A Stochastic Momentum Method for Min-max Bilevel Optimization

Quanqi Hu QUANQI-HU@UIOWA.EDU
Bokun Wang BOKUN-WANG@UIOWA.EDU
Tianbao Yang TIANBAO-YANG@UIOWA.EDU
The University of Iowa, USA

Abstract

In this paper, we study nonconvex min-max bilevel optimization problem where the outer objective function is non-convex and strongly concave and the inner objective function is strongly convex. This paper develops a single loop single timescale stochastic algorithm based on moving average estimator, which only requires a general unbiased stochastic oracle with bounded variance. To the best of our knowledge, the only existing work on min-max bilevel optimization focuses on the ones with an upper objective in certain structure and only achieves an oracle complexity of $\mathcal{O}(\epsilon^{-5})$. Under some mild assumptions on the partial derivatives of both outer and inner objective functions, we provide the first convergence guarantee with an oracle complexity of $\mathcal{O}(\epsilon^{-4})$ for a general class of min-max bilevel problems, which matches the optimal complexity order for solving stochastic nonconvex optimization under a general unbiased stochastic oracle model.

1. Introduction

We consider stochastic min-max bilevel optimization problems given by

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^{d_x}} \max_{\alpha\in\mathcal{A}} f(\mathbf{x},\alpha,\mathbf{y}(\mathbf{x})) = \mathbb{E}_\xi[f(\mathbf{x},\alpha,\mathbf{y}(\mathbf{x});\xi)] \qquad \text{(upper)} \\ &\mathbf{y}(\mathbf{x}) \in \arg\min_{\mathbf{y}\in\mathbb{R}^{d_y}} g(\mathbf{x},\mathbf{y}) = \mathbb{E}_\zeta[g(\mathbf{x},\mathbf{y};\zeta)] \qquad \text{(lower)} \end{aligned} \tag{1}$$

where f and g are smooth functions and $\mathcal{A} \subset \mathbb{R}^{d_\alpha}$ is a convex set. In particular, in this paper we assume that $f(\mathbf{x},\alpha,\mathbf{y})$ is nonconvex in the primal variable $\mathbf{x}$ but strongly concave in the dual variable α, $g(\mathbf{x},\mathbf{y})$ is strongly convex in $\mathbf{y}$. Problem (1) involves two optimization problems and has a two level structure. We refer to $\min_{\mathbf{x}\in\mathbb{R}^{d_x}} \max_{\alpha\in\mathcal{A}} f(\mathbf{x},\alpha,\mathbf{y}(\mathbf{x}))$ as the *upper problem* and $\min_{\mathbf{y}\in\mathbb{R}^{d_y}} g(\mathbf{x},\mathbf{y})$ as the *lower problem*. We call $f(\mathbf{x},\alpha,\mathbf{y}(\mathbf{x}))$ the *outer objective* and $g(\mathbf{x},\mathbf{y})$ the *inner objective*. Tackling the problem (1) is challenging as it involves solving a min-max problem and a coupled min problem simultaneously.

1.1. Related work

1.1.1. Min-max Bilevel optimization

To the best of our knowledge, the only existing work that provides a stochastic algorithm with provable convergence guarantee on min-max bilevel problems is [9]. They propose a single loop bi-time scale stochastic algorithm based on gradient descent ascent, and prove that it converges to an ϵ-stationary point with an oracle complexity of $\mathcal{O}(\epsilon^{-5})$. However, this convergence result is only established for a special case where $f(\mathbf{x},\cdot,\mathbf{y})$ is a linear function.

1.1.2. STOCHASTIC NONCONVEX STRONGLY CONCAVE MIN-MAX PROBLEMS

The considered problem is also closely related to non-convex strongly concave min-max problems, which have been studied extensively recently. To the best of our knowledge, [19] establishes the first results by proposing proximally guided stochastic mirror descent and variance-reduced gradient algorithms (PG-SMD/PG-SVRG) for solving non-smooth nonconvex concave min-max problems. They prove a convergence to nearly stationary point of the primal objective function with an oracle complexity in the order of $\mathcal{O}(\epsilon^{-6})$, which can be reduced to $\mathcal{O}(\epsilon^{-4})$ if the objective function is strongly concave in the dual variable and has certain special structure. The same order of oracle complexity is achieved in [22] without relying on any special structure. These two works use two-loop algorithms. There are some studies focusing on single-loop algorithms. [16] analyzes a single-loop stochastic gradient descent ascent method for smooth nonconvex strongly convave problem, which achieves $\mathcal{O}(\epsilon^{-4})$ complexity but with a large mini-batch size. In [11], the same order of complexity is achieved by a momentum method employing the stochastic moving average estimator (SEMA) without a large mini-batch size. Some recent works are trying to improve the complexity by leveraging the individual smoothness condition. In particular, an improved complexity of $\mathcal{O}(\epsilon^{-3})$ was achieved in several recent works under the Lipschitz continuous oracle model for the stochastic gradient [13, 17, 20].

1.1.3. STOCHASTIC NONCONVEX BILEVEL OPTIMIZATION

The considered problem belongs to a general family of non-convex bilevel optimization problems. Non-asymptotic convergence results for stochastic nonconvex bilevel optimization with strongly convex lower problem has been established in several recent studies [3, 8, 11, 12, 14]. As the one who gives the first results for this problem, [8] proposes a double-loop algorithm with $\mathcal{O}(\epsilon^{-6})$ oracle complexity for finding an ϵ-stationary point of the objective function. [14] improves the complexity order to $\mathcal{O}(\epsilon^{-4})$, but suffer from a large mini-batch size. [12] proposes a single-loop algorithm with two time-scale updates that achieves an oracle complexity of $\widetilde{\mathcal{O}}(\epsilon^{-5})$. Recently, [11] improves the oracle complexity to the state-of-the-art oracle complexity $\widetilde{\mathcal{O}}(\epsilon^{-4})$ by proposing a single-loop algorithm based on SEMA estimator. [4] unifies several SGD-type updates for stochastic nested problems into a single SGD approach and presents a new analysis showing that an improved sample complexity $\mathcal{O}(\epsilon^{-4})$ can be achieved for SGD-type methods. There are studies that try to further improve the complexity by leveraging the Lipschitz continuous conditions of stochastic oracles. In particular, by employing the variance reduction technique STORM [6] in gradient estimations, [10] achieves complexity of $\mathcal{O}(\epsilon^{-3})$ without large mini-batch size. [15] proposed a single timescale algorithm based on a double-momentum structure that achieves not only the same oracle complexity, but a reduced per-iteration complexity. However, none of these works tackle the min-max bilevel optimization problem directly.

1.2. Contributions

We present a single loop single timescale stochastic method based on the SEMA for solving a general form of min-max bilevel optimization problem under the nonconvex strongly concave (upper) strongly convex (lower) setting. Then we show, theoretically, that it converges to ϵ-stationary point with complexity $\mathcal{O}(\epsilon^{-4})$ under a general unbiased stochastic oracle model. This oracle complexity surpasses the existing work [9] and matches the optimal complexity order for solving stochastic

nonconvex optimization under a general unbiased stochastic oracle model [2].

2. Algorithm

Notations. Let $\|\cdot\|$ denote the Euclidean norm of a vector or the spectral norm of a matrix. For a twice differentiable function $f : X \times Y \to \mathbb{R}$, $\nabla_x f(x,y)$ (resp. $\nabla_y f(x,y)$) denotes its partial gradient taken with respect to x (resp. y), and $\nabla_{xy} f(x,y)$ (resp. $\nabla_{yy} f(x,y)$) denotes the Jacobian of $\nabla_x f(x,y)$ at y (resp. $\nabla_y f(x,y)$ at y). A mapping $f : X \to \mathbb{R}$ is L-Lipschitz continuous iff $\|f(x) - f(x')\| \leq L\|x - x'\|\ \forall x, x' \in X$. A function f is L-smooth iff its gradient $\nabla f(\cdot)$ is L-Lipschitz continuous. A function $g : X \to \mathbb{R}$ is λ-strongly convex iff $\forall x, x' \in X$, $g(x) \geq g(x') + \nabla g(x')^T(x - x') + \frac{\lambda}{2}\|x - x'\|^2$. A function $g : X \to \mathbb{R}$ is λ-strongly concave iff $-g(x)$ is λ-strongly convex. Let $\Pi_{\mathcal{A}}$ denote a projection onto a convex set $\mathcal{A}$.

We state the definition of ϵ-stationary point as following.

Definition 1 *Consider a differentiable function $F(\boldsymbol{x})$, a point $\boldsymbol{x}$ is called ϵ-stationary if $\|\nabla F(\boldsymbol{x})\| \leq \epsilon$. A stochastic algorithm is said to achieve an ϵ-stationary point in t iterations if $\mathbb{E}[\|\nabla F(\boldsymbol{x}_t)\|] \leq \epsilon$, where the expectation is taken over the stochasticity of the algorithm until the iteration t.*

Assumptions. Before presenting our algorithm, we make the following well-behaving assumptions.

Assumption 2 *For function f and g, we assume that the following conditions hold*

- *$f(\boldsymbol{x}, \alpha, \boldsymbol{y})$ is μ_f-strongly concave with respect to α for any fixed $\boldsymbol{x}, \boldsymbol{y}$, and $g(\boldsymbol{x}, \boldsymbol{y})$ is μ_g-strongly convex with respect to $\boldsymbol{y}$ for any fixed $\boldsymbol{x}$.*
- *$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \alpha, \boldsymbol{y})$, $\nabla_\alpha f(\boldsymbol{x}, \alpha, \boldsymbol{y})$, $\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \alpha, \boldsymbol{y})$ are L_f-Lipschitz continuous, with respect to $(\boldsymbol{x}, \alpha, \boldsymbol{y})$. $\nabla_{\boldsymbol{x}} g(\boldsymbol{x}, \boldsymbol{y})$, $\nabla_{\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y})$ are L_g-Lipschitz continuous, with respect to $(\boldsymbol{x}, \boldsymbol{y})$. $\nabla^2_{\boldsymbol{xy}} g(\boldsymbol{x}, \boldsymbol{y})$, $\nabla^2_{\boldsymbol{yy}} g(\boldsymbol{x}, \boldsymbol{y})$ are L_{gxy}, L_{gyy}-Lipschitz continuous respectively, with respect to $(\boldsymbol{x}, \boldsymbol{y})$.*
- *$\|\nabla_\alpha f(\boldsymbol{x}, \alpha, \boldsymbol{y})\|^2 \leq C^2_{f\alpha}$, $\|\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \alpha, \boldsymbol{y})\|^2 \leq C^2_{fy}$, $\|\nabla^2_{\boldsymbol{xy}} g(\boldsymbol{x}, \boldsymbol{y})\|^2 \leq C^2_{gxy}$, $0 \preceq \mathcal{O}_{\boldsymbol{gyy}}(\boldsymbol{x}, \boldsymbol{y}) \preceq C_{gyy} I$.*

Moreover, the gradients of functions f and g can be only accessed through unbiased oracles with bounded variance.

