsymbol{\varphi}_2 = \mathbf{x}_t$, $\boldsymbol{\delta}_1 = \eta_t\Big(\nabla f(\mathbf{x}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}})\Big)$, $\boldsymbol{\delta}_2 = \eta_t\Big(\nabla f(\mathbf{x}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t)\Big)$ as in Lemma 11 (with $\mathbf{z} = \tilde{\mathbf{x}}$), we have

$$
\begin{aligned}
&\eta_t \langle \nabla f(\mathbf{x}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle \\
&\le \frac{1}{2} \left(\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}\|^2 - \|\mathbf{x}_t - \tilde{\mathbf{x}}\|^2 - \|\mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\|^2 \right) \\
&\quad + \frac{\eta_t^2}{2} \|\nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) - \nabla_{\mathbf{x}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}})\|^2
\end{aligned}
$$

where Young's inequality combined with the martingale structure yields (also noting (27))

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) - \nabla_{\mathbf{x}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}})\|^2 \\
&= \mathbb{E}\|\nabla_{\mathbf{x}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}) - \nabla_{\mathbf{x}} H(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) - \boldsymbol{\Delta}^{1,t-\frac{1}{2}}_{\text{Bil}}\|^2 + \mathbb{E}\|\boldsymbol{\Delta}^{1,t}_{\text{Bil}}\|^2 \\
&\le (1+\beta) L_{\text{Bil}}^2 \mathbb{E}\|\mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\|^2 + (1+\tfrac{1}{\beta}) \mathbb{E}\|\boldsymbol{\Delta}^{1,t-\frac{1}{2}}_{\text{Bil}}\|^2 + \mathbb{E}\|\boldsymbol{\Delta}^{1,t}_{\text{Bil}}\|^2
\end{aligned}
$$

Combining the above two displays with expectation taken gives

$$
\begin{aligned}
&\eta_t \mathbb{E}\langle \nabla f(\mathbf{x}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle \\
&\le \frac{1}{2}\left(\mathbb{E}\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}\|^2 - \mathbb{E}\|\mathbf{x}_t - \tilde{\mathbf{x}}\|^2 - \mathbb{E}\|\mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\|^2 \right) \\
&\quad + \frac{\eta_t^2}{2}\left((1+\beta) L_{\mathrm{Bil}}^2 \mathbb{E}\|\mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\|^2 + (1+\tfrac{1}{\beta}) \mathbb{E}\|\boldsymbol{\Delta}_{\mathrm{Bil}}^{1,t-\frac{1}{2}}\|^2 + \mathbb{E}\|\boldsymbol{\Delta}_{\mathrm{Bil}}^{1,t}\|^2 \right)
\end{aligned}
$$

Analogously by setting the appropriate parameters, we have

$$
\begin{aligned}
&\eta_t \mathbb{E}\langle -\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\mathbf{y}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} \rangle \\
&\le \frac{1}{2}\left(\mathbb{E}\|\mathbf{y}_{t-1} - \tilde{\mathbf{y}}\|^2 - \mathbb{E}\|\mathbf{y}_t - \tilde{\mathbf{y}}\|^2 - \mathbb{E}\|\mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\|^2 \right) \\
&\quad + \frac{\eta_t^2}{2}\left((1+\beta) L_{\mathrm{Bil}}^2 \mathbb{E}\|\mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\|^2 + (1+\tfrac{1}{\beta}) \mathbb{E}\|\boldsymbol{\Delta}_{\mathrm{Bil}}^{2,t-\frac{1}{2}}\|^2 + \mathbb{E}\|\boldsymbol{\Delta}_{\mathrm{Bil}}^{2,t}\|^2 \right)
\end{aligned}
$$

Combining the last two displays gives

$$
\begin{aligned}
&\eta_t \mathbb{E}\langle \nabla f(\mathbf{x}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle \\
&\qquad\qquad + \eta_t \mathbb{E}\langle -\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\mathbf{y}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} \rangle \\
&\le \frac{1}{2}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \right) - \frac{1 - (1+\beta) L_{\mathrm{Bil}}^2 \eta_t^2}{2} \mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad + \frac{\eta_t^2}{2}\left((1+\tfrac{1}{\beta}) \mathbb{E}[\|\boldsymbol{\Delta}_{\mathrm{Bil}}^{1,t-\frac{1}{2}}\|^2 + \|\boldsymbol{\Delta}_{\mathrm{Bil}}^{2,t-\frac{1}{2}}\|^2] + \mathbb{E}[\|\boldsymbol{\Delta}_{\mathrm{Bil}}^{1,t}\|^2 + \|\boldsymbol{\Delta}_{\mathrm{Bil}}^{2,t}\|^2] \right) \qquad (33)
\end{aligned}
$$

Therefore plugging the above (33) into (28) of Lemma 14 with $\alpha_t = \frac{2}{t+1}$ and expectation taken, we have from (23) that

$$
\begin{aligned}
&\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}}, \mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \frac{t-1}{t+1}\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}}, \mathbf{y}^{\mathrm{ag}}_{t-\frac{3}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le \frac{2}{t+1}\mathbb{E}\langle \nabla F(\mathbf{x}^{\mathrm{md}}_{t-1}) + \nabla_{\mathbf{x}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle \\
&\quad + \frac{2}{t+1}\mathbb{E}\langle -\nabla_{\mathbf{y}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}) + \nabla G(\mathbf{y}^{\mathrm{md}}_{t-1}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle + \frac{2L_{\mathrm{Str}}}{(t+1)^2}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&= \frac{2}{t+1}\mathbb{E}\langle \nabla f(\mathbf{x}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle \\
&\quad + \frac{2}{t+1}\mathbb{E}\langle -\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\mathbf{y}^{\mathrm{md}}_{t-1}; \xi_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle + \frac{2L_{\mathrm{Str}}}{(t+1)^2}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad - \frac{2}{t+1}\mathbb{E}\langle \boldsymbol{\Delta}^{1,t-\frac{1}{2}}_{\mathrm{Str}} + \boldsymbol{\Delta}^{1,t}_{\mathrm{Bil}}, \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle - \frac{2}{t+1}\mathbb{E}\langle \boldsymbol{\Delta}^{2,t-\frac{1}{2}}_{\mathrm{Str}} + \boldsymbol{\Delta}^{2,t}_{\mathrm{Bil}}, \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle \\
&\le \frac{1}{(t+1)\eta_t}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) \\
&\quad - \frac{1}{(t+1)\eta_t}\left(1 - \frac{2L_{\mathrm{Str}}}{t+1}\eta_t - (1+\beta)L_{\mathrm{Bil}}^2\eta_t^2\right)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad + \frac{\eta_t}{t+1}\left((1+\tfrac{1}{\beta})\mathbb{E}\|\boldsymbol{\Delta}^{1,t-\frac{1}{2}}_{\mathrm{Bil}}\|^2 + \mathbb{E}\|\boldsymbol{\Delta}^{1,t}_{\mathrm{Bil}}\|^2\right) + \frac{\eta_t}{t+1}\left((1+\tfrac{1}{\beta})\mathbb{E}\|\boldsymbol{\Delta}^{2,t-\frac{1}{2}}_{\mathrm{Bil}}\|^2 + \mathbb{E}\|\boldsymbol{\Delta}^{2,t}_{\mathrm{Bil}}\|^2\right) \\
&\quad - \frac{2}{t+1}\mathbb{E}\langle \boldsymbol{\Delta}^{1,t-\frac{1}{2}}_{\mathrm{Str}} + \boldsymbol{\Delta}^{1,t}_{\mathrm{Bil}}, \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle - \frac{2}{t+1}\mathbb{E}\langle \boldsymbol{\Delta}^{2,t-\frac{1}{2}}_{\mathrm{Str}} + \boldsymbol{\Delta}^{2,t}_{\mathrm{Bil}}, \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle
\end{aligned}
$$

With some manipulations we obtain

$$
\begin{aligned}
&\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}}, \mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \frac{t-1}{t+1}\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}}, \mathbf{y}^{\mathrm{ag}}_{t-\frac{3}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le \frac{1}{(t+1)\eta_t}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) \\
&\quad - \frac{1}{(t+1)\eta_t}\left(r - \frac{2L_{\mathrm{Str}}}{t+1}\eta_t - (1+\beta)L_{\mathrm{Bil}}^2\eta_t^2\right)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad + \frac{\eta_t}{t+1}\left((1+\tfrac{1}{\beta})\mathbb{E}\|\boldsymbol{\Delta}^{1,t-\frac{1}{2}}_{\mathrm{Bil}}\|^2 + \mathbb{E}\|\boldsymbol{\Delta}^{1,t}_{\mathrm{Bil}}\|^2\right) + \frac{\eta_t}{t+1}\left((1+\tfrac{1}{\beta})\mathbb{E}\|\boldsymbol{\Delta}^{2,t-\frac{1}{2}}_{\mathrm{Bil}}\|^2 + \mathbb{E}\|\boldsymbol{\Delta}^{2,t}_{\mathrm{Bil}}\|^2\right) \\
&\quad - \frac{1-r}{(t+1)\eta_t}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] - \frac{2}{t+1}\mathbb{E}\langle \boldsymbol{\Delta}^{1,t-\frac{1}{2}}_{\mathrm{Str}}, \mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\rangle - \frac{2}{t+1}\mathbb{E}\langle \boldsymbol{\Delta}^{2,t-\frac{1}{2}}_{\mathrm{Str}}, \mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\rangle \\
&\quad - \frac{2}{t+1}\mathbb{E}\langle \boldsymbol{\Delta}^{1,t-\frac{1}{2}}_{\mathrm{Str}}, \mathbf{x}_{t-1} - \tilde{\mathbf{x}}\rangle - \frac{2}{t+1}\mathbb{E}\langle \boldsymbol{\Delta}^{2,t-\frac{1}{2}}_{\mathrm{Str}}, \mathbf{y}_{t-1} - \tilde{\mathbf{y}}\rangle \\
&\quad - \frac{2}{t+1}\mathbb{E}\langle \boldsymbol{\Delta}^{1,t}_{\mathrm{Bil}}, \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle - \frac{2}{t+1}\mathbb{E}\langle \boldsymbol{\Delta}^{2,t}_{\mathrm{Bil}}, \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle \\
&\equiv \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{II}_1 + \mathrm{II}_2 + \mathrm{III}_1 + \mathrm{III}_2
\end{aligned} \tag{34}
$$

