Stochastic Regret Guarantees for Online Zeroth- and First-Order Bilevel Optimization

Anonymous Authors[1]

Abstract

Online bilevel optimization (OBO) is a powerful framework for machine learning problems where both outer and inner objectives evolve over time, requiring dynamic updates. Current OBO approaches rely on deterministic *window-smoothed* regret minimization, which may not accurately reflect system performance when functions change rapidly. In this work, we introduce a novel search direction and show that both first- and zeroth-order (ZO) stochastic OBO algorithms leveraging this direction achieve sublinear stochastic bilevel regret without window smoothing. Beyond these guarantees, our framework enhances efficiency by: (i) reducing oracle dependence in hypergradient estimation, (ii) updating inner and outer variables alongside the linear system solution, and (iii) employing ZO-based estimation of Hessians, Jacobians, and gradients. Experiments on online parametric loss tuning and black-box adversarial attacks validate our approach.

1. Introduction

Bilevel optimization (BO) minimizes an outer objective dependent on an inner problem's solution. Originating in game theory (Stackelberg, 1952) and formalized in mathematical optimization (Bracken & McGill, 1973), BO finds applications in operations research, engineering, economics (Dempe, 2002), and image processing (Crockett et al., 2022). Recently, BO has gained traction in machine learning, including hyperparameter optimization (Franceschi et al., 2018), meta-learning (Finn et al., 2017), reinforcement learning (Stadie et al., 2020), and neural architecture search (Liu et al., 2018a).

In the *offline setting*, BO solves the following problem:

$$\begin{aligned} &\mathbf{x}^* \in \text{argmin}_{\mathbf{x}\in\mathbb{R}^{d_1}} f(\mathbf{x}, \mathbf{y}^*(\mathbf{x})) \\ &\text{subj. to} \quad \mathbf{y}^*(\mathbf{x}) = \text{argmin}_{\mathbf{y}\in\mathbb{R}^{d_2}} g(\mathbf{x}, \mathbf{y}), \end{aligned} \tag{BO}$$

where f and g are the outer and inner objectives, and $\mathbf{x}$ and $\mathbf{y}$ are their respective optimization variables.

OBO (Tarzanagh et al., 2024) addresses dynamic scenarios where objectives evolve over time, requiring the agent to update the outer decision in response to the optimal inner decision. Similar to online single-level optimization (OSO) (Zinkevich, 2003), OBO involves iterative decision-making without prior knowledge of outcomes (Tarzanagh et al., 2024; Lin et al., 2024; Bohne et al., 2024). Let T be the total number of rounds. Define $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^{d_1}$ as the decision variable and $f_t : \mathcal{X} \times \mathbb{R}^{d_2} \to \mathbb{R}$ as the outer function. Similarly, define $\mathbf{y}_t \in \mathbb{R}^{d_2}$ and $g_t : \mathcal{X} \times \mathbb{R}^{d_2} \to \mathbb{R}$ for the inner problem, where $\mathbf{y}_t^*(\mathbf{x}) = \text{argmin}_{\mathbf{y}\in\mathbb{R}^{d_2}} g_t(\mathbf{x}, \mathbf{y})$. OBO can be seen as a *single-player* problem, where the player selects $\mathbf{x}_t$ without knowing $\mathbf{y}_t^*(\mathbf{x})$, using $\mathbf{y}_t$ as an estimate based on g_t. Alternatively, it can be framed as a *two-player* game (Stackelberg, 1952), where the leader ($\mathbf{x}_t$) competes with the follower ($\mathbf{y}_t$), who selects $\mathbf{y}_t^*(\mathbf{x})$ based on limited knowledge of g_t; see Section 2. This framework includes online and adversarial variants of (BO), such as online actor-critic algorithms (Zhou et al., 2020), online meta-learning (Finn et al., 2019), and online hyperparameter optimization (Lin et al., 2024). The inner and outer functions may be time-varying, adversarial, unavailable *a priori*, and require *nonstationary* optimization.

1.1. Our Contributions

This paper addresses stochastic OBO, introducing novel first[1]- and zeroth-order methods to minimize stochastic bilevel regret. Key contributions are summarized below.