Assumption 3 *$\mathcal{O}_{f\boldsymbol{x}}, \mathcal{O}_{f\alpha}, \mathcal{O}_{f\boldsymbol{y}}, \mathcal{O}_{g\boldsymbol{y}}, \mathcal{O}_{g\boldsymbol{xy}}, \mathcal{O}_{g\boldsymbol{yy}}$ are unbiased stochastic oracles of $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \alpha, \boldsymbol{y})$, $\nabla_\alpha f(\boldsymbol{x}, \alpha, \boldsymbol{y})$, $\nabla_{\boldsymbol{y}} f(\boldsymbol{x}, \alpha, \boldsymbol{y})$, $\nabla_{\boldsymbol{y}} g(\boldsymbol{x}, \boldsymbol{y})$, $\nabla_{\boldsymbol{xy}} g(\boldsymbol{x}, \boldsymbol{y})$, $\nabla_{\boldsymbol{yy}} g(\boldsymbol{xy})$, and their variances are bounded by σ^2.*

The proposed algorithm is constructed by employing the SEMA for updating $\mathbf{x}, \mathbf{y}$ and α. Algorithms based on moving average estimators have achieved the state-of-the-art oracle complexity in both min-max and bilevel optimizations [11]. We first give a brief introduction to the SEMA estimator. For solving a nonconvex minimization problem $\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x})$ though an unbiased oracle $\mathcal{O}_F(\mathbf{x})$, i.e. $\mathbb{E}[O_F(\mathbf{x})] = \nabla F(\mathbf{x})$, the stochastic momentum method (stochastic heavy-ball method) that employs SEMA updates is given by

$$\begin{aligned} \mathbf{v}_{t+1} &= (1 - \beta)\mathbf{v}_t - \beta \mathcal{O}_F(\mathbf{x}_t) \\ \mathbf{x}_{t+1} &= \mathbf{x}_t - \eta \mathbf{v}_{t+1}, \end{aligned}$$

where β is the momentum parameter and η is known as step size or learning rate. The variance recursion property of the SEMA estimator (Lemma 9 in the appendix) plays a key role in the convergence analysis.

Algorithm 1: Stochastic Momentum Method for Min-max Bilevel Optimization

Input: $\mathbf{v}_0, \alpha_0, H_0, \mathbf{z}_0, \mathbf{x}_0, \mathbf{w}_0, y_0$

for *t=0,1,...,T-1* **do**

$\mathbf{v}_{t+1} = (1-\beta_\alpha)\mathbf{v}_t + \beta_\alpha \mathcal{O}_{f\alpha}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)$;

$\alpha_{t+1} = (1-\eta_\alpha)\alpha_t + \eta_\alpha \Pi_{\mathcal{A}}[\alpha_t + \tau_\alpha \mathbf{v}_{t+1}]$;

$\mathbf{w}_{t+1} = (1-\beta_y)\mathbf{w}_t + \beta_y \, \mathcal{O}_{g\mathbf{y}}(\mathbf{x}_t, \mathbf{y}_t)$;

$\mathbf{y}_{t+1} = \mathbf{y}_t - \eta_y \mathbf{w}_{t+1}$;

$H_{t+1} = \frac{k_t}{C_{gyy}} \prod_{i=1}^{q} \left(I - \frac{1}{C_{gyy}} \mathcal{O}_{g\mathbf{yy},i}(\mathbf{x}_t, \mathbf{y}_t) \right)$, q : uniformly sampled from $\{1, \ldots, k_t\}$;

$\mathcal{O}_t = \mathcal{O}_{f\mathbf{x}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathcal{O}_{g\mathbf{xy}}(\mathbf{x}_t, \mathbf{y}_t) H_{t+1} \mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)$;

$\mathbf{z}_{t+1} = (1-\beta_x)\mathbf{z}_t + \beta_x \mathcal{O}_t$;

$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_x \mathbf{z}_{t+1}$;

end

We present the proposed method for solving problem (1) in Algorithm 1. The procedure for each iteration is as following: first, update the dual variable α in the upper problem and the variable $\mathbf{y}$ in the lower problem using SEMA estimators. Second, we approximate the inverse of the Hessians $\nabla^2_{\mathbf{yy}} g(\mathbf{x}_t, \mathbf{y}_t)$ by H_{t+1}. Here $\mathcal{O}_{g\mathbf{yy},i}(\mathbf{x}_t, \mathbf{y}_t)$ denotes the output of the oracle $\mathcal{O}_{g\mathbf{yy}}(\mathbf{x}_t, \mathbf{y}_t)$ with a data point randomly sampled from the dataset. Note that such approximation has been widely used in previous studies to avoid directly computing matrix inverses [1, 8, 11]. Finally, we perform a SEMA update in $\mathbf{z}_{t+1} = (1-\beta_x)\mathbf{z}_t + \beta_x \mathcal{O}_t$ in order to reduce the variance of biased estimator $\mathcal{O}_t$.

To understand the biased estimator $\mathcal{O}_t$, first define $F(\mathbf{x}) := f(\mathbf{x}, \alpha(\mathbf{x}), \mathbf{y}(\mathbf{x}))$, where $\alpha(\mathbf{x}) = \arg\max_{\alpha \in \mathcal{A}} f(\mathbf{x}, \alpha, \mathbf{y}(\mathbf{x}))$, then the upper problem can be written equivalently as $\min_{\mathbf{x} \in \mathbb{R}^{d_x}} F(\mathbf{x})$. Due to the strong concavity of $f(\mathbf{x}, \cdot, \mathbf{y})$ and convexity of $\mathcal{A}$, one may apply Lemma A.5 in [18] to obtain

$$\begin{aligned}\nabla F(\mathbf{x}) &= \nabla_{\mathbf{x}} f(\mathbf{x}, \alpha(\mathbf{x}), \mathbf{y}(\mathbf{x})) + \nabla_{\mathbf{x}} \mathbf{y}(\mathbf{x})^T \nabla_{\mathbf{y}} f(\mathbf{x}, \alpha(\mathbf{x}), \mathbf{y}(\mathbf{x})) \\ &= \nabla_{\mathbf{x}} f(\mathbf{x}, \alpha(\mathbf{x}), \mathbf{y}(\mathbf{x})) - \nabla^2_{\mathbf{xy}} g(\mathbf{x}, \mathbf{y}(\mathbf{x}))[\nabla^2_{\mathbf{yy}} g(\mathbf{x}, \mathbf{y}(\mathbf{x}))]^{-1} \nabla_{\mathbf{y}} f(\mathbf{x}, \alpha(\mathbf{x}), \mathbf{y}(\mathbf{x})),\end{aligned}$$

where the standard result $\nabla_{\mathbf{x}} \mathbf{y}(\mathbf{x})^T = -\nabla^2_{\mathbf{xy}} g(\mathbf{x}, \mathbf{y}(\mathbf{x}))[\nabla^2_{\mathbf{yy}} g(\mathbf{x}, \mathbf{y}(\mathbf{x}))]^{-1}$ in the literature of bilevel optimization [8] is used in the second equality. Define

$$\nabla F(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) := \nabla_{\mathbf{x}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \nabla^2_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}_t)[\nabla^2_{\mathbf{yy}} g(\mathbf{x}_t, \mathbf{y}_t)]^{-1} \nabla_{\mathbf{y}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t).$$

as an approximation of $\nabla F(\mathbf{x}_t)$. Then, with unbiased estimator for each term in $\nabla F(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)$ except $[\nabla^2_{\mathbf{yy}} g(\mathbf{x}_t, \mathbf{y}_t)]^{-1}$, for which we have biased estimator H_{t+1}, we have the biased estimator $\mathcal{O}_t$ of $\nabla F(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)$. We have the following convergence result regarding to Algorithm 1.

Theorem 4 (Informal) *Under Assumption 2,3 and considering Algorithm 1, for all $\epsilon > 0$, with τ_α, τ_y small enough and step sizes $\eta_\alpha, \eta_y, \eta_x, \beta_\alpha, \beta_y, \beta_x = \mathcal{O}(\epsilon^2)$, we have the following convergence guarantee in $T = \mathcal{O}(\epsilon^{-4})$ iterations*

$$\frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|\nabla F(\boldsymbol{x}_t)\|^2] \leq \epsilon^2, \quad \frac{1}{T+1} \sum_{t=0}^{T} \mathbb{E}[\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{z}_{t+1}\|^2] \leq 2\epsilon^2.$$

Remark. It is clear that the oracle compelxity of Algorithm 1 is $\mathcal{O}(\epsilon^{-4})$, which matches the state-of-the-art complexity for solving nonconvex strongly concave min-max problems and nonconvex bilevel problems. In fact, as shown in [2], $\mathcal{O}(\epsilon^{-4})$ is the optimal oracle complexity for solving stochastic nonconvex optimization under a general unbiased stochastic oracle model.

2.1. Application in Robust Meta Learning

Meta learning aims to train a model on a variety of learning tasks such that a small number of training data from a new task will produce good generalization performance on that task. However, as shown in [5], having a simple average loss across all tasks as the objective function, like the popular Model-agnostic meta-learning (MAML) [7], the worst task performance is not well controlled. Motivated by this drawback, [5] considers the task-robust MAML in min-max optimization formulation. In particular, the average loss in the objective is replaced by a weighted loss, and the maximization is taken over the weights parameter. Similarly to the idea presented in [9], we consider K tasks, each of which has a corresponding loss function $f_i(\mathbf{x}, \mathbf{y}_i; \xi_i)$ and an inner loss function $g_i(\mathbf{x}, \mathbf{y}_i; \zeta_i)$. Here $\mathbf{x}$ is the shared parameter and $\mathbf{y}_i$ is the task specific parameter. The data samples ζ_i, ξ_i are taken from the training dataset $\mathcal{S}_i$ and testing dataset $\mathcal{D}_i$ respectively. Then the goal is to solve the following min-max bilevel optimization problem

$$\begin{aligned}
&\min_{\mathbf{x}} \max_{\alpha \in \Delta_K} \sum_{i=1}^{K} \alpha_i \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[f_i(\mathbf{x}, \mathbf{y}_i(\mathbf{x}); \xi_i)] - \lambda \mathrm{KL}(\alpha, \frac{\mathbf{1}}{K}) \\
&\mathbf{y}_i(\mathbf{x}) = \arg\min_{\mathbf{y}_i \in \mathbb{R}^{d_y}} \mathbb{E}_{\zeta_i \sim \mathcal{S}_i}[g_i(\mathbf{x}, \mathbf{y}_i; \zeta_i)] + R(\mathbf{y}_i),\ i = 1, 2, \ldots, K,
\end{aligned}$$

where $\Delta_K = \{\alpha \in \mathbb{R}^K : \sum_i \alpha_i = 1, \alpha_i \geq 0\}$ is a K-dimensional simplex, the negative KL-divergence term $-\lambda \mathrm{KL}(\alpha, \mathbf{1}/K) = -\lambda \sum_{i=1}^{K} \alpha_i \log(K\alpha_i)$ in the outer objective ensures the strong concavity of the outer objective in dual variable α that makes the function smooth in terms of $\mathbf{x}$, and $R(\cdot)$ is a strongly convex regularizer. Applying the proposed algorithm to the above problem yields a complexity of $O(1/\epsilon^4)$. In contrast, [9] considers the same problem with $\lambda = 0$ and suffers a complexity of $O(K/\epsilon^5)$. Although the two results are not directly comparable, we can see that adding the negative KL-divergence term not only increase the modeling flexbility but also enjoys a faster convergence speed by our algorithm.

3. Conclusion and Future Work

We have developed a new single loop single timescale stochastic algorithm for solving a family of min-max bilevel optimization problems. We showed that it achieves an oracle complexity of $\mathcal{O}(\epsilon^{-4})$, which surpasses the existing work and matches the optimal complexity order for solving stochastic nonconvex optimization under a general unbiased stochastic oracle model. Since our study is focused on the algorithm development and theoretical convergence analysis, this work is lack of cogent experimental results. In the future, we plan to conduct experiments to evaluate its empirical performance. At the same time, we hope our work inspires others to find more novel applications of our idea.