where for each line

$$
\mathrm{I}_1 + \mathrm{I}_2 \le \frac{1}{(t+1)\eta_t}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right)
$$

due to the stepsize condition (26) which in turn gives the factor in bracket $r - \frac{2L_{\rm Str}}{t+1}\eta_t - (1+\beta)L_{\rm Bil}^2\eta_t^2$ is nonnegative, and

$$\mathrm{II}_1 + \mathrm{II}_2 \le \frac{\eta_t}{t+1}\left((1+\tfrac{1}{\beta})\mathbb{E}\|\mathbf{\Delta}_{\rm Bil}^{t-\frac{1}{2}}\|^2 + \mathbb{E}\|\mathbf{\Delta}_{\rm Bil}^{t}\|^2\right) + \frac{\eta_t}{(1-r)(t+1)}\mathbb{E}\|\mathbf{\Delta}_{\rm Str}^{t-\frac{1}{2}}\|^2$$

due to the basic quadratic inequalities that $-\frac{1-r}{\eta_t}\|\mathbf{x}_{t-1} - \mathbf{x}_{t-\frac{1}{2}}\|^2 - 2\langle \mathbf{\Delta}_{\rm Str}^{1,t-\frac{1}{2}}, \mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\rangle \le \frac{\eta_t}{1-r}\|\mathbf{\Delta}_{\rm Str}^{1,t-\frac{1}{2}}\|^2$ and $-\frac{1-r}{\eta_t}\|\mathbf{y}_{t-1} - \mathbf{y}_{t-\frac{1}{2}}\|^2 - 2\langle \mathbf{\Delta}_{\rm Str}^{2,t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\rangle \le \frac{\eta_t}{1-r}\|\mathbf{\Delta}_{\rm Str}^{2,t-\frac{1}{2}}\|^2$, and finally

$$\mathrm{III}_1 = -\frac{2}{t+1}\mathbb{E}\langle \mathbf{\Delta}_{\rm Str}^{1,t-\frac{1}{2}}, \mathbf{x}_{t-1} - \tilde{\mathbf{x}}\rangle - \frac{2}{t+1}\mathbb{E}\langle \mathbf{\Delta}_{\rm Bil}^{1,t}, \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle = 0$$

and analogously $\mathrm{III}_2 = 0$, since each term in above is zero due to the law of iterated expectation applied to martingale difference conditions $\mathbb{E}[\mathbf{\Delta}_{\rm Str}^{i,t-\frac{1}{2}} \mid \mathcal{F}_{t-1}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{\Delta}_{\rm Bil}^{i,t} \mid \mathcal{F}_{t-\frac{1}{2}}] = \mathbf{0}$, $i = 1, 2$.

Multiplying both sides of (34) by $t(t+1)$ combined with the last three estimation bounds, and observing (5) and (6), we obtain for all $t = 1, \ldots, \mathscr{T}$

$$\begin{aligned}
&t(t+1)\mathbb{E}[V(\mathbf{x}_{t-\frac{1}{2}}^{\rm ag}, \mathbf{y}_{t-\frac{1}{2}}^{\rm ag} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - (t-1)t\mathbb{E}[V(\mathbf{x}_{t-\frac{3}{2}}^{\rm ag}, \mathbf{y}_{t-\frac{3}{2}}^{\rm ag} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le t(t+1)\left(\mathrm{I}_1 + \mathrm{I}_2 + \mathrm{II}_1 + \mathrm{II}_1 + \mathrm{III}_1 + \mathrm{III}_2\right) \\
&\le \frac{t}{\eta_t}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}}))]\right. \\
&\quad + t\eta_t\left(\tfrac{1}{1-r}\mathbb{E}\|\mathbf{\Delta}_{\rm Str}^{t-\frac{1}{2}}\|^2 + (1+\tfrac{1}{\beta})\mathbb{E}\|\mathbf{\Delta}_{\rm Bil}^{t-\frac{1}{2}}\|^2 + \mathbb{E}\|\mathbf{\Delta}_{\rm Bil}^{t}\|^2\right) \\
&\le \frac{t}{\eta_t}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}}))]\right. + \left(\tfrac{1}{1-r}\sigma_{\rm Str}^2 + (2+\tfrac{1}{\beta})\sigma_{\rm Bil}^2\right)t\eta_t
\end{aligned}$$

where in the last line above we applied (5) and (6) in Assumption 3, so by law of iterated expectations

$$\begin{aligned}
\mathbb{E}\|\mathbf{\Delta}_{\rm Str}^{t-\frac{1}{2}}\|^2 &= \mathbb{E}\left[\|\nabla f(\mathbf{x}_{t-1}^{\rm md}; \xi_{t-\frac{1}{2}}) - \nabla F(\mathbf{x}_{t-1}^{\rm md})\|^2 + \|\nabla g(\mathbf{y}_{t-1}^{\rm md}; \xi_{t-\frac{1}{2}}) - \nabla G(\mathbf{y}_{t-1}^{\rm md})\|^2\right] \le \sigma_{\rm Str}^2 \\
\mathbb{E}\|\mathbf{\Delta}_{\rm Bil}^{t-\frac{1}{2}}\|^2 &= \mathbb{E}\left[\|\nabla_{\mathbf{x}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) - \nabla_{\mathbf{x}} H(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2\right. \\
&\qquad \left. + \|-\nabla_{\mathbf{y}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) + \nabla_{\mathbf{y}} H(\mathbf{x}_{t-1}, \mathbf{y}_{t-1})\|^2\right] \le \sigma_{\rm Bil}^2 \\
\mathbb{E}\|\mathbf{\Delta}_{\rm Bil}^{t}\|^2 &= \mathbb{E}\left[\|\nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) - \nabla_{\mathbf{x}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}})\|^2\right. \\
&\qquad \left. + \|-\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla_{\mathbf{y}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}})\|^2\right] \le \sigma_{\rm Bil}^2
\end{aligned} \tag{35}$$

Now for a given $\mathcal{T} \in [1, \mathscr{T}]$, we finish the proof by telescope the above recursion for $t = 1, \ldots, \mathcal{T}$. We conclude from our choice of stepsize as in (9) that satisfies (26) so by denoting $\sigma \equiv \frac{1}{\sqrt{3}}\sqrt{\frac{1}{1-r}\sigma_{\rm Str}^2 + (2+\frac{1}{\beta})\sigma_{\rm Bil}^2}$, we have by Lemma 13(i)

$$\begin{aligned}
&\left(\tfrac{1}{1-r}\sigma_{\rm Str}^2 + (2+\tfrac{1}{\beta})\sigma_{\rm Bil}^2\right)\sum_{t=1}^{\mathcal{T}} t\eta_t = 3\sigma^2\sum_{t=1}^{\mathcal{T}} t\eta_t \le 3\sigma^2 \cdot \frac{1}{\square}\sum_{t=1}^{\mathcal{T}} t^2 \\
&= 3\sigma^2 \cdot \frac{\mathscr{C}\mathbb{E}[\mathcal{S}^{\frac{1}{2}}(\mathbf{x}_0, \mathbf{y}_0; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]}{\sigma[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{3} = \frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \sigma\mathscr{C}\mathbb{E}[\mathcal{S}^{\frac{1}{2}}(\mathbf{x}_0, \mathbf{y}_0; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]
\end{aligned}$$

where we recall in Lemma 13 that $\square \equiv \frac{\tilde{\sigma}[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}{\mathscr{C}\sqrt{\mathbb{E}[\|\mathbf{x}_0-\boldsymbol{\omega}_{\mathbf{x}}^\star\|^2+\|\mathbf{y}_0-\boldsymbol{\omega}_{\mathbf{y}}^\star\|^2]}}$. Finally

$$
\begin{aligned}
&\mathcal{T}(\mathcal{T}+1)\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}},\mathbf{y}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}}\mid\tilde{\mathbf{x}},\tilde{\mathbf{y}})]\\
&\le\sum_{t=1}^{\mathcal{T}}\frac{t}{\eta_t}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1},\mathbf{y}_{t-1};\tilde{\mathbf{x}},\tilde{\mathbf{y}})]-\mathbb{E}[\mathcal{S}(\mathbf{x}_t,\mathbf{y}_t;\tilde{\mathbf{x}},\tilde{\mathbf{y}})]\right)+\left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2+(2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right)\sum_{t=1}^{\mathcal{T}}t\eta_t\\
&=\frac{1}{\eta_1}\mathbb{E}[\mathcal{S}(\mathbf{x}_0,\mathbf{y}_0;\tilde{\mathbf{x}},\tilde{\mathbf{y}})]+\sum_{t=2}^{\mathcal{T}}\underbrace{\left(\frac{t}{\eta_t}-\frac{t-1}{\eta_{t-1}}\right)}_{=\sqrt{\frac{1+\beta}{r}}L_{\mathrm{Bil}}}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1},\mathbf{y}_{t-1};\tilde{\mathbf{x}},\tilde{\mathbf{y}})]-\frac{\mathcal{T}}{\eta_{\mathcal{T}}}\mathbb{E}[\mathcal{S}(\mathbf{x}_{\mathcal{T}},\mathbf{y}_{\mathcal{T}};\tilde{\mathbf{x}},\tilde{\mathbf{y}})]\\
&\quad+\frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}\cdot\mathscr{C}\sigma\mathbb{E}[\mathcal{S}^{\frac{1}{2}}(\mathbf{x}_0,\mathbf{y}_0;\tilde{\mathbf{x}},\tilde{\mathbf{y}})]
\end{aligned}
$$

Note in above derivations we applied Lemma 13(ii). Rearranging the terms along with Jensen's inequality proves (32).