• **Stochastic regret minimization without window-smoothing.** Existing OBO methods (Tarzanagh et al., 2024; Lin et al., 2024; Huang et al., 2023; Bohne et al., 2024) rely on deterministic *window-smoothed* regret minimization, which may not accurately reflect system performance when functions change rapidly. We address these limitations by introducing a novel search direction (Section 3) and proving that both first-order and ZO methods achieve sublinear *stochastic bilevel regret without window-smoothing* ($w = 1$); see Theorems 3.6 and 4.2 and Table 1.

• **OBO with function value oracle feedback.** In large-scale

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]First-order refers to the setting where only partial gradients of the leader objective f_t are accessible, while second-order information is still required for the follower objective g_t; refer to Section 3.

OBO Method	Window Size in Regret (w)	System Iters.	Stochastic Regret	Const. Regret Min.	Only Func. Feedback	Local Regret Bound
OAGD	$o(T)$	N.A. (Exact)	✗	✗	✗	$\frac{T}{w} + H_{1,T} + H_{2,T}$
SOBOW	$o(T)$	$\mathcal{O}(\kappa_g \log \kappa_g)$	✗	✗	✗	$\frac{T}{w} + V_T + H_{2,T}$
SOBBO	$o(T)$	$\mathcal{O}(\kappa_g \log \kappa_g)$	✓	✓	✗	$\frac{T}{w}\sigma^2 + V_T + H_{2,T}$
SOGD	1	1	✓	✓	✗	$T^{\frac{1}{3}}(\sigma^2 + \Delta_T) + T^{\frac{2}{3}}\Psi_T$
ZO-SOGD	1	1	✓	✓	✓	$(d_1+d_2)^{\frac{3}{4}} T^{\frac{1}{3}}(\hat{\sigma}^2 + \hat{\Delta}_T) + (d_1+d_2)^{\frac{3}{2}} T^{\frac{2}{3}}\hat{\Psi}_T$

Table 1. Comparison of OBO algorithms based on regret window size (w), system solver iterations, stochastic regret, constrained regret minimization, function feedback settings, and local regret bounds. Here, κ_g denotes the condition number of g_t, while V_T, $H_{p,T}$, Δ_T, Ψ_T, $\hat{\Delta}_T$, and $\hat{\Psi}_T$ are defined in (10), (13), and (25), respectively. The compared algorithms include OAGD (Tarzanagh et al., 2024), SOBOW (Lin et al., 2024), and SOBBO (Bohne et al., 2024).

and black-box settings (Chen et al., 2017; Nesterov, 2005), first- and second-order information is often unavailable or costly. Constructing accurate (hyper)-gradient estimators using only function value oracles is particularly challenging due to BO's nested structure. Existing methods rely on gradient, Hessian, and Jacobian oracles, limiting scalability (Franceschi et al., 2017; Ghadimi & Wang, 2018). We propose Algorithm 2, which estimates Hessians, Jacobians, and gradients using function value oracles, achieving sublinear local regret (Theorem 4.2).

• **OBO with one subproblem solver iteration.** A major challenge in BO is solving implicit systems to approximate the hypergradient (Ji et al., 2021; Chen et al., 2021). While efficient offline BO methods exist (Ji et al., 2021; Dagréou et al., 2022), extending them to OBO is difficult due to time-varying objectives. SOBOW (Lin et al., 2024) partially addresses this using a conjugate gradient (CG) algorithm with increasing iterations (Table 1). We improve upon SOBOW by introducing Algorithms 1 and 2, which require only a *single* subproblem solver iteration.

2. Preliminaries

Notation. $\mathbb{R}^d$ denotes the d-dimensional real space, with $\mathbb{R}^d_+$ and $\mathbb{R}^d_{++}$ as its positive and negative orthants. Vectors are bold lower-case letters (e.g., $\mathbf{x}, \mathbf{y}$), with $\langle \mathbf{x}, \mathbf{y}\rangle$ for inner product and $\|\cdot\|$ for Euclidean norm. A gradient is $\nabla_{\mathbf{x}}$, with $\nabla^2_{\mathbf{xy}} = \nabla_{\mathbf{x}}\nabla_{\mathbf{y}}$. A function is L-smooth if its gradient is L-Lipschitz. The Euclidean projection onto a convex set $\mathcal{X}$ is $\Pi_{\mathcal{X}}(\mathbf{z}) = \text{argmin}_{\mathbf{x}\in\mathcal{X}}(1/2)\|\mathbf{x}-\mathbf{z}\|^2$. The set $\{1,\ldots,T\}$ is denoted by $[T]$, and $\mathbb{E}[\cdot]$ represents expectation. Lastly, $\mathcal{O}(\cdot)$ hides problem-independent constants.