References

[1] Naman Agarwal, Brian Bullins, and Elad Hazan. Second-order stochastic optimization for machine learning in linear time. *J. Mach. Learn. Res.*, 18:116:1–116:40, 2017.

[2] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

[3] Tianyi Chen, Yuejiao Sun, and Wotao Yin. A single-timescale stochastic bilevel optimization method. *arXiv preprint arXiv:2102.04671*, 2021.

[4] Tianyi Chen, Yuejiao Sun, and Wotao Yin. Tighter analysis of alternating stochastic gradient method for stochastic nested problems, 2021.

[5] Liam Collins, Aryan Mokhtari, and Sanjay Shakkottai. Task-robust model-agnostic meta-learning, 2020.

[6] Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pages 15236–15245, 2019.

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR, 2017. URL http://proceedings.mlr.press/v70/finn17a.html.

[8] Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

[9] Alex Gu, Songtao Lu, Parikshit Ram, and Lily Weng. Nonconvex min-max bilevel optimization for task robust meta learning. In *Beyond first order methods in machine learning systems*, 2021.

[10] Zhishuai Guo, Quanqi Hu, Lijun Zhang, and Tianbao Yang. Randomized stochastic variance-reduced methods for multi-task stochastic bilevel optimization. *arXiv preprint arXiv:2105.02266*, 2021.

[11] Zhishuai Guo, Yi Xu, Wotao Yin, Rong Jin, and Tianbao Yang. On stochastic moving-average estimators for non-convex optimization. *CoRR*, abs/2104.14840, 2021. URL https://arxiv.org/abs/2104.14840.

[12] Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

[13] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order momentum methods from mini to minimax optimization. *arXiv preprint arXiv:2008.08170*, 2020.

[14] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Provably faster algorithms for bilevel optimization and applications to meta-learning. *arXiv preprint arXiv:2010.07962*, 2020.

[15] Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum, 2021.

[16] Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

[17] Luo Luo, Haishan Ye, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *CoRR*, abs/2001.03724, 2020.

[18] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D. Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods, 2019.

[19] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *CoRR*, abs/1810.02060, 2018.

[20] Quoc Tran-Dinh, Deyi Liu, and Lam M. Nguyen. Hybrid variance-reduced sgd algorithms for minimax problems with nonconvex-linear function. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[21] Mengdi Wang, Ethan X Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Mathematical Programming*, 161(1-2):419–449, 2017.

[22] Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

Appendix A. Convergence analysis

First of all, we give the formal statement of Theorem 4.

Theorem 5 *Let $F(\boldsymbol{x}_0) - F(\boldsymbol{x}^*) \leq \Delta_F$. Under Assumption 2,3 and considering Algorithm 1, with*

$\tau_\alpha \leq 1/(3L_f)$, $\tau_y \leq 1/(3L_g)$, $\beta_x \leq \frac{\epsilon^2}{108C_1}$, $\beta_\alpha \leq \frac{\epsilon^2\mu_f^2}{13824C_\alpha\sigma^2}$, $\beta_y \leq \frac{\epsilon^2\mu_g^2}{6912(2C_\alpha L_\alpha^2+C_{\boldsymbol{y}})\sigma^2}$,

$$\eta_\alpha^2 \leq \min\left\{\frac{\mu_f\beta_\alpha^2}{64\tau_\alpha L_f^2}, \frac{\beta_x C_\alpha}{6\tau_\alpha\mu_f L_F^2}, \frac{4}{\tau_\alpha^2\mu_f^2}\right\},$$

$$\eta_y^2 \leq \min\left\{\frac{\mu_f\beta_\alpha^2(6C_\alpha L_\alpha^2+3C_{\boldsymbol{y}})}{768\tau_y\mu_g C_\alpha L_f^2}, \frac{\eta_\alpha^2\tau_\alpha^2\mu_f^2(6C_\alpha L_\alpha^2+3C_{\boldsymbol{y}})}{768\tau_y\mu_g C_\alpha L_\alpha^2}, \frac{\mu_g\beta_y^2}{128\tau_y L_g^2}, \frac{\beta_x(2C_\alpha L_\alpha^2+C_{\boldsymbol{y}})}{24\tau_y\mu_g L_F^2}, \frac{4}{\tau_y^2\mu_g^2}\right\},$$

$$\eta_x \leq \min\left\{\frac{\mu_f\beta_\alpha}{62\sqrt{C_\alpha}L_f}, \frac{\eta_\alpha\tau_\alpha\mu_f}{62\sqrt{C_\alpha}L_\alpha}, \frac{\mu_g\beta_y}{44\sqrt{2C_\alpha L_\alpha^2+C_{\boldsymbol{y}}}L_g}, \frac{\eta_y\tau_y\mu_g}{44\sqrt{2C_\alpha L_\alpha^2+C_{\boldsymbol{y}}}L_y}, \frac{\sqrt{\beta_x}}{19L_F}, \frac{1}{2L_F}\right\},$$

$$k_t \geq \frac{C_{gyy}}{2\mu_g}\log\left(\frac{16C_{gxy}^2C_{fy}^2}{\mu_g^2\epsilon^2}\right),$$

$$T \geq \max\left\{\frac{144\Delta_F}{\eta_x\epsilon^2}, \frac{1728C_\alpha\|\alpha_0-\alpha(\boldsymbol{x}_0,\boldsymbol{y}_0)\|^2}{\eta_\alpha\tau_\alpha\mu_f\epsilon^2}, \frac{13824C_\alpha\delta_{f\alpha,0}}{\mu_f^2\beta_\alpha\epsilon^2}, \frac{864(2C_\alpha L_\alpha^2+C_{\boldsymbol{y}})\delta_{\boldsymbol{y},0}}{\eta_y\tau_y\mu_g\epsilon^2}, \frac{6912(2C_\alpha L_\alpha^2+C_{\boldsymbol{y}})\delta_{gy,0}}{\mu_g^2\beta_y\epsilon^2}, \frac{216\delta_{z,0}}{\beta_x\epsilon^2}\right\},$$

where C_1 is a constant defined in the proof of this Theorem and $C_\alpha, C_{\boldsymbol{y}}$ are constants defined in the proof of Lemma 7, we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\boldsymbol{x}_t)\|^2] \leq \epsilon^2, \quad \frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{z}_{t+1}\|^2] \leq 2\epsilon^2$$

To prove Theorem 5, we need the following lemmas.

Lemma 6 *Consider the update in Algorithm 1, where $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_{\boldsymbol{x}}\boldsymbol{z}_{t+1}$. With $\eta_{\boldsymbol{x}}L_F \leq \frac{1}{2}$, we have*

$$F(\boldsymbol{x}_{t+1}) \leq F(\boldsymbol{x}_t) + \frac{\eta_{\boldsymbol{x}}}{2}\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{z}_{t+1}\|^2 - \frac{\eta_{\boldsymbol{x}}}{2}\|\nabla F(\boldsymbol{x}_t)\|^2 - \frac{\eta_{\boldsymbol{x}}}{4}\|\boldsymbol{z}_{t+1}\|^2.$$

Lemma 7 *Under Assumption 2 and considering Algorithm 1, for all nonnegative integer t, we have*

$$\|\nabla F(\boldsymbol{x}_t) - \boldsymbol{z}_{t+1}\|^2 \leq 3C_\alpha\|\alpha_t - \alpha(\boldsymbol{x}_t)\|^2 + 3C_{\boldsymbol{y}}\|\boldsymbol{y}_t - \boldsymbol{y}(\boldsymbol{x}_t)\|^2 + 3\|\nabla F(\boldsymbol{x}_t, \alpha_t, \boldsymbol{y}_t) - z_{t+1}\|^2, \quad (2)$$

where $C_\alpha C_{\boldsymbol{y}}$ are constants defined in the proof.

Lemma 8 *[Lemma 4.3 [16]] Under Assumption 2, $\boldsymbol{y}(\boldsymbol{x})$ is $L_{\boldsymbol{y}}$-Lipschitz-continuous with $L_{\boldsymbol{y}} = L_g/\mu_g$. Define $\alpha(\boldsymbol{x},\boldsymbol{y}) := \arg\max_{\alpha\in\mathcal{A}} f(\boldsymbol{x},\alpha,\boldsymbol{y})$. Then $\alpha(\boldsymbol{x},\boldsymbol{y})$ is $L_\alpha = L_f/\mu_f$-Lipschitz continuous in the sense that*

$$\|\alpha(\boldsymbol{x}_t,\boldsymbol{y}_t) - \alpha(\boldsymbol{x}_{t+1},\boldsymbol{y}_{t+1})\|^2 \leq L_\alpha^2(\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 + \|\boldsymbol{y}_t - \boldsymbol{y}_{t+1}\|^2).$$

Lemma 9 *[Lemma 2 [21]] (Variance recursion of SEMA) Considering a moving average sequence $\boldsymbol{z}_{t+1} = (1-\beta_t)\boldsymbol{z}_t + \beta_t\mathcal{O}_h(\boldsymbol{x}_t)$ for tracking $h(\boldsymbol{x}_t)$, where $\mathbb{E}_t[\mathcal{O}_h(\boldsymbol{x}_t)] = h(\boldsymbol{x}_t)$ and h is a L-Lipschitz continuous mapping. Then we have*

$$\mathbb{E}_t[\|\boldsymbol{z}_{t+1} - h(\boldsymbol{x}_t)\|^2] \leq (1-\beta_t)\|\boldsymbol{z}_t - h(\boldsymbol{x}_{t-1})\|^2 + 2\beta_t^2\mathbb{E}_t[\|\mathcal{O}_h(\boldsymbol{x}_t) - h(\boldsymbol{x}_t)\|^2] + \frac{L^2\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\|^2}{\beta_t},$$

where $\mathbb{E}_t$ denotes the expectation conditioned on all randomness before $\mathcal{O}_h(\boldsymbol{x}_t)$.