Step 3. We conduct the following "bootstrapping" argument to arrive at our final theorem. Starting from the recursion (32) we have by setting $\tilde{\mathbf{x}}=\boldsymbol{\omega}_{\mathbf{x}}^\star$, $\tilde{\mathbf{y}}=\boldsymbol{\omega}_{\mathbf{y}}^\star$, Lemma 12 implies that its first summand on the left hand $\mathcal{T}(\mathcal{T}+1)\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}},\mathbf{y}^{\mathrm{ag}}_{\mathcal{T}-\frac{1}{2}}\mid\boldsymbol{\omega}_{\mathbf{x}}^\star,\boldsymbol{\omega}_{\mathbf{y}}^\star)]$ is nonnegative, and hence we can drop it and have for any $\mathcal{T}=1,\ldots,\mathscr{T}$

$$
\begin{aligned}
&\frac{\mathcal{T}}{\eta_{\mathcal{T}}}\mathbb{E}[\mathcal{S}(\mathbf{x}_{\mathcal{T}},\mathbf{y}_{\mathcal{T}};\boldsymbol{\omega}_{\mathbf{x}}^\star,\boldsymbol{\omega}_{\mathbf{y}}^\star)]\le\frac{1}{\eta_1}\mathbb{E}[\mathcal{S}(\mathbf{x}_0,\mathbf{y}_0;\boldsymbol{\omega}_{\mathbf{x}}^\star,\boldsymbol{\omega}_{\mathbf{y}}^\star)]\\
&\quad+\sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}}\sum_{t=2}^{\mathcal{T}}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1},\mathbf{y}_{t-1};\boldsymbol{\omega}_{\mathbf{x}}^\star,\boldsymbol{\omega}_{\mathbf{y}}^\star)]+\frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}\cdot\mathscr{C}\sigma\sqrt{\mathbb{E}[\mathcal{S}(\mathbf{x}_0,\mathbf{y}_0;\boldsymbol{\omega}_{\mathbf{x}}^\star,\boldsymbol{\omega}_{\mathbf{y}}^\star)]}\\
&=(\tfrac{2}{r}L_{\mathrm{Str}}+\square)\mathbb{E}[\mathcal{S}(\mathbf{x}_0,\mathbf{y}_0;\boldsymbol{\omega}_{\mathbf{x}}^\star,\boldsymbol{\omega}_{\mathbf{y}}^\star)]\\
&\quad+\sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}}\underbrace{\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1},\mathbf{y}_{t-1};\boldsymbol{\omega}_{\mathbf{x}}^\star,\boldsymbol{\omega}_{\mathbf{y}}^\star)]}_{\equiv\mathcal{Q}_{\mathcal{T}-1}}+\frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}\cdot\mathscr{C}\sigma\sqrt{\mathbb{E}[\mathcal{S}(\mathbf{x}_0,\mathbf{y}_0;\boldsymbol{\omega}_{\mathbf{x}}^\star,\boldsymbol{\omega}_{\mathbf{y}}^\star)]}
\end{aligned}
\tag{36}
$$

Converting (36) to a version of partial sum $\mathcal{Q}_{\mathcal{T}-1}\equiv\sum_{t=1}^{\mathcal{T}}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1},\mathbf{y}_{t-1};\boldsymbol{\omega}_{\mathbf{x}}^\star,\boldsymbol{\omega}_{\mathbf{y}}^\star)]$ that for all $\mathcal{T}=1,\ldots,\mathscr{T}$

$$
\begin{aligned}
&\frac{\mathcal{T}}{\eta_{\mathcal{T}}}\mathbb{E}[\mathcal{S}(\mathbf{x}_{\mathcal{T}},\mathbf{y}_{\mathcal{T}};\boldsymbol{\omega}_{\mathbf{x}}^\star,\boldsymbol{\omega}_{\mathbf{y}}^\star)]=\frac{\mathcal{T}}{\eta_{\mathcal{T}}}(\mathcal{Q}_{\mathcal{T}}-\mathcal{Q}_{\mathcal{T}-1})\\
&\le\sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}}\mathcal{Q}_{\mathcal{T}-1}+\frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}\cdot\mathscr{C}\sigma\sqrt{\mathcal{Q}_0}+(\tfrac{2}{r}L_{\mathrm{Str}}+\square)\mathcal{Q}_0
\end{aligned}
\tag{37}
$$

(37) is equivalently written as

$$
\frac{\mathcal{T}}{\eta_{\mathcal{T}}}\mathcal{Q}_{\mathcal{T}}\le\frac{\mathcal{T}+1}{\eta_{\mathcal{T}+1}}\mathcal{Q}_{\mathcal{T}-1}+\frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}\cdot\mathscr{C}\sigma\sqrt{\mathcal{Q}_0}+(\tfrac{2}{r}L_{\mathrm{Str}}+\square)\mathcal{Q}_0
$$

From here and onwards, we denote $\kappa_t \equiv \frac{t}{\eta_t} = \frac{2}{r}L_{\mathrm{Str}} \vee \square + \sqrt{\frac{1+\beta}{r}}L_{\mathrm{Bil}}t$ for each $t = 1, \ldots, \mathscr{T}$. 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

$$\frac{\mathcal{Q}_{\mathcal{T}}}{\kappa_{\mathcal{T}+1}} \le \frac{\mathcal{Q}_{\mathcal{T}-1}}{\kappa_{\mathcal{T}}} + \frac{\frac{\mathcal{T}(\mathcal{T}+\frac{1}{2})(\mathcal{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0} + (\frac{2}{r}L_{\mathrm{Str}} + \square)\mathcal{Q}_0}{\kappa_{\mathcal{T}} \cdot \kappa_{\mathcal{T}+1}}$$

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

$$\begin{aligned}
\frac{\mathcal{Q}_{\mathcal{T}-1}}{\kappa_{\mathcal{T}}} &\le \frac{\mathcal{Q}_0}{\kappa_1} + \sum_{T=1}^{\mathcal{T}-1} \frac{\frac{T(T+\frac{1}{2})(T+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0} + (\frac{2}{r}L_{\mathrm{Str}} + \square)\mathcal{Q}_0}{\kappa_T \cdot \kappa_{T+1}} \\
&\le \frac{\mathcal{Q}_0}{\kappa_1} + \left[\frac{\mathscr{T}(\mathscr{T}+\frac{1}{2})(\mathscr{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0} + (\tfrac{2}{r}L_{\mathrm{Str}} + \square)\mathcal{Q}_0\right] \sum_{T=1}^{\mathcal{T}-1} \frac{1}{\kappa_T \cdot \kappa_{T+1}}
\end{aligned}$$

where we applied Lemma 13(ii) that for all $T = 1, \ldots, \mathcal{T} - 1$ we have $\kappa_{T+1} - \kappa_T = \sqrt{\frac{1+\beta}{r}}L_{\mathrm{Bil}}$. This yields

$$\sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}} \sum_{T=1}^{\mathcal{T}-1} \frac{1}{\kappa_T \cdot \kappa_{T+1}} = \sum_{T=1}^{\mathcal{T}-1} \left[\frac{1}{\kappa_T} - \frac{1}{\kappa_{T+1}}\right] = \frac{1}{\kappa_1} - \frac{1}{\kappa_{\mathcal{T}}}$$