Stochastic OBO Setting. Let T be the total rounds (Tarzanagh et al., 2024). Define $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^{d_1}$ as the decision variable and $f_t : \mathcal{X} \times \mathbb{R}^{d_2}$ as the outer objective. The inner decision variable and objective are $\mathbf{y}_t \in \mathbb{R}^{d_2}$ and $g_t : \mathcal{X} \times \mathbb{R}^{d_2}$, where the optimal inner decision is:

$$\mathbf{y}^*_t(\mathbf{x}) \in \underset{\mathbf{y}\in\mathbb{R}^{d_2}}{\text{argmin}} \left\{ g_t(\mathbf{x},\mathbf{y}) := \underset{\zeta_t\sim\mathcal{D}_g}{\mathbb{E}} \left[g_t(\mathbf{x},\mathbf{y};\zeta_t)\right] \right\}. \quad (1)$$

Further, we have

$$f_t(\mathbf{x}, \mathbf{y}^*_t(\mathbf{x})) := \mathbb{E}_{\xi_t\sim\mathcal{D}_f}\left[f_t(\mathbf{x},\mathbf{y}^*_t(\mathbf{x});\xi_t)\right].$$

Here, $(\mathcal{D}_f, \mathcal{D}_g)$ are data distributions. Note that our setting is stochastic, and only noisy evaluations of the function, gradient, and Hessian are accessible.

Unlike OSO, where true losses are revealed immediately, in OBO, the outer function $f_t(\mathbf{x}, \mathbf{y}^*_t(\mathbf{x}))$ is unavailable for updating $\mathbf{x}_t$. Moreover, $f_t(\mathbf{x}, \mathbf{y}^*_t(\mathbf{x}))$ is typically non-convex in $\mathbf{x}$, making standard regret definitions from online convex optimization (Hazan, 2016b) inapplicable.

Given a sequence $\{\alpha_t \in \mathbb{R}_{++}\}_{t=1}^T$, we define the following notion of *bilevel local regret*:

$$\text{BL-Reg}_T := \sum_{t=1}^T \mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}^*_t(\mathbf{x}_t))\right)\right\|^2\right], \quad (2a)$$

with

$$\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}^*_t(\mathbf{x}_t))\right) = \frac{1}{\alpha_t}\Big(\mathbf{x}_t - \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \alpha_t \nabla f_t(\mathbf{x}_t, \mathbf{y}^*_t(\mathbf{x}_t))\right]\Big). \quad (2b)$$

The local regret (2) compares the leader's decision $\mathbf{x}_t$ to the stationary points $\mathbf{x}^*_t$ satisfying $\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}^*_t; \nabla f_t(\mathbf{x}^*_t, \mathbf{y}^*_t(\mathbf{x}^*_t))\right) = 0$. This can also be viewed as dynamic local regret, as the baseline corresponds to a stationary point of the leader's objective f_t.

Previous work on (nonconvex) OBO examined unconstrained local regret using window-smoothed objectives: $F_{t,w}(\mathbf{x},\mathbf{y}) = (1/w)\sum_{i=0}^{w-1} f_{t-i}(\mathbf{x},\mathbf{y})$. For $w = 1$ and $\mathcal{X} = \mathbb{R}^{d_1}$, this reduces to (2). Tarzanagh et al. (2024); Lin et al. (2024) showed that $w = o(T)$ ensures sublinear regret under slow variations in $\{F_{t,w}\}_{t=1}^T$, while rapid changes can lead to deviations. However, smoothing may misrepresent regret (Figure 1). This paper introduces a new projection-based local regret notion (2) without smoothing, and establishes sublinear regret for constrained OBO.