Lemma 10 *[Lemma 5 [11]] Consider update* $\alpha_{t+1} = (1-\eta_\alpha)\alpha_t + \eta_\alpha \Pi_{\mathcal{A}}[\alpha_t + \tau_\alpha \boldsymbol{v}_{t+1}]$ *in Algorithm 1. With* $\tau_\alpha \leq 1/(3L_f)$ *and* $\eta_\alpha \tau_\alpha \mu_f \leq 2$*, we have*

$$\begin{aligned}\|\alpha_{t+1} - \alpha(\boldsymbol{x}_{t+1}, \boldsymbol{y}_{t+1})\|^2 \leq \left(1 - \frac{\eta_\alpha \tau_\alpha \mu_f}{4}\right)\|\alpha_t - \alpha(\boldsymbol{x}_t, \boldsymbol{y}_t)\|^2 + \frac{8\eta_\alpha \tau_\alpha}{\mu_f}\|\nabla_\alpha f(\boldsymbol{x}_t, \alpha_t, \boldsymbol{y}_t) - \boldsymbol{v}_{t+1}\|^2 \\ - \frac{2\tau_\alpha}{\eta_\alpha}(1 + \frac{\eta_\alpha \tau_\alpha \mu_f}{4})(\frac{1}{2\tau_\alpha} - \frac{3L_f}{4})\|\alpha_t - \alpha_{t+1}\|^2 + \frac{8L_\alpha^2 \eta_x^2}{\eta_\alpha \tau_\alpha \mu_f}\|\boldsymbol{z}_{t+1}\|^2 \\ + \frac{8L_\alpha^2}{\eta_\alpha \tau_\alpha \mu_f}\|\boldsymbol{y}_t - \boldsymbol{y}_{t+1}\|^2.\end{aligned}$$

Lemma 11 *[Lemma 13 [11]] Consider update* $\boldsymbol{y}_{t+1} = (1-\eta_y)\boldsymbol{y}_t + \eta_y \Pi_{\mathcal{Y}}[\boldsymbol{y}_t - \tau_y \boldsymbol{w}_{t+1}]$ *where* $\mathcal{Y} \subset \mathbb{R}^{d_y}$ *is convex. With* $\tau_y \leq 1/(3L_g)$*,* $\eta_y \tau_y \mu_g \leq 2$*, we have*

$$\begin{aligned}\|\boldsymbol{y}_{t+1} - \boldsymbol{y}(\boldsymbol{x}_{t+1})\|^2 \leq \left(1 - \frac{\eta_y \tau_y \mu_g}{4}\right)\|\boldsymbol{y}_t - \boldsymbol{y}(\boldsymbol{x}_t)\|^2 + \frac{8\eta_y \tau_y}{\mu_g}\|\nabla_{\boldsymbol{y}} g(\boldsymbol{x}_t, \boldsymbol{y}_t) - \boldsymbol{w}_{t+1}\|^2 \\ - \frac{2\tau_y}{\eta_y}(1 + \frac{\eta_y \tau_y \mu_g}{4})(\frac{1}{2\tau_y} - \frac{3L_g}{4})\|\boldsymbol{y}_t - \boldsymbol{y}_{t+1}\|^2 + \frac{8L_{\boldsymbol{y}}^2 \eta_x^2}{\eta_y \tau_y \mu_g}\|\boldsymbol{z}_{t+1}\|^2.\end{aligned}$$

Note that in Algorithm 1, since $\mathcal{Y} = \mathbb{R}^{d_y}$, the projection is not needed. Thus the update for $\mathbf{y}_t$ can be simplified to $\mathbf{y}_{t+1} = \mathbf{y}_t - \eta_y \mathbf{w}_{t+1}$, where the constant τ_y is absorbed in the step size η_y.

Lemma 12 *[[11][8]] Under Assumption 2,3 and considering update of Algorithm 1, we have*

$$\|\mathbb{E}[H_{t+1}] - [\nabla^2_{\boldsymbol{yy}} g(\boldsymbol{x}_t, \boldsymbol{y}_t)]^{-1}\| \leq \frac{1}{\mu_g}\left(1 - \frac{\mu_g}{C_{gyy}}\right)^{k_t}$$

$$\|H_{t+1}\| \leq \frac{k_t}{C_{gyy}}, \quad \|[\nabla^2_{\mathbf{yy}} g(\mathbf{x}_t, \mathbf{y}_t)]^{-1} - H_{t+1}\| \leq \frac{1}{\mu_g} + \frac{k_t}{C_{gyy}}.$$

A.1. Proof of Theorem 5

Proof First, we bound the last term of the RHS of 2. Since H_{t+1} is an biased estimator of $[\nabla^2_{\mathbf{yy}} g(\mathbf{x}_t, \mathbf{y}_t)]^{-1}$, we cannot directly apply Lemma 9 to $\mathbf{z}_{t+1}$. Define

$$\widehat{\nabla} F(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) = \nabla_{\mathbf{x}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}_t)\mathbb{E}[H_{t+1}]\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t).$$

Denote $\delta_{\mathbf{z},t} := \|\nabla F(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{z}_{t+1}\|^2$, $\delta_{\alpha,t} := \|\alpha_t - \alpha(\mathbf{x}_t)\|^2$ and $\delta_{\mathbf{y},t} := \|\mathbf{y}_t - \mathbf{y}(\mathbf{x}_t)\|^2$. By the update rule of $\mathbf{z}_{t+1}$,

$$
\begin{aligned}
\mathbb{E}[\delta_{\mathbf{z},t}] &= \mathbb{E}[\|\nabla F(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{z}_{t+1}\|^2] \\
&= \mathbb{E}[\|\nabla F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - (1-\beta_x)\mathbf{z}_t - \beta_x(\mathcal{O}_{f\mathbf{x}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathcal{O}_{g\mathbf{xy}}(\mathbf{x}_t, \mathbf{y}_t)H_{t+1}\mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t))\|^2] \\
&= \mathbb{E}[\|(1-\beta_x)(\nabla F(\mathbf{x}_{t-1}, \alpha_{t-1}, \mathbf{y}_{t-1}) - \mathbf{z}_t) + (1-\beta_x)(\nabla F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - \nabla F(\mathbf{x}_{t-1}, \alpha_{t-1}\, \mathbf{y}_{t-1})) \\
&\quad + \beta_x(\nabla F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - \widehat{\nabla} F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t)) + \beta_x(\widehat{\nabla} F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - (\mathcal{O}_{f\mathbf{x}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) \\
&\quad - \mathcal{O}_{g\mathbf{xy}}(\mathbf{x}_t, \mathbf{y}_t)H_{t+1}\mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)))\|^2] \\
&= \mathbb{E}[\|(1-\beta_x)(\nabla F(\mathbf{x}_{t-1}, \alpha_{t-1}, \mathbf{y}_{t-1}) - \mathbf{z}_t) + (1-\beta_x)(\nabla F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - \nabla F(\mathbf{x}_{t-1}, \alpha_{t-1}\, \mathbf{y}_{t-1})) \\
&\quad + \beta_x(\nabla F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - \widehat{\nabla} F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t))\|^2] + \beta_x^2\mathbb{E}[\|\widehat{\nabla} F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - (\mathcal{O}_{f\mathbf{x}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) \\
&\quad - \mathcal{O}_{g\mathbf{xy}}(\mathbf{x}_t, \mathbf{y}_t)H_{t+1}\mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t))\|^2] \\
&\le (1+\beta_x)(1-\beta_x)^2\mathbb{E}[\|(\nabla F(\mathbf{x}_{t-1}, \alpha_{t-1}, \mathbf{y}_{t-1}) - \mathbf{z}_t\|^2] \\
&\quad + 2\left(1 + \frac{1}{\beta_x}\right)\mathbb{E}[\|\nabla F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - \nabla F(\mathbf{x}_{t-1}, \alpha_{t-1}\, \mathbf{y}_{t-1})\|^2 + \beta_x^2\|\nabla F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - \widehat{\nabla} F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t)\|^2] \\
&\quad + \beta_x^2\mathbb{E}[\|\widehat{\nabla} F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - (\mathcal{O}_{f\mathbf{x}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathcal{O}_{g\mathbf{xy}}(\mathbf{x}_t, \mathbf{y}_t)H_{t+1}\mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t))\|^2]
\end{aligned} \tag{3}
$$

where the last equality follows from the definition of $\widehat{\nabla} F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t)$, in the sense that

$$
\widehat{\nabla} F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) = \mathbb{E}[\mathcal{O}_{f\mathbf{x}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) + \mathcal{O}_{g\mathbf{xy}}(\mathbf{x}_t, \mathbf{y}_t)H_{t+1}\mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)].
$$

The last two terms of 3 can be bounded as

$$
\begin{aligned}
\mathbb{E}[\|\nabla F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - \widehat{\nabla} F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t)\|^2] &\le C_{gxy}^2\|\mathbb{E}[H_{t+1}] - [\nabla_{\mathbf{yy}}^2 g(\mathbf{x}_t, \mathbf{y}_t)]^{-1}\|^2 C_{fy}^2 \\
&\le \frac{C_{gxy}^2 C_{fy}^2}{\mu_g^2}\left(1 - \frac{\mu_g}{C_{gyy}}\right)^{2k_t},
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}[\|\widehat{\nabla} F(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - (\mathcal{O}_{f\mathbf{x}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathcal{O}_{g\mathbf{xy}}(\mathbf{x}_t, \mathbf{y}_t)H_{t+1}\mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t))\|^2] \\
&\quad \le 2\mathbb{E}[\|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \alpha_t\, \mathbf{y}_t) - (\mathcal{O}_{f\mathbf{x}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)\|^2] \\
&\qquad + 6\mathbb{E}[\|\nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}_t)\mathbb{E}[H_{t+1}](\nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) - \mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t))\|^2] \\
&\qquad + 6\mathbb{E}[\|\nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}_t)(\mathbb{E}[H_{t+1}] - H_{t+1})\mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)\|^2] \\
&\qquad + 6\mathbb{E}[\|(\nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}_t) - \mathcal{O}_{g\mathbf{xy}}(\mathbf{x}_t, \mathbf{y}_t))H_{t+1}\mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)\|^2] \\
&\quad \le 2\sigma^2 + 6C_{gxy}^2\frac{k_t^2}{C_{gyy}^2}\sigma^2 + 24C_{gxy}^2\frac{k_t^2}{C_{gyy}^2}(C_{fy}^2 + \sigma^2) + 6\sigma^2\frac{k_t^2}{C_{gyy}^2}(C_{fy}^2 + \sigma^2) =: C_1,
\end{aligned}
$$

where we use Lemma 12 and $\mathbb{E}[\|\mathcal{O}_{f\mathbf{y}}(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)\|^2] \le C_{fy}^2 + \sigma^2$. Thus,

$$
\begin{aligned}
\mathbb{E}[\delta_{\mathbf{z},t}] \le (1-\beta_x)\mathbb{E}[\delta_{\mathbf{z},t-1}] &+ \frac{4}{\beta_x}L_F^2\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2 + \|\alpha_t - \alpha_{t-1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t-1}\|^2] \\
&+ 4\beta_x\frac{C_{gxy}^2 C_{fy}^2}{\mu_g^2}\left(1 - \frac{\mu_g}{C_{gyy}}\right)^{2k_t} + \beta_x^2 C_1,
\end{aligned}
$$

where $L_F^2 = 2L_f^2 + 6\left(C_{gxy}^2 \frac{1}{\mu_g^2} L_f^2 + C_{gxy}^2 \frac{L_{gyy}^2}{\mu_g^4} C_{fy}^2 + L_{gxy}^2 \frac{1}{\mu_g^2} C_{fy}^2\right)$.