and hence

$$\begin{aligned}
&\sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}} \frac{\mathcal{Q}_{\mathcal{T}-1}}{\kappa_{\mathcal{T}}} \\
&\le \sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}} \frac{\mathcal{Q}_0}{\kappa_1} + \left[\frac{\mathscr{T}(\mathscr{T}+\frac{1}{2})(\mathscr{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0} + (\tfrac{2}{r}L_{\mathrm{Str}} + \square)\mathcal{Q}_0\right] \sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}} \sum_{T=1}^{\mathcal{T}-1} \frac{1}{\kappa_T \cdot \kappa_{T+1}} \\
&= \sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}} \frac{\mathcal{Q}_0}{\kappa_1} + \left[\frac{\mathscr{T}(\mathscr{T}+\frac{1}{2})(\mathscr{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0} + (\tfrac{2}{r}L_{\mathrm{Str}} + \square)\mathcal{Q}_0\right] \left(\frac{1}{\kappa_1} - \frac{1}{\kappa_{\mathcal{T}}}\right) \\
&\le \frac{\frac{\mathscr{T}(\mathscr{T}+\frac{1}{2})(\mathscr{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0} + (\frac{2}{r}L_{\mathrm{Str}} + \square + \sqrt{\frac{1+\beta}{r}}L_{\mathrm{Bil}})\mathcal{Q}_0}{\frac{2}{r}L_{\mathrm{Str}} + \square + \sqrt{\frac{1+\beta}{r}}L_{\mathrm{Bil}}} - \frac{\frac{\mathscr{T}(\mathscr{T}+\frac{1}{2})(\mathscr{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0} + (\frac{2}{r}L_{\mathrm{Str}} + \square)\mathcal{Q}_0}{\kappa_{\mathcal{T}}} \\
&= \mathcal{Q}_0 + \frac{\frac{\mathscr{T}(\mathscr{T}+\frac{1}{2})(\mathscr{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0}}{\kappa_1} - \frac{\frac{\mathscr{T}(\mathscr{T}+\frac{1}{2})(\mathscr{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0} + (\frac{2}{r}L_{\mathrm{Str}} + \square)\mathcal{Q}_0}{\kappa_{\mathcal{T}}}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[\mathcal{S}(\mathbf{x}_{\mathcal{T}}, \mathbf{y}_{\mathcal{T}}; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] &\le \sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}} \frac{\mathcal{Q}_{\mathcal{T}-1}}{\kappa_{\mathcal{T}}} + \frac{\frac{\mathscr{T}(\mathscr{T}+\frac{1}{2})(\mathscr{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0} + (\frac{2}{r}L_{\mathrm{Str}} + \square)\mathcal{Q}_0}{\kappa_{\mathcal{T}}} \\
&\le \mathcal{Q}_0 + \frac{\frac{\mathscr{T}(\mathscr{T}+\frac{1}{2})(\mathscr{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma\sqrt{\mathcal{Q}_0}}{\kappa_1} \\
&\le \left(1 + \frac{\mathscr{C}\sigma[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}{\kappa_1\sqrt{\mathcal{Q}_0}}\right)\mathcal{Q}_0 = \mathcal{A}(\sigma; \mathscr{T}, \mathscr{C}, r, \beta)\mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)]
\end{aligned} \tag{38}$$

where the prefactor $\mathcal{A}(\sigma; \mathscr{T}, \mathscr{C}, r, \beta)$ defined as in (11) lies in $[1, 1+\mathscr{C}^2]$ and reduces to 1 when the argument is set as 0.[7]

Now we drop the second summand on the left hand of (32) with $\tilde{\mathbf{x}} = \boldsymbol{\omega}_{\mathbf{x}}^\star$, $\tilde{\mathbf{y}} = \boldsymbol{\omega}_{\mathbf{y}}^\star$, $\mathcal{T} = \mathscr{T}$. Combining with (38) ($\mathcal{T} = \mathscr{T}$) gives

$$\begin{aligned}
&\mathscr{T}(\mathscr{T}+1)\mathbb{E}[V(\mathbf{x}^{\text{ag}}_{\mathscr{T}-\frac12}, \mathbf{y}^{\text{ag}}_{\mathscr{T}-\frac12} \mid \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \\
&\le \kappa_1 \mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] + \sqrt{\tfrac{1+\beta}{r}} L_{\text{Bil}} \sum_{t=2}^{\mathscr{T}} \mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \\
&\quad + \frac{\mathscr{T}(\mathscr{T}+\frac12)(\mathscr{T}+1)}{[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}} \cdot \mathscr{C}\sigma \sqrt{\mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)]} \\
&\le \left(\tfrac{2}{r} L_{\text{Str}} + \frac{\sigma[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}{\mathscr{C}\sqrt{\mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)]}} + \sqrt{\tfrac{1+\beta}{r}} L_{\text{Bil}} \right) \mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \\
&\quad + \sqrt{\tfrac{1+\beta}{r}} L_{\text{Bil}} (\mathscr{T}-1) \cdot \mathcal{A}(\sigma; \mathscr{T}, \mathscr{C}, r, \beta) \mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \\
&\quad + \mathscr{C}\sigma[\mathscr{T}(\mathscr{T}+1)^2]^{1/2} \sqrt{\mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)]} \\
&\le \left(\tfrac{2}{r} L_{\text{Str}} + \mathcal{A}(\sigma; \mathscr{T}, \mathscr{C}, r, \beta) \sqrt{\tfrac{1+\beta}{r}} L_{\text{Bil}} \mathscr{T} \right) \mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \\
&\quad + (\tfrac{1}{\mathscr{C}} + \mathscr{C})\sigma[\mathscr{T}(\mathscr{T}+1)^2]^{1/2} \sqrt{\mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)]}
\end{aligned}$$

Using (24) and (25) in Lemma 12 again lower bounds the left hand in the last display as

$$\mathscr{T}(\mathscr{T}+1)\mathbb{E}[V(\mathbf{x}^{\text{ag}}_{\mathscr{T}-\frac12}, \mathbf{y}^{\text{ag}}_{\mathscr{T}-\frac12} \mid \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \ge \frac{\mu_{\text{Str}}}{2} \mathscr{T}(\mathscr{T}+1)\mathbb{E}[\mathcal{S}(\mathbf{x}^{\text{ag}}_{\mathscr{T}-\frac12}, \mathbf{y}^{\text{ag}}_{\mathscr{T}-\frac12}; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \ge 0$$

Dividing both sides by $\frac{\mu_{\text{Str}}}{2}\mathscr{T}(\mathscr{T}+1)$ concludes

$$\begin{aligned}
\mathbb{E}[\mathcal{S}(\mathbf{x}^{\text{ag}}_{\mathscr{T}-\frac12}, \mathbf{y}^{\text{ag}}_{\mathscr{T}-\frac12}; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \le{} & \frac{2\left(\tfrac{2}{r} L_{\text{Str}} + \mathcal{A}(\sigma; \mathscr{T}, \mathscr{C}, r, \beta) \sqrt{\tfrac{1+\beta}{r}} L_{\text{Bil}} \mathscr{T} \right)}{\mu_{\text{Str}} \mathscr{T}(\mathscr{T}+1)} \mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \\
& + \frac{2(\frac{1}{\mathscr{C}} + \mathscr{C})\sigma}{\mu_{\text{Str}} \mathscr{T}^{1/2}} \sqrt{\mathbb{E}[\mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)]}
\end{aligned}$$

and hence concludes (10) and the whole proof of Theorem 6.

B.4. Proof of Theorem 9

Using a scaling reduction argument analogous to the one in §B.1 we only need to prove the case of $\mathscr{R} = 1$. We overload function notations F, H, G to the new group accordingly where $F \leftarrow F(\mathbf{x}) - \frac{\mu_\star}{2}\|\mathbf{x} - \mathbf{x}_0\|^2$ and $G \leftarrow G(\mathbf{y}) - \frac{\mu_\star}{2}\|\mathbf{y} - \mathbf{y}_0\|^2$ are nonstrongly convex and $H \leftarrow \frac{\mu_\star}{2}\|\mathbf{x} -$

[7] Indeed, we have from the definition (11) of the prefactor $\mathcal{A}(\tilde{\sigma}; \mathscr{T}, \mathscr{C}, r, \beta) = 1 + \frac{\mathscr{C}\tilde{\sigma}[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}{\kappa_1\sqrt{\mathbb{E}[\|\mathbf{x}_0 - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \|\mathbf{y}_0 - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2]}} \ge 1$ and also by Lemma 13(i) we have $\kappa_1 \ge \frac{\tilde{\sigma}[\mathscr{T}(\mathscr{T}+1)^2]^{1/2}}{\mathscr{C}\sqrt{\mathbb{E}[\|\mathbf{x}_0 - \boldsymbol{\omega}_{\mathbf{x}}^\star\|^2 + \|\mathbf{y}_0 - \boldsymbol{\omega}_{\mathbf{y}}^\star\|^2]}}$ and hence it satisfies $\mathcal{A}(\tilde{\sigma}; \mathscr{T}, \mathscr{C}, r, \beta) \le 1 + \mathscr{C}^2$.