Online Gradient Descent (OGD). One of the most widely used algorithms for online (single-level) optimization is

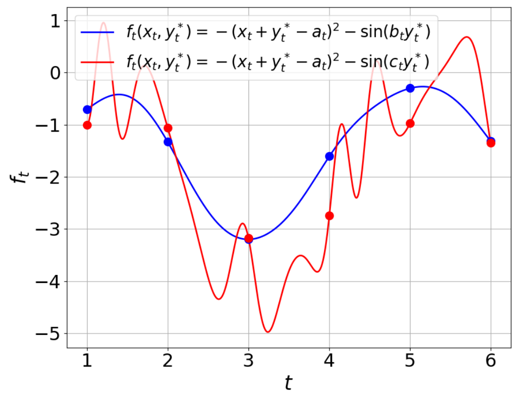

Figure 1. Smoothly and rapidly changing f_t in OBO with $g_t(x_t, y_t) = (y_t - \cos(x_t))^2$, $a_t = 1 + 0.5\sin(t)$, $b_t = 1 + \sin(0.5t)$, and $c_t = 10b_t$.

OGD (Zinkevich, 2003). The procedure for OGD is as follows: For each $t \in [T]$, the algorithm selects $\mathbf{x}_t \in \mathcal{X}$, observes the function $f_t : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$, and updates according to

$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\big(\mathbf{x}_t - \alpha_t \nabla f_t(\mathbf{x}_t)\big), \quad \alpha_t > 0. \tag{OGD}$$

In the following, we adapt OGD to OBO and introduce a novel framework that requires limited feedback and can utilize ZO updates within a single-loop structure.

3. Stochastic OBO with Access to First and Second Order Oracles

To adapt OGD to OBO, Tarzanagh et al. (2024); Lin et al. (2024); Bohne et al. (2024) developed a variant alternating between inner and outer OGD, achieving sublinear bilevel regret bounds. We introduce a new search direction that enables sublinear bilevel regret without window smoothing.

To compute the hypergradient $\nabla f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))$ where $\mathbf{y}_t^*(\mathbf{x})$ is defined in (1), since $\nabla_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) = 0$, using the implicit function theorem, yields

$$\begin{aligned}\nabla f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) &= \nabla_{\mathbf{x}} f_t\left(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})\right) \\ &\quad + \nabla \mathbf{y}_t^*(\mathbf{x}) \nabla_{\mathbf{y}} f_t\left(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})\right),\end{aligned} \tag{3}$$

where $\nabla \mathbf{y}_t^*(\mathbf{x}) \nabla^2_{\mathbf{y}} g_t\left(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})\right) + \nabla^2_{\mathbf{xy}} g_t\left(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})\right) = 0$.

As the exact $\mathbf{y}_t^*(\mathbf{x})$ is not available, we estimate the hypergradient of f_t at $(\mathbf{x}, \mathbf{y})$ by

$$\tilde{\nabla} f_t(\mathbf{x}, \mathbf{y}) := \nabla_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}) + \nabla^2_{\mathbf{xy}} g_t\left(\mathbf{x}, \mathbf{y}\right) \mathbf{v}_t^*(\mathbf{x}), \tag{4a}$$

where

$$\nabla^2_{\mathbf{y}} g_t\left(\mathbf{x}, \mathbf{y}\right) \mathbf{v}_t^*(\mathbf{x}) + \nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}) = 0. \tag{4b}$$

An accurate solution of (4b) is crucial for tight regret bounds. Tarzanagh et al. (2024) assumes an exact solution, which is restrictive in large-scale settings. To address this, Lin et al. (2024) proposed an efficient OBO algorithm with window averaging, using CG methods to solve (4b), which

Algorithm 1 SOGD

Require: $(\mathbf{x}_1, \mathbf{y}_1, \mathbf{v}_1) \in \mathcal{X} \times \mathbb{R}^{d_2} \times \mathbb{R}^{d_2}$; $T \in \mathbb{N}$; $p \in \mathbb{R}_{++}$; stepsizes $\{(\alpha_t, \beta_t, \delta_t) \in \mathbb{R}^3_{++}\}_{t=1}^T$; parameters $\{(\gamma_t, \lambda_t, \eta_t)\}_{t=1}^T \in (0, 1)$; $\mathbf{z}_t := (\mathbf{x}_t, \mathbf{y}_t)$.