Setting $k_t \geq \frac{C_{gyy}}{2\mu_g} \log(16 C_{gxy}^2 C_{fy}^2/(\mu_g^2 \epsilon^2))$, we have $4C_{gxy}^2 C_{fy}^2/\mu_g \left(1 - \frac{\mu_g}{C_{gyy}}\right)^{2k_t} \leq \epsilon^2/4$. Therefore,

$$\sum_{t=0}^{T} \mathbb{E}[\delta_{\mathbf{z},t}] \leq \frac{\delta_{\mathbf{z},0}}{\beta_x} + \frac{4L_F^2}{\beta_x^2} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\alpha_t - \alpha_{t+1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2] + \frac{\epsilon^2 T}{4} + \beta_x C_1 T.$$

Denote $\delta_{g\mathbf{y},t} := \|\mathbf{w}_{t+1} - \nabla_{\mathbf{y}} g(\mathbf{x}_t, \mathbf{y}_t)\|^2$ and $\delta_{f\alpha,t} := \|\mathbf{v}_{t+1} - \nabla_{\alpha} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)\|^2$. Applying Lemma 9 to $\mathbf{w}_{t+1}$ and $\mathbf{v}_{t+1}$, we get

$$\mathbb{E}[\delta_{g\mathbf{y},t+1}] \leq (1-\beta_y)\mathbb{E}[\delta_{g\mathbf{y},t}] + 2\beta_y^2\sigma^2 + \frac{L_g^2(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2)}{\beta_y}$$

and

$$\mathbb{E}[\delta_{f\alpha,t+1}] \leq (1-\beta_\alpha)\mathbb{E}[\delta_{f\alpha,t}] + 2\beta_\alpha^2\sigma^2 + \frac{L_f^2(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\alpha_t - \alpha_{t+1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2)}{\beta_\alpha}.$$

By taking telescopic sum, we obtain

$$\sum_{t=0}^{T} \mathbb{E}[\delta_{g\mathbf{y},t}] \leq \frac{\delta_{g\mathbf{y},0}}{\beta_y} + 2\beta_y\sigma^2 T + \frac{L_g^2\eta_x^2}{\beta_y^2} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{z}_{t+1}\|^2] + \frac{L_g^2}{\beta_y^2} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2]$$

and

$$\begin{aligned}\sum_{t=0}^{T} \mathbb{E}[\delta_{f\alpha,t}] \leq \frac{\delta_{f\alpha,0}}{\beta_\alpha} + 2\beta_\alpha\sigma^2 T + \frac{L_f^2\eta_x^2}{\beta_\alpha^2} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{z}_{t+1}\|^2] + \frac{L_f^2}{\beta_\alpha^2} \sum_{t=0}^{T-1} \mathbb{E}[\|\alpha_t - \alpha_{t+1}\|^2] \\ + \frac{L_f^2}{\beta_\alpha^2} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2]\end{aligned}$$

Then Lemma 11 and Lemma 10 with telescopic sum give

$$\begin{aligned}\sum_{t=0}^{T} \mathbb{E}[\delta_{\mathbf{y},t}] &\leq \frac{4}{\eta_y\tau_y\mu_g}\delta_{\mathbf{y},0} + \frac{32}{\mu_g^2} \sum_{t=1}^{T} \mathbb{E}[\delta_{g\mathbf{y},t}] + \frac{32L_{\mathbf{y}}^2\eta_x^2}{\eta_y^2\tau_y^2\mu_g^2} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{z}_{t+1}\|^2] - \frac{1}{\eta_y^2\tau_y\mu_g} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2] \\ &\leq \frac{4}{\eta_y\tau_y\mu_g}\delta_{\mathbf{y},0} + \frac{32}{\mu_g^2\beta_y}\delta_{g\mathbf{y},0} + \frac{64\beta_y\sigma^2 T}{\mu_g^2} + \left(\frac{32L_g^2}{\mu_g^2\beta_y^2} - \frac{1}{\eta_y^2\tau_y\mu_g}\right) \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2] \\ &\quad + \left(\frac{32L_g^2\eta_x^2}{\mu_g^2\beta_y^2} + \frac{32L_{\mathbf{y}}^2\eta_x^2}{\eta_y^2\tau_y^2\mu_g^2}\right) \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{z}_{t+1}\|^2]\end{aligned}$$

and

$$\begin{aligned}
\sum_{t=0}^{T}\mathbb{E}[\|\alpha_t-\alpha(\mathbf{x}_t,\mathbf{y}_t)\|^2] &\le \frac{4}{\eta_\alpha\tau_\alpha\mu_f}\|\alpha_0-\alpha(\mathbf{x}_0,\mathbf{y}_0)\|^2+\frac{32}{\mu_f^2}\sum_{t=1}^{T}\mathbb{E}[\delta_{f\alpha,t}]+\frac{32L_\alpha^2\eta_x^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2]\\
&\quad-\frac{1}{\eta_\alpha^2\tau_\alpha\mu_f}\sum_{t=0}^{T-1}\mathbb{E}[\|\alpha_t-\alpha_{t+1}\|^2]+\frac{32L_\alpha^2\eta_x^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2]\\
&\le \frac{4}{\eta_\alpha\tau_\alpha\mu_f}\|\alpha_0-\alpha(\mathbf{x}_0,\mathbf{y}_0)\|^2+\frac{32}{\mu_f^2\beta_\alpha}\delta_{f\alpha,0}+\frac{64\beta_\alpha\sigma^2T}{\mu_f^2}\\
&\quad+\left(\frac{32L_f^2}{\mu_f^2\beta_\alpha^2}-\frac{1}{\eta_\alpha^2\tau_\alpha\mu_f}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\alpha_t-\alpha_{t+1}\|^2]+\left(\frac{32L_f^2}{\mu_f^2\beta_\alpha^2}+\frac{32L_\alpha^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2]\\
&\quad+\left(\frac{32L_f^2\eta_x^2}{\mu_f^2\beta_\alpha^2}+\frac{32L_\alpha^2\eta_x^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2].
\end{aligned}$$

Note that $\alpha(\mathbf{x}) = \alpha(\mathbf{x}, \mathbf{y}(\mathbf{x}))$. By the L_α-smoothness of $\alpha(\mathbf{x}, \mathbf{y})$, we get

$$
\begin{aligned}
\sum_{t=0}^{T} \mathbb{E}[\delta_{\alpha,t}] &\leq 2\sum_{t=0}^{T} \mathbb{E}[\|\alpha_t - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2] + 2\sum_{t=0}^{T} E[\|\alpha(\mathbf{x}_t, \mathbf{y}_t) - \alpha(\mathbf{x}_t)\|^2] \\
&\leq \frac{8}{\eta_\alpha \tau_\alpha \mu_f}\|\alpha_0 - \alpha(\mathbf{x}_0, \mathbf{y}_0)\|^2 + \frac{64}{\mu_f^2 \beta_\alpha}\delta_{f\alpha,0} + \frac{128\beta_\alpha \sigma^2 T}{\mu_f^2} \\
&\quad + \left(\frac{64L_f^2}{\mu_f^2\beta_\alpha^2} - \frac{2}{\eta_\alpha^2\tau_\alpha\mu_f}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\alpha_t - \alpha_{t+1}\|^2] + \left(\frac{64L_f^2}{\mu_f^2\beta_\alpha^2} + \frac{64L_\alpha^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2] \\
&\quad + \left(\frac{64L_f^2\eta_x^2}{\mu_f^2\beta_\alpha^2} + \frac{64L_\alpha^2\eta_x^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2] + 2L_\alpha^2\sum_{t=0}^{T}\mathbb{E}[\delta_{\mathbf{y},t}] \\
&\leq \frac{8}{\eta_\alpha \tau_\alpha \mu_f}\|\alpha_0 - \alpha(\mathbf{x}_0, \mathbf{y}_0)\|^2 + \frac{64}{\mu_f^2 \beta_\alpha}\delta_{f\alpha,0} + \frac{128\beta_\alpha \sigma^2 T}{\mu_f^2} \\
&\quad + \left(\frac{64L_f^2}{\mu_f^2\beta_\alpha^2} - \frac{2}{\eta_\alpha^2\tau_\alpha\mu_f}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\alpha_t - \alpha_{t+1}\|^2] + \left(\frac{64L_f^2}{\mu_f^2\beta_\alpha^2} + \frac{64L_\alpha^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2] \\
&\quad + \left(\frac{64L_f^2\eta_x^2}{\mu_f^2\beta_\alpha^2} + \frac{64L_\alpha^2\eta_x^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2] + 2L_\alpha^2\left(\frac{4}{\eta_y\tau_y\mu_g}\delta_{\mathbf{y},0} + \frac{32}{\mu_g^2\beta_y}\delta_{g\mathbf{y},0} + \frac{64\beta_y\sigma^2 T}{\mu_g^2}\right. \\
&\quad \left. + \left(\frac{32L_g^2}{\mu_g^2\beta_y^2} - \frac{1}{\eta_y^2\tau_y\mu_g}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2] + \left(\frac{32L_g^2\eta_x^2}{\mu_g^2\beta_y^2} + \frac{32L_{\mathbf{y}}^2\eta_x^2}{\eta_y^2\tau_y^2\mu_g^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2]\right) \\
&\leq \frac{8}{\eta_\alpha \tau_\alpha \mu_f}\|\alpha_0 - \alpha(\mathbf{x}_0, \mathbf{y}_0)\|^2 + \frac{64}{\mu_f^2 \beta_\alpha}\delta_{f\alpha,0} + \frac{128\beta_\alpha \sigma^2 T}{\mu_f^2} + \frac{8L_\alpha^2}{\eta_y\tau_y\mu_g}\delta_{\mathbf{y},0} \\
&\quad + \frac{64L_\alpha^2}{\mu_g^2\beta_y}\delta_{g\mathbf{y},0} + \frac{128L_\alpha^2\beta_y\sigma^2 T}{\mu_g^2} + \left(\frac{64L_f^2}{\mu_f^2\beta_\alpha^2} - \frac{2}{\eta_\alpha^2\tau_\alpha\mu_f}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\alpha_t - \alpha_{t+1}\|^2] \\
&\quad + \left(\frac{64L_f^2}{\mu_f^2\beta_\alpha^2} + \frac{64L_\alpha^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2} + \frac{64L_\alpha^2L_g^2}{\mu_g^2\beta_y^2} - \frac{2L_\alpha^2}{\eta_y^2\tau_y\mu_g}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2] \\
&\quad + \left(\frac{64L_f^2\eta_x^2}{\mu_f^2\beta_\alpha^2} + \frac{64L_\alpha^2\eta_x^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2} + \frac{64L_\alpha^2L_g^2\eta_x^2}{\mu_g^2\beta_y^2} + \frac{64L_\alpha^2L_{\mathbf{y}}^2\eta_x^2}{\eta_y^2\tau_y^2\mu_g^2}\right)\sum_{t=0}^{T-1}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2].
\end{aligned}
$$