$\mathbf{x}_0\|^2 + \mathbf{x}^\top \mathbf{B}\mathbf{y} - \mathbf{x}^\top \mathbf{u}_\mathbf{x} + \mathbf{u}_\mathbf{y}^\top \mathbf{y} - \frac{\mu_\star}{2}\|\mathbf{y}-\mathbf{y}_0\|^2$ is an isotropic quadratic. For convenience we repeat the iterates of Algorithm 2 with $\mathscr{R} = 1$ as

$$
\begin{aligned}
\mathbf{x}_{t-\frac{1}{2}} &= \mathbf{x}_{t-1} - \eta_t \left(\nabla f(\mathbf{x}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_\mathbf{x} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) \right), \\
\mathbf{y}_{t-\frac{1}{2}} &= \mathbf{y}_{t-1} - \eta_t \left(-\nabla_\mathbf{y} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) + \nabla g(\mathbf{y}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) \right), \\
\mathbf{x}^{\text{ag}}_{t-\frac{1}{2}} &= (1-\alpha_t)\mathbf{x}^{\text{ag}}_{t-\frac{3}{2}} + \alpha_t \mathbf{x}_{t-\frac{1}{2}}, \\
\mathbf{y}^{\text{ag}}_{t-\frac{1}{2}} &= (1-\alpha_t)\mathbf{y}^{\text{ag}}_{t-\frac{3}{2}} + \alpha_t \mathbf{y}_{t-\frac{1}{2}}, \\
\mathbf{x}_t &= \mathbf{x}_{t-1} - \eta_t \left(\nabla f(\mathbf{x}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) + \nabla_\mathbf{x} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) \right), \\
\mathbf{y}_t &= \mathbf{y}_{t-1} - \eta_t \left(-\nabla_\mathbf{y} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\mathbf{y}^{\text{md}}_{t-1}; \xi_{t-\frac{1}{2}}) \right), \\
\mathbf{x}^{\text{md}}_t &= (1-\alpha_{t+1})\mathbf{x}^{\text{ag}}_{t-\frac{1}{2}} + \alpha_{t+1}\mathbf{x}_t, \\
\mathbf{y}^{\text{md}}_t &= (1-\alpha_{t+1})\mathbf{y}^{\text{ag}}_{t-\frac{1}{2}} + \alpha_{t+1}\mathbf{y}_t
\end{aligned}
$$

with the initialization $\mathbf{x}_0 = \mathbf{x}^{\text{md}}_0 = \mathbf{x}^{\text{ag}}_{-\frac{1}{2}} \in \mathbb{R}^n$, $\mathbf{y}_0 = \mathbf{y}^{\text{md}}_0 = \mathbf{y}^{\text{ag}}_{-\frac{1}{2}} \in \mathbb{R}^m$. We continue to assume the noise-related setting as in (27), and continue to denote $\mathcal{S}(\mathbf{x}, \mathbf{y}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \equiv \|\mathbf{x}-\tilde{\mathbf{x}}\|^2 + \|\mathbf{y}-\tilde{\mathbf{y}}\|^2$. Our proof proceeds in the following steps:

Step 1. We prove the following generalization of Lemma 14:

Lemma 15 *For arbitrary $\tilde{\mathbf{x}} \in \mathbb{R}^n, \tilde{\mathbf{y}} \in \mathbb{R}^m$ and $\alpha_t \in (0,1]$ the iterates of Algorithm 2 satisfy almost surely*

$$
\begin{aligned}
&V(\mathbf{x}^{\text{ag}}_{t-\frac{1}{2}}, \mathbf{y}^{\text{ag}}_{t-\frac{1}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) - (1-\alpha_t)V(\mathbf{x}^{\text{ag}}_{t-\frac{3}{2}}, \mathbf{y}^{\text{ag}}_{t-\frac{3}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \\
&\le \alpha_t \langle \nabla F(\mathbf{x}^{\text{md}}_{t-1}) + \nabla_\mathbf{x} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle + \alpha_t \langle -\nabla_\mathbf{y} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}) + \nabla G(\mathbf{y}^{\text{md}}_{t-1}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} \rangle \\
&\quad + \frac{\alpha_t^2 L_{\text{Str}}}{2} \mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1}) - \alpha_t \mu_\star \mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})
\end{aligned}
\tag{39}
$$

The proof goes in an analogous fashion as the proof of Lemma 14, except that the display above (31) is replaced by

$$
\begin{aligned}
&\langle \nabla_\mathbf{x} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle + \langle -\nabla_\mathbf{y} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} \rangle \\
&\le \langle \nabla_\mathbf{x} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle + \langle -\nabla_\mathbf{y} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} \rangle - \mu_\star \mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})
\end{aligned}
$$

due to our H being a $\mu_\star$-strongly-convex-$\mu_\star$-strongly-concave isotropic quadratic function after scaling reduction. Hence (31) becomes

$$
\begin{aligned}
&\langle \nabla_\mathbf{x} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}^{\text{ag}}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle - (1-\alpha_t) \langle \nabla_\mathbf{x} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x}^{\text{ag}}_{t-\frac{3}{2}} - \tilde{\mathbf{x}} \rangle \\
&\qquad + \langle -\nabla_\mathbf{y} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}^{\text{ag}}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} \rangle - (1-\alpha_t) \langle -\nabla_\mathbf{y} H(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y}^{\text{ag}}_{t-\frac{3}{2}} - \tilde{\mathbf{y}} \rangle \\
&\le \alpha_t \left[\langle \nabla_\mathbf{x} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}} \rangle + \langle -\nabla_\mathbf{y} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}} \rangle - \mu_\star \mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \right]
\end{aligned}
\tag{40}
$$

This concludes (39) and the whole lemma.

Step 2. Analogous to (33) in Step 2 in the proof of Theorem 6 in §B.3 we conclude for all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$,

$$\begin{aligned}
&\eta_t \mathbb{E}\langle \nabla f(\mathbf{x}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle \\
&\qquad + \eta_t \mathbb{E}\langle -\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\mathbf{y}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle \\
&\le \frac{1}{2}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) \\
&\quad - \frac{1-(1+\beta)L_{\mathrm{Bil}}^2\eta_t^2}{2}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] + \frac{\eta_t^2}{2}(2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2
\end{aligned}$$

To show this, note that

$$\begin{aligned}
&\eta_t \langle \nabla f(\mathbf{x}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle \\
&\le \frac{1}{2}\left(\|\mathbf{x}_{t-1} - \tilde{\mathbf{x}}\|^2 - \|\mathbf{x}_t - \tilde{\mathbf{x}}\|^2 - \|\mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\|^2\right) \\
&\quad + \frac{\eta_t^2}{2}\|\nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) - \nabla_{\mathbf{x}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}})\|^2
\end{aligned}$$

and analogously

$$\begin{aligned}
&\eta_t \langle -\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\mathbf{y}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle \\
&\le \frac{1}{2}\left(\|\mathbf{y}_{t-1} - \tilde{\mathbf{y}}\|^2 - \|\mathbf{y}_t - \tilde{\mathbf{y}}\|^2 - \|\mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\|^2\right) \\
&\quad + \frac{\eta_t^2}{2}\|\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) - \nabla_{\mathbf{y}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}})\|^2
\end{aligned}$$

To handle the stochastic terms, Young's inequality combined with the martingale structure, along with the definition of L_{Bil}, indicates

$$\begin{aligned}
&\mathbb{E}\left\|\begin{bmatrix} \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) - \nabla_{\mathbf{x}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) \\ \nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) - \nabla_{\mathbf{y}} h(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \zeta_{t-\frac{1}{2}}) \end{bmatrix}\right\|^2 \\
&= \mathbb{E}\left\|\begin{bmatrix} \nabla_{\mathbf{x}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}) - \nabla_{\mathbf{x}} H(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) - \mathbf{\Delta}_{\mathrm{Bil}}^{1,t-\frac{1}{2}} \\ \nabla_{\mathbf{y}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}) - \nabla_{\mathbf{y}} H(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}) - \mathbf{\Delta}_{\mathrm{Bil}}^{2,t-\frac{1}{2}} \end{bmatrix}\right\|^2 + \mathbb{E}\left\|\begin{bmatrix} \mathbf{\Delta}_{\mathrm{Bil}}^{1,t} \\ \mathbf{\Delta}_{\mathrm{Bil}}^{2,t} \end{bmatrix}\right\|^2 \\
&\le (1+\beta)\mathbb{E}\left\|\begin{bmatrix} \mu_\star \mathbf{I} & \mathbf{B} \\ -\mathbf{B}^\top & \mu_\star \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1} \\ \mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1} \end{bmatrix}\right\|^2 + (1+\tfrac{1}{\beta})\mathbb{E}\left\|\begin{bmatrix} \mathbf{\Delta}_{\mathrm{Bil}}^{1,t-\frac{1}{2}} \\ \mathbf{\Delta}_{\mathrm{Bil}}^{2,t-\frac{1}{2}} \end{bmatrix}\right\|^2 + \mathbb{E}\left\|\begin{bmatrix} \mathbf{\Delta}_{\mathrm{Bil}}^{1,t} \\ \mathbf{\Delta}_{\mathrm{Bil}}^{2,t} \end{bmatrix}\right\|^2 \\
&\le (1+\beta)L_{\mathrm{Bil}}^2\left(\mathbb{E}\|\mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\|^2 + \mathbb{E}\|\mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\|^2\right) + (1+\tfrac{1}{\beta})\mathbb{E}\|\mathbf{\Delta}_{\mathrm{Bil}}^{t-\frac{1}{2}}\|^2 + \mathbb{E}\|\mathbf{\Delta}_{\mathrm{Bil}}^t\|^2
\end{aligned}$$

Combining the last three displays gives

$$\begin{aligned}
&\eta_t \mathbb{E}\langle \nabla f(\mathbf{x}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle \\
&\qquad + \eta_t \mathbb{E}\langle -\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\mathbf{y}_{t-1}^{\mathrm{md}}; \xi_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle \\
&\le \frac{1}{2}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) - \frac{1-(1+\beta)L_{\mathrm{Bil}}^2\eta_t^2}{2}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})]
\end{aligned}$$

$$+ \frac{\eta_t^2}{2}\left((1+\tfrac{1}{\beta})\mathbb{E}\|\boldsymbol{\Delta}_{\text{Bil}}^{t-\frac{1}{2}}\|^2 + \mathbb{E}\|\boldsymbol{\Delta}_{\text{Bil}}^{t}\|^2\right). \tag{33}$$

Combining this with Lemma 15, we have

$$\begin{aligned}
&\mathbb{E}[V(\mathbf{x}_{t-\frac{1}{2}}^{\text{ag}}, \mathbf{y}_{t-\frac{1}{2}}^{\text{ag}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - (1-\alpha_t)\mathbb{E}[V(\mathbf{x}_{t-\frac{3}{2}}^{\text{ag}}, \mathbf{y}_{t-\frac{3}{2}}^{\text{ag}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le \alpha_t \mathbb{E}\langle \nabla F(\mathbf{x}_{t-1}^{\text{md}}) + \nabla_{\mathbf{x}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle + \alpha_t \mathbb{E}\langle -\nabla_{\mathbf{y}} H(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}) + \nabla G(\mathbf{y}_{t-1}^{\text{md}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle \\
&\quad + \tfrac{\alpha_t^2 L_{\text{Str}}}{2}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] - \alpha_t\mu_\star \mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&= \alpha_t \mathbb{E}\langle \nabla f(\mathbf{x}_{t-1}^{\text{md}}; \xi_{t-\frac{1}{2}}) + \nabla_{\mathbf{x}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t), \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle \\
&\quad + \alpha_t \mathbb{E}\langle -\nabla_{\mathbf{y}} h(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \zeta_t) + \nabla g(\mathbf{y}_{t-1}^{\text{md}}; \xi_{t-\frac{1}{2}}), \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle \\
&\quad - \alpha_t \mathbb{E}\langle \boldsymbol{\Delta}_{\text{Str}}^{1,t-\frac{1}{2}} + \boldsymbol{\Delta}_{\text{Bil}}^{1,t}, \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle - \alpha_t \mathbb{E}\langle \boldsymbol{\Delta}_{\text{Str}}^{2,t-\frac{1}{2}} + \boldsymbol{\Delta}_{\text{Bil}}^{2,t}, \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle \\
&\quad + \tfrac{\alpha_t^2 L_{\text{Str}}}{2}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] - \alpha_t\mu_\star \mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le \frac{\alpha_t}{\eta_t}\Bigg(\frac{1}{2}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) \\
&\qquad\qquad - \frac{1-(1+\beta)L_{\text{Bil}}^2\eta_t^2}{2}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] + \frac{\eta_t^2}{2}(2+\tfrac{1}{\beta})\sigma_{\text{Bil}}^2\Bigg) \\
&\quad - \alpha_t \mathbb{E}\langle \boldsymbol{\Delta}_{\text{Str}}^{1,t-\frac{1}{2}} + \boldsymbol{\Delta}_{\text{Bil}}^{1,t}, \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle - \alpha_t \mathbb{E}\langle \boldsymbol{\Delta}_{\text{Str}}^{2,t-\frac{1}{2}} + \boldsymbol{\Delta}_{\text{Bil}}^{2,t}, \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle \\
&\quad + \tfrac{\alpha_t^2 L_{\text{Str}}}{2}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] - \alpha_t\mu_\star \mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]
\end{aligned}$$

Continuing this estimation gives (note Young's inequality applies, and $\mathbb{E}\langle \boldsymbol{\Delta}_{\text{Str}}^{1,t-\frac{1}{2}} + \boldsymbol{\Delta}_{\text{Bil}}^{1,t}, \mathbf{x}_{t-\frac{1}{2}} - \tilde{\mathbf{x}}\rangle = \mathbb{E}\langle \boldsymbol{\Delta}_{\text{Str}}^{1,t-\frac{1}{2}}, \mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\rangle$ and analogously $\mathbb{E}\langle \boldsymbol{\Delta}_{\text{Str}}^{2,t-\frac{1}{2}} + \boldsymbol{\Delta}_{\text{Bil}}^{2,t}, \mathbf{y}_{t-\frac{1}{2}} - \tilde{\mathbf{y}}\rangle = \mathbb{E}\langle \boldsymbol{\Delta}_{\text{Str}}^{2,t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\rangle$)

$$\begin{aligned}
&\mathbb{E}[V(\mathbf{x}_{t-\frac{1}{2}}^{\text{ag}}, \mathbf{y}_{t-\frac{1}{2}}^{\text{ag}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - (1-\alpha_t)\mathbb{E}[V(\mathbf{x}_{t-\frac{3}{2}}^{\text{ag}}, \mathbf{y}_{t-\frac{3}{2}}^{\text{ag}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le \frac{\alpha_t}{2\eta_t}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) \\
&\quad - \frac{\alpha_t}{2\eta_t}\left(r - \alpha_t L_{\text{Str}}\eta_t - (1+\beta)L_{\text{Bil}}^2\eta_t^2\right)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad + \frac{\alpha_t\eta_t}{2}(2+\tfrac{1}{\beta})\sigma_{\text{Bil}}^2 - \frac{\alpha_t(1-r)}{2\eta_t}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad - \alpha_t \mathbb{E}\langle \boldsymbol{\Delta}_{\text{Str}}^{1,t-\frac{1}{2}}, \mathbf{x}_{t-\frac{1}{2}} - \mathbf{x}_{t-1}\rangle - \alpha_t \mathbb{E}\langle \boldsymbol{\Delta}_{\text{Str}}^{2,t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}} - \mathbf{y}_{t-1}\rangle - \alpha_t\mu_\star \mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le \frac{\alpha_t}{2\eta_t}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) \\
&\quad - \frac{\alpha_t}{2\eta_t}\left(r - \alpha_t L_{\text{Str}}\eta_t - (1+\beta)L_{\text{Bil}}^2\eta_t^2\right)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad + \frac{\alpha_t\eta_t}{2}(2+\tfrac{1}{\beta})\sigma_{\text{Bil}}^2 + \frac{\alpha_t\eta_t}{2(1-r)}\mathbb{E}[\|\boldsymbol{\Delta}_{\text{Str}}^{1,t-\frac{1}{2}}\|^2 + \|\boldsymbol{\Delta}_{\text{Str}}^{2,t-\frac{1}{2}}\|^2] - \alpha_t\mu_\star \mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le \frac{\alpha_t}{2\eta_t}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right)
\end{aligned}$$

$$- \frac{\alpha_t}{2\eta_t}\left(r - \alpha_t L_{\mathrm{Str}}\eta_t - (1+\beta)L_{\mathrm{Bil}}^2\eta_t^2\right)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})]$$
$$- \alpha_t\mu_\star\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] + \frac{\alpha_t\eta_t}{2}\left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right)$$

This yields, applying Young's inequality,

$$\begin{aligned}
&\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}}, \mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - (1-\alpha_t)\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}}, \mathbf{y}^{\mathrm{ag}}_{t-\frac{3}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le \frac{\alpha_t}{2\eta_t}\left(\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) \\
&\quad - \frac{\alpha_t}{2\eta_t}\left(r - \alpha_t L_{\mathrm{Str}}\eta_t - (1+\beta)L_{\mathrm{Bil}}^2\eta_t^2\right)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad - \alpha_t\mu_\star\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] + \frac{\alpha_t\eta_t}{2}\left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right) \\
&\le \frac{\alpha_t}{2\eta_t}\left((1-\alpha_t)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) \\
&\quad - \frac{\alpha_t}{2\eta_t}\left(r - \alpha_t L_{\mathrm{Str}}\eta_t - (1+\beta)L_{\mathrm{Bil}}^2\eta_t^2\right)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad + \frac{\alpha_t^2}{2\eta_t}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \alpha_t\mu_\star\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] + \frac{\alpha_t\eta_t}{2}\left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right) \\
&\le \frac{\alpha_t}{2\eta_t}\left((1-\alpha_t)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] - \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) \\
&\quad - \frac{\alpha_t}{2\eta_t}\left(r - \alpha_t L_{\mathrm{Str}}\eta_t - (1+\beta)L_{\mathrm{Bil}}^2\eta_t^2\right)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad + \eta_t\mu_\star^2\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] + \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}$$

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

$$\begin{aligned}
&\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}}, \mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] + \frac{\mu_\star}{2}\mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\quad - (1-\alpha_t)\left(\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}}, \mathbf{y}^{\mathrm{ag}}_{t-\frac{3}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] + \frac{\mu_\star}{2}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) \\
&\le -\frac{\mu_\star}{2}\left(r - 2\alpha_t - \left(\tfrac{L_{\mathrm{Str}}}{\mu_\star} + \tfrac{(1+\beta)L_{\mathrm{Bil}}^2}{\mu_\star^2}\right)\alpha_t^2\right)\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-\frac{1}{2}}, \mathbf{y}_{t-\frac{1}{2}}; \mathbf{x}_{t-1}, \mathbf{y}_{t-1})] \\
&\quad + \frac{\alpha_t^2}{2\mu_\star}\left(\tfrac{1}{1-r}\sigma_{\mathrm{Str}}^2 + (2+\tfrac{1}{\beta})\sigma_{\mathrm{Bil}}^2\right)
\end{aligned}$$