Then Plug the bounds we obtained for $\sum_{t=0}^{T}\mathbb{E}[\delta_{\alpha,t}]$, $\sum_{t=0}^{T}\mathbb{E}[\delta_{\mathbf{y},t}]$ and $\sum_{t=0}^{T}\mathbb{E}[\delta_{\mathbf{z},t}]$ into Lemma 7 to get

$$\begin{aligned}
\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2] &\le 3C_\alpha\sum_{t=0}^{T}\mathbb{E}[\delta_{\alpha,t}]+3C_{\mathbf{y}}\sum_{t=0}^{T}\mathbb{E}[\delta_{\mathbf{y},t}]+3\sum_{t=0}^{T}\mathbb{E}[\delta_{\mathbf{z},t}]\\
&\le \frac{24C_\alpha}{\eta_\alpha\tau_\alpha\mu_f}\|\alpha_0-\alpha(\mathbf{x}_0,\mathbf{y}_0)\|^2+\frac{192C_\alpha}{\mu_f^2\beta_\alpha}\delta_{f\alpha,0}+\frac{384C_\alpha\beta_\alpha\sigma^2T}{\mu_f^2}\\
&\quad+\frac{24C_\alpha L_\alpha^2+12C_{\mathbf{y}}}{\eta_y\tau_y\mu_g}\delta_{\mathbf{y},0}+\frac{192C_\alpha L_\alpha^2+96C_{\mathbf{y}}}{\mu_g^2\beta_y}\delta_{g\mathbf{y},0}+\frac{(384C_\alpha L_\alpha^2+192C_{\mathbf{y}})\beta_y\sigma^2T}{\mu_g^2}\\
&\quad+\frac{3}{\beta_x}\delta_{\mathbf{z},0}+\frac{3\epsilon^2T}{4}+3\beta_xC_1T\\
&\quad+\left(\frac{192C_\alpha L_f^2}{\mu_f^2\beta_\alpha^2}-\frac{6C_\alpha}{\eta_\alpha^2\tau_\alpha\mu_f}+\frac{12L_F^2}{\beta_x^2}\right)\sum_{t=0}^{T}\mathbb{E}[\|\alpha_t-\alpha_{t+1}\|^2]\\
&\quad+\left(\frac{192C_\alpha L_f^2}{\mu_f^2\beta_\alpha^2}+\frac{192C_\alpha L_\alpha^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}+\frac{96(2C_\alpha L_\alpha^2+C_{\mathbf{y}})L_g^2}{\mu_g^2\beta_y^2}-\frac{6C_\alpha L_\alpha^2+3C_{\mathbf{y}}}{\eta_y^2\tau_y\mu_g}\right.\\
&\qquad\left.+\frac{12L_F^2}{\beta_x^2}\right)\sum_{t=0}^{T}\mathbb{E}[\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2]\\
&\quad+\left(\frac{192C_\alpha L_f^2\eta_x^2}{\mu_f^2\beta_\alpha^2}+\frac{192C_\alpha L_\alpha^2\eta_x^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}+\frac{96(2C_\alpha L_\alpha^2+C_{\mathbf{y}})L_g^2\eta_x^2}{\mu_g^2\beta_y^2}\right.\\
&\qquad\left.+\frac{96(2C_\alpha L_\alpha^2+C_{\mathbf{y}})L_{\mathbf{y}}^2\eta_x^2}{\eta_y^2\tau_y^2\mu_g^2}+\frac{12L_F^2\eta_x^2}{\beta_x^2}\right)\sum_{t=0}^{T}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2].
\end{aligned}$$

Applying Lemma 6,

$$\begin{aligned}
\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2] &\le \frac{2F(\mathbf{x}_0)-2F(\mathbf{x}_{T+1})}{\eta_x T}+\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2]-\frac{1}{2T}\sum_{t=0}^{T}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2] \\
&\le \frac{2F(\mathbf{x}_0)-2F(\mathbf{x}^*)}{\eta_x T}+\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2]-\frac{1}{2T}\sum_{t=0}^{T}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2] \\
&\le \frac{2F(\mathbf{x}_0)-2F(\mathbf{x}^*)}{\eta_x T}+\frac{1}{T}\left(\frac{24C_\alpha}{\eta_\alpha\tau_\alpha\mu_f}\|\alpha_0-\alpha(\mathbf{x}_0,\mathbf{y}_0)\|^2+\frac{192C_\alpha}{\mu_f^2\beta_\alpha}\delta_{f\alpha,0}\right. \\
&\quad \left.+\frac{24C_\alpha L_\alpha^2+12C_{\mathbf{y}}}{\eta_y\tau_y\mu_g}\delta_{\mathbf{y},0}+\frac{192C_\alpha L_\alpha^2+96C_{\mathbf{y}}}{\mu_g^2\beta_y}\delta_{g\mathbf{y},0}+\frac{3}{\beta_x}\delta_{\mathbf{z},0}\right) \\
&\quad +\left(3\beta_x C_1+\frac{384C_\alpha\beta_\alpha\sigma^2}{\mu_f^2}+\frac{(384C_\alpha L_\alpha^2+192C_{\mathbf{y}})\beta_y\sigma^2}{\mu_g^2}\right)+\frac{3\epsilon^2}{4} \\
&\quad +\frac{1}{T}\left(\frac{192C_\alpha L_f^2}{\mu_f^2\beta_\alpha^2}-\frac{6C_\alpha}{\eta_\alpha^2\tau_\alpha\mu_f}+\frac{12L_F^2}{\beta_x^2}\right)\sum_{t=0}^{T}\mathbb{E}[\|\alpha_t-\alpha_{t+1}\|^2] \\
&\quad +\frac{1}{T}\left(\frac{192C_\alpha L_f^2}{\mu_f^2\beta_\alpha^2}+\frac{192C_\alpha L_\alpha^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}+\frac{96(2C_\alpha L_\alpha^2+C_{\mathbf{y}})L_g^2}{\mu_g^2\beta_y^2}-\frac{6C_\alpha L_\alpha^2+3C_{\mathbf{y}}}{\eta_y^2\tau_y\mu_g}\right. \\
&\quad \left.+\frac{12L_F^2}{\beta_x^2}\right)\sum_{t=0}^{T}\mathbb{E}[\|\mathbf{y}_t-\mathbf{y}_{t+1}\|^2] \\
&\quad +\frac{1}{T}\left(\frac{192C_\alpha L_f^2\eta_x^2}{\mu_f^2\beta_\alpha^2}+\frac{192C_\alpha L_\alpha^2\eta_x^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2}+\frac{96(2C_\alpha L_\alpha^2+C_{\mathbf{y}})L_g^2\eta_x^2}{\mu_g^2\beta_y^2}\right. \\
&\quad \left.+\frac{96(2C_\alpha L_\alpha^2+C_{\mathbf{y}})L_{\mathbf{y}}^2\eta_x^2}{\eta_y^2\tau_y^2\mu_g^2}+\frac{12L_F^2\eta_x^2}{\beta_x^2}-\frac{1}{2}\right)\sum_{t=0}^{T}\mathbb{E}[\|\mathbf{z}_{t+1}\|^2]. \qquad (4)
\end{aligned}$$

By setting

$$\begin{aligned}
\eta_\alpha^2 &\le \min\left\{\frac{\mu_f\beta_\alpha^2}{64\tau_\alpha L_f^2},\frac{\beta_x^2 C_\alpha}{4\tau_\alpha\mu_f L_F^2}\right\} \\
\eta_y^2 &\le \min\left\{\frac{\mu_f\beta_\alpha^2(6C_\alpha L_\alpha^2+3C_{\mathbf{y}})}{768\tau_y\mu_g C_\alpha L_f^2},\frac{\eta_\alpha^2\tau_\alpha^2\mu_f^2(6C_\alpha L_\alpha^2+3C_{\mathbf{y}})}{768\tau_y\mu_g C_\alpha L_\alpha^2},\frac{\mu_g\beta_y^2}{128\tau_y L_g^2},\frac{\beta_x(2C_\alpha L_\alpha^2+C_{\mathbf{y}})}{16\tau_y\mu_g L_F^2}\right\} \\
\eta_x &\le \min\left\{\frac{\mu_f\beta_\alpha}{62\sqrt{C_\alpha}L_f},\frac{\eta_\alpha\tau_\alpha\mu_f}{62\sqrt{C_\alpha}L_\alpha},\frac{\mu_g\beta_y}{44\sqrt{2C_\alpha L_\alpha^2+C_{\mathbf{y}}}L_g},\frac{\eta_y\tau_y\mu_g}{44\sqrt{2C_\alpha L_\alpha^2+C_{\mathbf{y}}}L_{\mathbf{y}}},\frac{\beta_x}{16L_F}\right\},
\end{aligned}$$

we have

$$\frac{192C_\alpha L_f^2}{\mu_f^2\beta_\alpha^2} - \frac{6C_\alpha}{\eta_\alpha^2\tau_\alpha\mu_f} + \frac{12L_F^2}{\beta_x^2} \leq 0$$

$$\frac{192C_\alpha L_f^2}{\mu_f^2\beta_\alpha^2} + \frac{192C_\alpha L_\alpha^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2} + \frac{96(2C_\alpha L_\alpha^2 + C_\mathbf{y})L_g^2}{\mu_g^2\beta_y^2} - \frac{6C_\alpha L_\alpha^2 + 3C_\mathbf{y}}{\eta_y^2\tau_y\mu_g} + \frac{12L_F^2}{\beta_x^2} \leq 0$$

$$\frac{192C_\alpha L_f^2\eta_x^2}{\mu_f^2\beta_\alpha^2} + \frac{192C_\alpha L_\alpha^2\eta_x^2}{\eta_\alpha^2\tau_\alpha^2\mu_f^2} + \frac{96(2C_\alpha L_\alpha^2 + C_\mathbf{y})L_g^2\eta_x^2}{\mu_g^2\beta_y^2} + \frac{96(2C_\alpha L_\alpha^2 + C_\mathbf{y})L_\mathbf{y}^2\eta_x^2}{\eta_y^2\tau_y^2\mu_g^2} + \frac{12L_F^2\eta_x^2}{\beta_x^2} - \frac{1}{2} \leq 0,$$

Which implies that the last three terms of RHS of inequality (4) are less or equal to 0. Hence,

$$\begin{aligned}\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2] \leq \frac{2F(\mathbf{x}_0) - 2F(\mathbf{x}^*)}{\eta_x T} + \frac{1}{T}\Bigg(&\frac{24C_\alpha}{\eta_\alpha\tau_\alpha\mu_f}\|\alpha_0 - \alpha(\mathbf{x}_0,\mathbf{y}_0)\|^2 + \frac{192C_\alpha}{\mu_f^2\beta_\alpha}\delta_{f\alpha,0} \\ &+ \frac{24C_\alpha L_\alpha^2 + 12C_\mathbf{y}}{\eta_y\tau_y\mu_g}\delta_{\mathbf{y},0} + \frac{192C_\alpha L_\alpha^2 + 96C_\mathbf{y}}{\mu_g^2\beta_y}\delta_{g\mathbf{y},0} + \frac{3}{\beta_x}\delta_{\mathbf{z},0}\Bigg) \\ &+ \left(3\beta_x C_1 + \frac{384C_\alpha\beta_\alpha\sigma^2}{\mu_f^2} + \frac{(384C_\alpha L_\alpha^2 + 192C_\mathbf{y})\beta_y\sigma^2}{\mu_g^2}\right) + \frac{3\epsilon^2}{4}.\end{aligned}$$

With

$$\begin{aligned}&\beta_x \leq \frac{\epsilon^2}{108C_1}, \quad \beta_\alpha \leq \frac{\epsilon^2\mu_f^2}{13824C_\alpha\sigma^2}, \quad \beta_y \leq \frac{\epsilon^2\mu_g^2}{6912(2C_\alpha L_\alpha^2 + C_\mathbf{y})\sigma^2} \\ &T \geq \max\Bigg\{\frac{144(F(\mathbf{x}_0) - F(\mathbf{x}^*))}{\eta_x\epsilon^2}, \frac{1728C_\alpha\|\alpha_0 - \alpha(\mathbf{x}_0,\mathbf{y}_0)\|^2}{\eta_\alpha\tau_\alpha\mu_f\epsilon^2}, \frac{13824C_\alpha\delta_{f\alpha,0}}{\mu_f^2\beta_\alpha\epsilon^2}, \\ &\qquad \frac{864(2C_\alpha L_\alpha^2 + C_\mathbf{y})\delta_{\mathbf{y},0}}{\eta_y\tau_y\mu_g\epsilon^2}, \frac{6912(2C_\alpha L_\alpha^2 + C_\mathbf{y})\delta_{g\mathbf{y},0}}{\mu_g^2\beta_y\epsilon^2}, \frac{216\delta_{\mathbf{z},0}}{\beta_x\epsilon^2}\Bigg\},\end{aligned}$$

we have

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)\|^2] \leq \frac{1}{12}\epsilon^2 + \frac{1}{12}\epsilon^2 + \frac{3}{4}\epsilon^2 \leq \epsilon^2$$