Step 3. By the definition α_t we have $r - 2\alpha_t - \left(\frac{L_{\mathrm{Str}}}{\mu_\star} + \frac{(1+\beta)L_{\mathrm{Bil}}^2}{\mu_\star^2}\right)\alpha_t^2 \ge 0$, so we obtain regularity condition $\alpha_t \le \bar{\alpha} = \frac{r}{1+\sqrt{1+r\left(\frac{L_{\mathrm{Str}}}{\mu_\star} + \frac{(1+\beta)L_{\mathrm{Bil}}^2}{\mu_\star^2}\right)}}$ of Theorem 9. 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

$$\begin{aligned}
&\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{t-\frac{1}{2}}, \mathbf{y}^{\mathrm{ag}}_{t-\frac{1}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] + \frac{\mu_\star}{2}\mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] \\
&\le (1-\alpha_t)\left(\mathbb{E}[V(\mathbf{x}^{\mathrm{ag}}_{t-\frac{3}{2}}, \mathbf{y}^{\mathrm{ag}}_{t-\frac{3}{2}} \mid \tilde{\mathbf{x}}, \tilde{\mathbf{y}})] + \frac{\mu_\star}{2}\mathbb{E}[\mathcal{S}(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}})]\right) + \frac{3\alpha_t^2}{2\mu_\star}\sigma^2
\end{aligned}$$

Plugging in $\tilde{\mathbf{x}} = \boldsymbol{\omega}_{\mathbf{x}}^\star, \tilde{\mathbf{y}} = \boldsymbol{\omega}_{\mathbf{y}}^\star$ gives

$$\begin{aligned}
&\mathbb{E}[V(\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \mid \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \\
&= F(\tilde{\mathbf{x}}) + G(\tilde{\mathbf{y}}) - F(\boldsymbol{\omega}_{\mathbf{x}}^\star) - G(\boldsymbol{\omega}_{\mathbf{y}}^\star) + \langle \nabla_{\mathbf{x}} H(\boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star), \tilde{\mathbf{x}} - \boldsymbol{\omega}_{\mathbf{x}}^\star \rangle + \langle -\nabla_{\mathbf{y}} H(\boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star), \tilde{\mathbf{y}} - \boldsymbol{\omega}_{\mathbf{y}}^\star \rangle \\
&\ge \langle \nabla F(\boldsymbol{\omega}_{\mathbf{x}}^\star) + \nabla_{\mathbf{x}} H(\boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star), \tilde{\mathbf{x}} - \boldsymbol{\omega}_{\mathbf{x}}^\star \rangle + \langle \nabla G(\boldsymbol{\omega}_{\mathbf{y}}^\star) - \nabla_{\mathbf{y}} H(\boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star), \tilde{\mathbf{y}} - \boldsymbol{\omega}_{\mathbf{y}}^\star \rangle = 0
\end{aligned}$$

and also

$$\begin{aligned}
&\mathbb{E}[V(\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \mid \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \\
&\le \langle \nabla F(\boldsymbol{\omega}_{\mathbf{x}}^\star) + \nabla_{\mathbf{x}} H(\boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star), \tilde{\mathbf{x}} - \boldsymbol{\omega}_{\mathbf{x}}^\star \rangle + \langle \nabla G(\boldsymbol{\omega}_{\mathbf{y}}^\star) - \nabla_{\mathbf{y}} H(\boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star), \tilde{\mathbf{y}} - \boldsymbol{\omega}_{\mathbf{y}}^\star \rangle + \tfrac{L_{\mathrm{Str}}}{2} \mathcal{S}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star) \\
&= \tfrac{L_{\mathrm{Str}}}{2} \mathcal{S}(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)
\end{aligned}$$

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

$$\begin{aligned}
&\frac{\mu_\star}{2} \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \le \mathbb{E}[V(\mathbf{x}_{t-\frac{1}{2}}^{\mathrm{ag}}, \mathbf{y}_{t-\frac{1}{2}}^{\mathrm{ag}} \mid \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star) + \frac{\mu_\star}{2} \mathbb{E}[\mathcal{S}(\mathbf{x}_t, \mathbf{y}_t; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star)] \\
&\le \left(V(\mathbf{x}_{-\frac{1}{2}}^{\mathrm{ag}}, \mathbf{y}_{-\frac{1}{2}}^{\mathrm{ag}} \mid \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star) + \frac{\mu_\star}{2} \mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star) \right) \prod_{\tau=1}^{t} (1 - \alpha_\tau) + \sum_{\tau=1}^{t} \frac{3\alpha_\tau^2}{2\mu_\star} \left[\prod_{\tau'=\tau+1}^{t} (1 - \alpha_{\tau'}) \right] \sigma^2 \\
&\le \mathcal{S}(\mathbf{x}_0, \mathbf{y}_0; \boldsymbol{\omega}_{\mathbf{x}}^\star, \boldsymbol{\omega}_{\mathbf{y}}^\star) \frac{L_{\mathrm{Str}} + \mu_\star}{2} \prod_{\tau=1}^{t} (1 - \alpha_\tau) + \frac{3\sigma^2}{2\mu_\star} \sum_{\tau=1}^{t} \alpha_\tau^2 \prod_{\tau'=\tau+1}^{t} (1 - \alpha_{\tau'})
\end{aligned}$$

Dividing both sides by $\frac{\mu_\star}{2}$ gives (17) and our theorem.

B.5. Proof of Theorem 4

Before the proof we first adopt the scaling reduction argument as in §B.1, to argue that we only need to prove the result for the case of bilinear games centered at zero, i.e. $F(\mathbf{x}) = 0 = G(\mathbf{y})$ where from (4) we have $L_{\mathrm{Str}} = \mu_{\mathrm{Str}} = \mu_F = 0$. We set the iteration symbol $\mathbf{z} \equiv \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{x} - \boldsymbol{\omega}_{\mathbf{x}}^\star \\ \mathbf{y} - \boldsymbol{\omega}_{\mathbf{y}}^\star \end{bmatrix}$ and also $\hat{\mathscr{F}}(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = \hat{\mathbf{x}}^\top \mathbf{B} \hat{\mathbf{y}}$, with $\hat{\mathscr{F}}(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ being equal to $\mathscr{F}(\mathbf{x}, \mathbf{y})$ defined as in (7) up to an additive constant. Our scaling-reduction argument hence applies.

Proof [Proof of Theorem 4] From the update rule we have

$$\mathbf{z}_{t-\frac{1}{2}} = \mathbf{z}_{t-1} - \eta \mathbf{J} \mathbf{z}_{t-1} + \eta \boldsymbol{\varepsilon}_{t-\frac{1}{2}}, \tag{41a}$$

$$\mathbf{z}_{t-\frac{1}{2}}^{\mathrm{ag}} = \tfrac{t-1}{t+1} \mathbf{z}_{t-\frac{3}{2}}^{\mathrm{ag}} + \tfrac{2}{t+1} \mathbf{z}_{t-\frac{1}{2}}, \tag{41b}$$

$$\mathbf{z}_t = \mathbf{z}_{t-1} - \eta \mathbf{J} \mathbf{z}_{t-\frac{1}{2}} + \eta \boldsymbol{\varepsilon}_t. \tag{41c}$$

Note the $[\mathbf{x}_t^{\mathrm{md}}; \mathbf{y}_t^{\mathrm{md}}]$ sequence becomes irrelevant in this update 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

$$\mathbb{E}\|\mathbf{z}_t\|^2 \le \mathbb{E}\|\mathbf{z}_0\|^2 + 2t\eta^2 \sigma_{\mathrm{Bil}}^2 \tag{42}$$

Note (41a) and (41c) together gives

$$\mathbf{z}_t = \left(\mathbf{I} - \eta\mathbf{J} + \eta^2\mathbf{J}^2\right)\mathbf{z}_{t-1} - \eta^2\mathbf{J}\boldsymbol{\varepsilon}_{t-\frac{1}{2}} + \eta\boldsymbol{\varepsilon}_t \tag{43}$$

Taking squared norm on both sides of (43), 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

$$\begin{aligned}
&\mathbb{E}\|\mathbf{z}_t\|^2 = \mathbb{E}\left[(\mathbf{z}_{t-1})^\top\left(\mathbf{I} + \eta^2\mathbf{J}^2 + \eta^4\mathbf{J}^4\right)\mathbf{z}_{t-1}\right] + \mathbb{E}\left\|-\eta^2\mathbf{J}\boldsymbol{\varepsilon}_{t-\frac{1}{2}} + \eta\boldsymbol{\varepsilon}_t\right\|^2 \\
&\le \mathbb{E}\|\mathbf{z}_{t-1}\|^2 + \mathbb{E}\left\|\eta^2\mathbf{J}\boldsymbol{\varepsilon}_{t-\frac{1}{2}}\right\|^2 + \mathbb{E}\left\|\eta\boldsymbol{\varepsilon}_t\right\|^2 \\
&\le \mathbb{E}\|\mathbf{z}_{t-1}\|^2 + \eta^2\left(1 + \eta^2\lambda_{\max}(\mathbf{B}^\top\mathbf{B})\right)\sigma_{\mathrm{Bil}}^2 \le \mathbb{E}\|\mathbf{z}_{t-1}\|^2 + 2\eta^2\sigma_{\mathrm{Bil}}^2
\end{aligned} \tag{44}$$

Recursively applying the above concludes (42).