Furthermore, to show the second part of the theorem, we have

$$\begin{aligned}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t) - \mathbf{z}_{t+1}\|^2] \leq \Bigg(&\frac{24C_\alpha}{\eta_\alpha\tau_\alpha\mu_f}\|\alpha_0 - \alpha(\mathbf{x}_0,\mathbf{y}_0)\|^2 + \frac{192C_\alpha}{\mu_f^2\beta_\alpha}\delta_{f\alpha,0} + \frac{24C_\alpha L_\alpha^2 + 12C_\mathbf{y}}{\eta_y\tau_y\mu_g}\delta_{\mathbf{y},0} \\ &+ \frac{192C_\alpha L_\alpha^2 + 96C_\mathbf{y}}{\mu_g^2\beta_y}\delta_{g\mathbf{y},0} + \frac{3}{\beta_x}\delta_{\mathbf{z},0}\Bigg) \\ &+ T\left(\frac{3\epsilon^2}{4} + 3\beta_x C_1 + \frac{384C_\alpha\beta_\alpha\sigma^2}{\mu_f^2} + \frac{(384C_\alpha L_\alpha^2 + 192C_\mathbf{y})\beta_y\sigma^2}{\mu_g^2}\right) \\ &+ \frac{1}{2}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x})\|^2 + \|\nabla F(\mathbf{x}) - \mathbf{z}_{t+1}\|^2].\end{aligned}$$

With parameters set above, it follows that

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}[\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2]\leq 2\epsilon^2$$

■

A.2. Proof of Lemma 6

Proof By L_F-smoothness of $F(\mathbf{x})$, with $\eta_x \leq \frac{1}{2L_F}$, we have

$$\begin{aligned}
F(\mathbf{x}_{t+1}) &\leq F(\mathbf{x}_t)+\nabla F(\mathbf{x}_t)^T(\mathbf{x}_{t+1}-\mathbf{x}_t)+\frac{L_F}{2}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\\
&= F(\mathbf{x}_t)-\eta_x\nabla F(\mathbf{x}_t)^T\mathbf{z}_{t+1}+\frac{L_F}{2}\eta_x^2\|\mathbf{z}_{t+1}\|^2\\
&= F(\mathbf{x}_t)+\frac{\eta_x}{2}\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2-\frac{\eta_x}{2}\|\nabla F(\mathbf{x}_t)\|^2+\left(\frac{L_F}{2}\eta_x^2+\frac{\eta_x}{2}\right)\|\mathbf{z}_{t+1}\|^2.
\end{aligned}$$

■

A.3. Proof of Lemma 7

Proof By the standard inequality $\|a+b+c\|^2\leq 3\|a\|^2+3\|b\|^2+3\|c\|^2$ we can split $\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2$ into three parts. For simplicity, we denote the first two terms obtained as A_1, A_2.

$$\begin{aligned}
\|\nabla F(\mathbf{x}_t)-\mathbf{z}_{t+1}\|^2 &\leq 3\|\nabla F(\mathbf{x}_t,\alpha(\mathbf{x}_t),\mathbf{y}(\mathbf{x}_t))-\nabla F(\mathbf{x}_t,\alpha_t,\mathbf{y}(\mathbf{x}_t))\|^2+3\|\nabla F(\mathbf{x}_t,\alpha_t,\mathbf{y}(\mathbf{x}_t))-\nabla F(\mathbf{x}_t,\alpha_t,\mathbf{y}_t)\|^2\\
&\quad+3\|\nabla F(\mathbf{x}_t,\alpha_t,\mathbf{y}_t)-z_{t+1}\|^2\\
&= 3A_1+3A_2+3\|\nabla F(\mathbf{x}_t,\alpha_t,\mathbf{y}_t)-z_{t+1}\|^2
\end{aligned}$$

The first term can be bounded as following

$$\begin{aligned}
A_1 &= \|\nabla_{\mathbf{x}} f(\mathbf{x}_t,\alpha(\mathbf{x}_t),\mathbf{y}(\mathbf{x}_t))-\nabla_{\mathbf{x}} f(\mathbf{x}_t,\alpha_t,\mathbf{y}(\mathbf{x}_t))\\
&\quad+\nabla_{\mathbf{xy}} g(\mathbf{x}_t,\mathbf{y}(\mathbf{x}_t))\nabla_{\mathbf{yy}}^{-1} g(\mathbf{x}_t,\mathbf{y}(\mathbf{x}_t))[\nabla_{\mathbf{y}} f(\mathbf{x}_t,\alpha_t,\mathbf{y}(\mathbf{x}_t))-\nabla_{\mathbf{y}} f(\mathbf{x}_t,\alpha(\mathbf{x}_t),\mathbf{y}(\mathbf{x}_t))]\|^2\\
&\leq 2\|\nabla_{\mathbf{x}} f(\mathbf{x}_t,\alpha(\mathbf{x}_t),\mathbf{y}(\mathbf{x}_t))-\nabla_{\mathbf{x}} f(\mathbf{x}_t,\alpha_t,\mathbf{y}(\mathbf{x}_t))\|^2\\
&\quad+2\|\nabla_{\mathbf{xy}} g(\mathbf{x}_t,\mathbf{y}(\mathbf{x}_t))\nabla_{\mathbf{yy}}^{-1} g(\mathbf{x}_t,\mathbf{y}(\mathbf{x}_t))[\nabla_{\mathbf{y}} f(\mathbf{x}_t,\alpha_t,\mathbf{y}(\mathbf{x}_t))-\nabla_{\mathbf{y}} f(\mathbf{x}_t,\alpha(\mathbf{x}_t),\mathbf{y}(\mathbf{x}_t))]\|^2\\
&\leq \left(2L_f^2+2\frac{C_{gxy}^2L_f^2}{\mu_g^2}\right)\|\alpha_t-\alpha(\mathbf{x}_t)\|^2\\
&=: C_\alpha\|\alpha_t-\alpha(\mathbf{x}_t)\|^2.
\end{aligned}$$

The second term can be bounded in a similar way,

$$\begin{aligned}
A_2 &= \|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}(\mathbf{x}_t)) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) + \nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}_t) \nabla_{\mathbf{yy}}^{-1} g(\mathbf{x}_t, \mathbf{y}_t) \nabla_{\mathbf{y}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) \\
&\quad - \nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t)) \nabla_{\mathbf{yy}}^{-1} g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t)) \nabla_{\mathbf{y}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}(\mathbf{x}_t))\|^2 \\
&\leq 2\|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}(\mathbf{x}_t)) - \nabla_{\mathbf{x}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)\|^2 + 2\|\nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}_t) \nabla_{\mathbf{yy}}^{-1} g(\mathbf{x}_t, \mathbf{y}_t) \nabla_{\mathbf{y}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) \\
&\quad - \nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t)) \nabla_{\mathbf{yy}}^{-1} g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t)) \nabla_{\mathbf{y}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}(\mathbf{x}_t))\|^2 \\
&\leq 2L_f^2 \|\mathbf{y}_t - \mathbf{y}(\mathbf{x}_t)\|^2 + 6\|[\nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t))] \nabla_{\mathbf{yy}}^{-1} g(\mathbf{x}_t, \mathbf{y}_t) \nabla_{\mathbf{y}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)\|^2 \\
&\quad + 6\|\nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t)) [\nabla_{\mathbf{yy}}^{-1} g(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{yy}}^{-1} g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t))] \nabla_{\mathbf{y}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)\|^2 \\
&\quad + 6\|\nabla_{\mathbf{xy}} g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t)) \nabla_{\mathbf{yy}}^{-1} g(\mathbf{x}_t, \mathbf{y}(\mathbf{x}_t)) [\nabla_{\mathbf{y}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t, \alpha_t, \mathbf{y}(\mathbf{x}_t))]\|^2 \\
&\leq \left(2L_f^2 + 6\frac{L_{g\mathbf{xy}}^2 C_{fy}^2}{\mu_g^2} + 6\frac{C_{gxy}^2 L_{g\mathbf{yy}}^2 C_{f_y}^2}{\mu_g^4} + 6\frac{C_{gxy}^2 L_f^2}{\mu_g^2} \right) \|\mathbf{y}_t - \mathbf{y}(\mathbf{x}_t)\|^2 \\
&=: C_{\mathbf{y}} \|\mathbf{y}_t - \mathbf{y}(\mathbf{x}_t)\|^2.
\end{aligned}$$

Combining the three inequalities above gives the desired result. ■

A.4. Proof of Lemma 10

Proof Define $\tilde{\alpha}_{t+1} = \Pi_{\mathcal{A}}[\alpha_t + \tau_\alpha v_{t+1}]$. Then $\alpha_{t+1} = \alpha_t + \eta_\alpha(\tilde{\alpha}_{t+1} - \alpha_t)$, and

$$\begin{aligned}
\|\alpha_{t+1} - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 &= \|\alpha_t + \eta_\alpha(\tilde{\alpha}_{t+1} - \alpha_t) - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&= \|\alpha_t - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \eta_\alpha^2 \|\tilde{\alpha}_{t+1} - \alpha_t\|^2 + 2\eta_\alpha (\tilde{\alpha}_{t+1} - \alpha_t)^T (\alpha_t - \alpha(\mathbf{x}_t, \mathbf{y}_t)).
\end{aligned}$$

As a result,

$$(\tilde{\alpha}_{t+1} - \alpha_t)^T (\alpha_t - \alpha(\mathbf{x}_t, \mathbf{y}_t)) = \frac{1}{2\eta_\alpha} (\|\alpha_{t+1} - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 - \|\alpha_t - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 - \eta_\alpha^2 \|\tilde{\alpha}_{t+1} - \alpha_t\|^2) \quad (5)$$

Due to L_f-smoothness of $f(\mathbf{x}, \alpha, \mathbf{y})$ with respect to α, we have

$$-f(\mathbf{x}_t, \tilde{\alpha}_{t+1}, \mathbf{y}_t) \leq -f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)^T (\tilde{\alpha}_{t+1} - \alpha_t) + \frac{L_f}{2} \|\tilde{\alpha}_{t+1} - \alpha_t\|^2.$$

Hence

$$-f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) \geq -f(\mathbf{x}_t, \tilde{\alpha}_{t+1}, \mathbf{y}_t) + \nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)^T (\tilde{\alpha}_{t+1} - \alpha_t) - \frac{L_f}{2} \|\tilde{\alpha}_{t+1} - \alpha_t\|^2.$$

Due to the μ_α-strong concavity of $f(\mathbf{x}, \alpha, \mathbf{y})$ in α, we have

$$\begin{aligned}
-f(\mathbf{x}_t, \alpha, \mathbf{y}_t) &\geq -f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)^T (\alpha - \alpha_t) + \frac{\mu_\alpha}{2} \|\alpha - \alpha_t\|^2 \\
&= -f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)^T (\alpha - \tilde{\alpha}_{t+1}) - \nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)^T (\tilde{\alpha}_{t+1} - \alpha_t) \\
&\quad + \frac{\mu_\alpha}{2} \|\alpha - \alpha_t\|^2 \\
&= -f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{v}_{t+1}^T (\alpha - \tilde{\alpha}_{t+1}) - (\nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{v}_{t+1})^T (\alpha - \tilde{\alpha}_{t+1}) \\
&\quad - \nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t)^T (\tilde{\alpha}_{t+1} - \alpha_t) + \frac{\mu_\alpha}{2} \|\alpha - \alpha_t\|^2.
\end{aligned}$$

Combine the above inequalities and we have

$$\begin{aligned}-f(\mathbf{x}_t,\alpha,\mathbf{y}_t) &\geq -f(\mathbf{x}_t,\alpha_t,\mathbf{y}_t) - \mathbf{v}_{t+1}^T(\alpha-\tilde{\alpha}_{t+1}) - (\nabla_\alpha f(\mathbf{x}_t,\alpha_t,\mathbf{y}_t)-\mathbf{v}_{t+1})^T(\alpha-\tilde{\alpha}_{t+1})\\ &\quad - \nabla_\alpha f(\mathbf{x}_t,\alpha_t,\mathbf{y}_t)^T(\tilde{\alpha}_{t+1}-\alpha_t) + \frac{\mu_\alpha}{2}\|\alpha-\alpha_t\|^2\\ &\geq -f(\mathbf{x}_t,\tilde{\alpha}_{t+1},\mathbf{y}_t) - \mathbf{v}_{t+1}^T(\alpha-\tilde{\alpha}_{t+1}) - (\nabla_\alpha f(\mathbf{x}_t,\alpha_t,\mathbf{y}_t)-\mathbf{v}_{t+1})^T(\alpha-\tilde{\alpha}_{t+1})\\ &\quad - \frac{L_f}{2}\|\tilde{\alpha}_{t+1}-\alpha_t\|^2 + \frac{\mu_\alpha}{2}\|\alpha-\alpha_t\|^2\end{aligned} \tag{6}$$

Note that $\tilde{\alpha}_{t+1} = \Pi_{\mathcal{A}}[\alpha_t + \tau_\alpha v_{t+1}] = \arg\min_{\alpha\in\mathcal{A}} \frac{1}{2}\|\alpha-\alpha_t-\tau_\alpha \mathbf{v}_{t+1}\|^2$. Since $\mathcal{A}$ is a convex set and the function $\frac{1}{2}\|\alpha-\alpha_t-\tau_\alpha\mathbf{v}_{t+1}\|^2$ is convex in α, according to the first order optimality condition of convex function, we have

$$(\tilde{\alpha}_{t+1}-\alpha_t-\tau_\alpha\mathbf{v}_{t+1})^T(\alpha-\tilde{\alpha}_{t+1}) \geq 0, \quad \forall \alpha\in\mathcal{A}.$$

Then we obtain

$$\begin{aligned}\mathbf{v}_{t+1}^T(\alpha-\tilde{\alpha}_{t+1}) &\leq \frac{1}{\tau_\alpha}(\tilde{\alpha}_{t+1}-\alpha_t)^T(\alpha-\tilde{\alpha}_{t+1})\\ &= \frac{1}{\tau_\alpha}(\tilde{\alpha}_{t+1}-\alpha_t)^T(\alpha-\alpha_t) + \frac{1}{\tau_\alpha}(\tilde{\alpha}_{t+1}-\alpha_t)^T(\alpha_t-\tilde{\alpha}_{t+1})\\ &= \frac{1}{\tau_\alpha}(\tilde{\alpha}_{t+1}-\alpha_t)^T(\alpha-\alpha_t) - \frac{1}{\tau_\alpha}\|\alpha_t-\tilde{\alpha}_{t+1}\|^2\end{aligned} \tag{7}$$

Combining (6) and (7) yields

$$\begin{aligned}-f(\mathbf{x}_t,\alpha,\mathbf{y}_t) \geq -f(\mathbf{x}_t,\tilde{\alpha}_{t+1},\mathbf{y}_t) &- \frac{1}{\tau_\alpha}(\tilde{\alpha}_{t+1}-\alpha_t)^T(\alpha-\alpha_t) - (\nabla_\alpha f(\mathbf{x}_t,\alpha_t,\mathbf{y}_t)-\mathbf{v}_{t+1})^T(\alpha-\tilde{\alpha}_{t+1})\\ &+ \frac{1}{\tau_\alpha}\|\alpha_t-\tilde{\alpha}_{t+1}\|^2 - \frac{L_f}{2}\|\tilde{\alpha}_{t+1}-\alpha_t\|^2 + \frac{\mu_\alpha}{2}\|\alpha-\alpha_t\|^2\end{aligned}$$

Take $\alpha = \alpha(\mathbf{x}_t, \mathbf{y}_t)$ and we obtain

$$\begin{aligned}
-f(\mathbf{x}_t, \tilde{\alpha}_{t+1}, \mathbf{y}_t) &\geq -f(\mathbf{x}_t, \alpha(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}_t) \\
&\geq -f(\mathbf{x}_t, \tilde{\alpha}_{t+1}, \mathbf{y}_t) - \frac{1}{\tau_\alpha}(\tilde{\alpha}_{t+1} - \alpha_t)^T(\alpha(\mathbf{x}_t, \mathbf{y}_t) - \alpha_t) \\
&\quad - (\nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{v}_{t+1})^T(\alpha(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\alpha}_{t+1}) + \frac{1}{\tau_\alpha}\|\alpha_t - \tilde{\alpha}_{t+1}\|^2 \\
&\quad - \frac{L_f}{2}\|\tilde{\alpha}_{t+1} - \alpha_t\|^2 + \frac{\mu_\alpha}{2}\|\alpha(\mathbf{x}_t, \mathbf{y}_t) - \alpha_t\|^2 \\
&\geq -f(\mathbf{x}_t, \tilde{\alpha}_{t+1}, \mathbf{y}_t) - \frac{1}{\tau_\alpha}(\tilde{\alpha}_{t+1} - \alpha_t)^T(\alpha(\mathbf{x}_t, \mathbf{y}_t) - \alpha_t) \\
&\quad - \frac{2}{\mu_\alpha}\|\nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{v}_{t+1}\|^2 - \frac{\mu_\alpha}{4}\|\alpha(\mathbf{x}_t, \mathbf{y}_t) - \alpha_t\|^2 - \frac{\mu_\alpha}{4}\|\alpha_t - \tilde{\alpha}_{t+1}\|^2 \\
&\quad + \frac{1}{\tau_\alpha}\|\alpha_t - \tilde{\alpha}_{t+1}\|^2 - \frac{L_f}{2}\|\tilde{\alpha}_{t+1} - \alpha_t\|^2 + \frac{\mu_\alpha}{2}\|\alpha(\mathbf{x}_t, \mathbf{y}_t) - \alpha_t\|^2 \\
&= -f(\mathbf{x}_t, \tilde{\alpha}_{t+1}, \mathbf{y}_t) + \frac{1}{2\eta_\alpha\tau_\alpha}(\|\alpha_{t+1} - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 - \|\alpha_t - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 - \eta_\alpha^2\|\tilde{\alpha}_{t+1} - \alpha_t\|^2) \\
&\quad - \frac{2}{\mu_\alpha}\|\nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{v}_{t+1}\|^2 - \frac{\mu_\alpha}{4}\|\alpha(\mathbf{x}_t, \mathbf{y}_t) - \alpha_t\|^2 - \frac{\mu_\alpha}{4}\|\alpha_t - \tilde{\alpha}_{t+1}\|^2 \\
&\quad + \frac{1}{\tau_\alpha}\|\alpha_t - \tilde{\alpha}_{t+1}\|^2 - \frac{L_f}{2}\|\tilde{\alpha}_{t+1} - \alpha_t\|^2 + \frac{\mu_\alpha}{2}\|\alpha(\mathbf{x}_t, \mathbf{y}_t) - \alpha_t\|^2,
\end{aligned}$$

where the equality follows from 5.
Hence we have

$$\begin{aligned}
\|\alpha_{t+1} - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 &\leq \left(1 - \frac{\mu_\alpha\eta_\alpha\tau_\alpha}{2}\right)\|\alpha_t - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \frac{4\eta_\alpha\tau_\alpha}{\mu_\alpha}\|\nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{v}_{t+1}\|^2 \\
&\quad + 2\eta_\alpha\tau_\alpha\left(\frac{\eta_\alpha}{2\tau_\alpha} + \frac{\mu_\alpha}{4} - \frac{1}{\tau_\alpha} + \frac{L_f}{2}\right)\|\alpha_t - \tilde{\alpha}_{t+1}\|^2 \\
&\leq \left(1 - \frac{\mu_\alpha\eta_\alpha\tau_\alpha}{2}\right)\|\alpha_t - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \frac{4\eta_\alpha\tau_\alpha}{\mu_\alpha}\|\nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{v}_{t+1}\|^2 \\
&\quad + 2\eta_\alpha\tau_\alpha\left(\frac{3L_f}{4} - \frac{1}{2\tau_\alpha}\right)\|\alpha_t - \tilde{\alpha}_{t+1}\|^2,
\end{aligned}$$

where we use $\eta_\alpha \leq 1/2$, $\mu_\alpha \leq L_f$ and $\tau_\alpha \leq 1/(3L_f)$.
As a result,

$$\begin{aligned}
\|\alpha_{t+1} - \alpha(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 &= \|\alpha_{t+1} - \alpha(\mathbf{x}_t, \mathbf{y}_t) + \alpha(\mathbf{x}_t, \mathbf{y}_t) - \alpha(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
&\leq \left(1 + \frac{\eta_\alpha \tau_\alpha \mu_\alpha}{4}\right) \|\alpha_{t+1} - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\quad + \left(1 + \frac{4}{\eta_\alpha \tau_\alpha \mu_\alpha}\right) \|\alpha(\mathbf{x}_t, \mathbf{y}_t) - \alpha(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \|^2 \\
&\leq \left(1 + \frac{\eta_\alpha \tau_\alpha \mu_\alpha}{4}\right) \|\alpha_{t+1} - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\quad + \left(1 + \frac{4}{\eta_\alpha \tau_\alpha \mu_\alpha}\right) L_\alpha^2 (\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2) \\
&\leq \left(1 - \frac{\mu_\alpha \eta_\alpha \tau_\alpha}{4}\right) \|\alpha_t - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \frac{8\eta_\alpha \tau_\alpha}{\mu_\alpha} \|\nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{v}_{t+1}\|^2 \\
&\quad + \left(1 + \frac{\eta_\alpha \tau_\alpha \mu_\alpha}{4}\right) 2\eta_\alpha \tau_\alpha \left(\frac{3L_f}{4} - \frac{1}{2\tau_\alpha}\right) \|\alpha_t - \tilde{\alpha}_{t+1}\|^2 \\
&\quad + \frac{8L_\alpha^2}{\eta_\alpha \tau_\alpha \mu_\alpha} (\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2) \\
&\leq \left(1 - \frac{\mu_\alpha \eta_\alpha \tau_\alpha}{4}\right) \|\alpha_t - \alpha(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \frac{8\eta_\alpha \tau_\alpha}{\mu_\alpha} \|\nabla_\alpha f(\mathbf{x}_t, \alpha_t, \mathbf{y}_t) - \mathbf{v}_{t+1}\|^2 \\
&\quad + \left(1 + \frac{\eta_\alpha \tau_\alpha \mu_\alpha}{4}\right) \frac{2\tau_\alpha}{\eta_\alpha} \left(\frac{3L_f}{4} - \frac{1}{2\tau_\alpha}\right) \|\alpha_t - \alpha_{t+1}\|^2 \\
&\quad + \frac{8L_\alpha^2}{\eta_\alpha \tau_\alpha \mu_\alpha} (\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2),
\end{aligned}$$

where we use $(1 + \epsilon/2)(1 - \epsilon) \leq (1 - \epsilon/2 - \epsilon^2) \leq 1 - \epsilon/2$ and Lemma 8. ■