Step 2. We start from the update rule (41b) 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 $t = 1, \ldots, \mathscr{T}$, so

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

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

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

which yields

$$\mathbf{z}^{\mathrm{ag}}_{\mathscr{T}-\frac{1}{2}} = \frac{2}{(\mathscr{T}+1)\mathscr{T}}\sum_{t=1}^{\mathscr{T}} t\mathbf{z}_{t-\frac{1}{2}} = \frac{2}{-\eta(\mathscr{T}+1)\mathscr{T}}\mathbf{J}^{-1}\left(\mathscr{T}\mathbf{z}_{\mathscr{T}} - \sum_{t=1}^{\mathscr{T}}\mathbf{z}_{t-1} - \eta\sum_{t=1}^{\mathscr{T}} t\boldsymbol{\varepsilon}_t\right) \tag{45}$$

Obviously the least singular value of the matrix $\mathbf{J}$ can be lower-bounded as $\sigma_{\min}(\mathbf{J}) \ge \sqrt{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$. We conclude from (45) along with Young's inequality that

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

where applying the last-iterate bound (42) together with some elementary estimates leads to

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

and, using the property of square-integrable martingales,

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

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

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

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

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

Dividing both sides by $\sqrt{\lambda_{\min}(\mathbf{B}\mathbf{B}^\top)}$ and taking squares conclude (8) and hence the theorem. ■

Appendix C. Proof of auxiliary lemmas

C.1. Proof of Lemma 11

The analysis in this subsection is partially motivated by Lemma 2 of Chen et al. (2017).
Proof [Proof of Lemma 11] By definition of $\boldsymbol{\delta}_1,\boldsymbol{\delta}_2$, we have for any $\mathbf{z}\in\mathbb{R}^d$

$$
\langle\boldsymbol{\delta}_1,\boldsymbol{\varphi}_1-\mathbf{z}\rangle=\frac{1}{2}\left[\|\boldsymbol{\theta}-\mathbf{z}\|^2-\|\boldsymbol{\theta}-\boldsymbol{\varphi}_1\|^2-\|\boldsymbol{\varphi}_1-\mathbf{z}\|^2\right]\tag{46}
$$

$$
\langle\boldsymbol{\delta}_2,\boldsymbol{\varphi}_2-\mathbf{z}\rangle=\frac{1}{2}\left[\|\boldsymbol{\theta}-\mathbf{z}\|^2-\|\boldsymbol{\theta}-\boldsymbol{\varphi}_2\|^2-\|\boldsymbol{\varphi}_2-\mathbf{z}\|^2\right]\tag{47}
$$

Specifically, letting $\mathbf{z}=\boldsymbol{\varphi}_2$ in (46) we have

$$
\langle\boldsymbol{\delta}_1,\boldsymbol{\varphi}_1-\boldsymbol{\varphi}_2\rangle=\frac{1}{2}\left[\|\boldsymbol{\theta}-\boldsymbol{\varphi}_2\|^2-\|\boldsymbol{\theta}-\boldsymbol{\varphi}_1\|^2-\|\boldsymbol{\varphi}_1-\boldsymbol{\varphi}_2\|^2\right]\tag{48}
$$

Now, combining inequalities (47) and (48) we have

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

which in turn gives

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

An application of the Young and Cauchy-Schwartz inequalities gives

$$\begin{aligned}
&\langle \boldsymbol{\delta}_2, \boldsymbol{\varphi}_1 - \mathbf{z} \rangle \le \|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\| \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\| + \frac{1}{2}\left[\|\boldsymbol{\theta} - \mathbf{z}\|^2 - \|\boldsymbol{\varphi}_2 - \mathbf{z}\|^2 - \|\boldsymbol{\theta} - \boldsymbol{\varphi}_1\|^2 - \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|^2\right] \\
&\le \frac{1}{2}\|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\|^2 + \frac{1}{2}\|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|^2 + \frac{1}{2}\left[\|\boldsymbol{\theta} - \mathbf{z}\|^2 - \|\boldsymbol{\varphi}_2 - \mathbf{z}\|^2 - \|\boldsymbol{\theta} - \boldsymbol{\varphi}_1\|^2 - \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|^2\right] \\
&= \frac{1}{2}\|\boldsymbol{\delta}_2 - \boldsymbol{\delta}_1\|^2 + \frac{1}{2}\left[\|\boldsymbol{\theta} - \mathbf{z}\|^2 - \|\boldsymbol{\varphi}_2 - \mathbf{z}\|^2 - \|\boldsymbol{\theta} - \boldsymbol{\varphi}_1\|^2\right]
\end{aligned} \tag{49}$$

This establishes (21) and hence Lemma 11. ■

C.2. Proof of Lemma 12

Proof [Proof of Lemma 12] It is straightforward to verify that $F(\mathbf{x})$ and $G(\mathbf{y})$ are L_{Str}-smooth and μ_{Str}-strongly convex. For the rest of this proof, we observe that the saddle definition of $\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}$ satisfies the first-order stationary condition for problem (1):

$$\nabla_{\mathbf{x}} \mathscr{F}(\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}) = \nabla F(\boldsymbol{\omega}_{\mathbf{x}}^{\star}) + \nabla_{\mathbf{x}} H(\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}) = 0, \quad \nabla_{\mathbf{y}} \mathscr{F}(\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}) = \nabla_{\mathbf{y}} H(\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}) - \nabla G(\boldsymbol{\omega}_{\mathbf{y}}^{\star}) = 0 \tag{50}$$

Since both $f(\mathbf{x})$ and $g(\mathbf{y})$ are μ_{Str}-strongly convex, we have

$$\begin{aligned}
&F(\mathbf{x}) - F(\boldsymbol{\omega}_{\mathbf{x}}^{\star}) + \left\langle \nabla_{\mathbf{x}} H(\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}), \mathbf{x} - \boldsymbol{\omega}_{\mathbf{x}}^{\star} \right\rangle \\
&\ge \langle \nabla F(\boldsymbol{\omega}_{\mathbf{x}}^{\star}), \mathbf{x} - \boldsymbol{\omega}_{\mathbf{x}}^{\star} \rangle + \frac{\mu_{\text{Str}}}{2}\|\mathbf{x} - \boldsymbol{\omega}_{\mathbf{x}}^{\star}\|^2 + \left\langle \nabla_{\mathbf{x}} H(\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}), \mathbf{x} - \boldsymbol{\omega}_{\mathbf{x}}^{\star} \right\rangle \\
&= \left\langle \nabla F(\boldsymbol{\omega}_{\mathbf{x}}^{\star}) + \nabla_{\mathbf{x}} H(\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}), \mathbf{x} - \boldsymbol{\omega}_{\mathbf{x}}^{\star} \right\rangle + \frac{\mu_{\text{Str}}}{2}\|\mathbf{x} - \boldsymbol{\omega}_{\mathbf{x}}^{\star}\|^2 = \frac{\mu_{\text{Str}}}{2}\|\mathbf{x} - \boldsymbol{\omega}_{\mathbf{x}}^{\star}\|^2
\end{aligned}$$

and

$$\begin{aligned}
&G(\mathbf{y}) - G(\boldsymbol{\omega}_{\mathbf{y}}^{\star}) - \left\langle \nabla_{\mathbf{y}} H(\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}), \mathbf{y} - \boldsymbol{\omega}_{\mathbf{y}}^{\star} \right\rangle \\
&\ge \left\langle \nabla G(\boldsymbol{\omega}_{\mathbf{y}}^{\star}), \mathbf{y} - \boldsymbol{\omega}_{\mathbf{y}}^{\star} \right\rangle + \frac{\mu_{\text{Str}}}{2}\left\|\mathbf{y} - \boldsymbol{\omega}_{\mathbf{y}}^{\star}\right\|^2 - \left\langle \nabla_{\mathbf{y}} H(\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}), \mathbf{y} - \boldsymbol{\omega}_{\mathbf{y}}^{\star} \right\rangle \\
&= -\left\langle \nabla_{\mathbf{y}} H(\boldsymbol{\omega}_{\mathbf{x}}^{\star}, \boldsymbol{\omega}_{\mathbf{y}}^{\star}) - \nabla G(\boldsymbol{\omega}_{\mathbf{y}}^{\star}), \mathbf{y} - \boldsymbol{\omega}_{\mathbf{y}}^{\star} \right\rangle + \frac{\mu_{\text{Str}}}{2}\left\|\mathbf{y} - \boldsymbol{\omega}_{\mathbf{y}}^{\star}\right\|^2 = \frac{\mu_{\text{Str}}}{2}\left\|\mathbf{y} - \boldsymbol{\omega}_{\mathbf{y}}^{\star}\right\|^2
\end{aligned}$$

where in both of the two displays, the inequality holds due to the μ_{Str}-strong convexity of F and G, and the equality holds due to the first-order stationary condition (50). This completes the proof. ■

C.3. Proof of Lemma 13

Proof [Proof of Lemma 13] Items (i)—(iii) are straightforward. For the proof of (26) in item (iv), we note that $\eta_t = \bar{\eta}_t(\sigma; \mathscr{T}, \mathscr{C}, r, \beta) \le \frac{t}{\frac{2}{r}L_{\mathrm{Str}} + \sqrt{\frac{1+\beta}{r}}L_{\mathrm{Bil}}t} \le \frac{1}{\sqrt{\frac{1+\beta}{r}}L_{\mathrm{Bil}}}$ which gives

$$r - \frac{2L_{\mathrm{Str}}}{t+1}\eta_t - (1+\beta)L_{\mathrm{Bil}}^2\eta_t^2 \ge \frac{r}{t}\left(t - \left(\frac{2}{r}L_{\mathrm{Str}} + \sqrt{\tfrac{1+\beta}{r}}L_{\mathrm{Bil}}t\right)\eta_t\right) \ge 0$$

and hence completes the proof. ■