For $t = 1$ **to** T **do**:

S1. Draw samples $\mathcal{B}_t$ and $\bar{\mathcal{B}}_t$ with batch sizes b and $\bar{b}$. Get search directions $\mathbf{d}_t^{\mathbf{y}}$, $\mathbf{d}_t^{\mathbf{v}}$, and $\mathbf{d}_t^{\mathbf{x}}$:

$$\begin{aligned}&\mathbf{d}_t^{\mathbf{yy}}(\mathbf{z}_t; \bar{\mathcal{B}}_t) = \nabla_{\mathbf{y}} g_t(\mathbf{z}_t; \bar{\mathcal{B}}_t), \\ &\mathbf{d}_t^{\mathbf{y}} = \mathbf{d}_t^{\mathbf{yy}}(\mathbf{z}_t; \bar{\mathcal{B}}_t) + (1 - \gamma_t)(\mathbf{d}_{t-1}^{\mathbf{y}} - \mathbf{d}_t^{\mathbf{yy}}(\mathbf{z}_{t-1}; \bar{\mathcal{B}}_t)),\end{aligned} \tag{7a}$$

$$\begin{aligned}&\mathbf{d}_t^{\mathbf{vv}}(\mathbf{z}_t; \mathcal{B}_t) = \nabla_{\mathbf{y}} f_t(\mathbf{z}_t; \mathcal{B}_t) + \nabla^2_{\mathbf{y}} g_t\left(\mathbf{z}_t; \bar{\mathcal{B}}_t\right) \mathbf{v}_t, \\ &\mathbf{d}_t^{\mathbf{v}} = \mathbf{d}_t^{\mathbf{vv}}(\mathbf{z}_t; \mathcal{B}_t) + (1 - \lambda_t)(\mathbf{d}_{t-1}^{\mathbf{v}} - \mathbf{d}_t^{\mathbf{vv}}(\mathbf{z}_{t-1}; \mathcal{B}_t)),\end{aligned} \tag{7b}$$

$$\begin{aligned}&\mathbf{d}_t^{\mathbf{xx}}(\mathbf{z}_t; \mathcal{B}_t) = \nabla_{\mathbf{x}} f_t(\mathbf{z}_t; \mathcal{B}_t) + \nabla^2_{\mathbf{xy}} g_t\left(\mathbf{z}_t; \bar{\mathcal{B}}_t\right) \mathbf{v}_t, \\ &\mathbf{d}_t^{\mathbf{x}} = \mathbf{d}_t^{\mathbf{xx}}(\mathbf{z}_t; \mathcal{B}_t) + (1 - \eta_t)(\mathbf{d}_{t-1}^{\mathbf{x}} - \mathbf{d}_t^{\mathbf{xx}}(\mathbf{z}_{t-1}; \mathcal{B}_t)).\end{aligned} \tag{7c}$$

S2. Update inner, system, and outer solutions:

$$\begin{aligned}\mathbf{y}_{t+1} = \mathbf{y}_t - \beta_t \mathbf{d}_t^{\mathbf{y}}, \;\; &\mathbf{v}_{t+1} = \Pi_{\mathcal{Z}_p}\left[\mathbf{v}_t - \delta_t \mathbf{d}_t^{\mathbf{v}}\right], \\ &\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}[\mathbf{x}_t - \alpha_t \mathbf{d}_t^{\mathbf{x}}].\end{aligned}$$

is equivalent to:

$$\min_{\mathbf{v}_t \in \mathbb{R}^{d_2}} \; (1/2)\|\nabla^2_{\mathbf{y}} g_t\left(\mathbf{x}, \mathbf{y}\right) \mathbf{v}_t + \nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y})\|^2. \tag{5}$$

New Search Direction for OBO. Next, we introduce a novel search direction that enables both first- and ZO stochastic OBO algorithms to achieve sublinear bilevel regret without smoothing. We first state the following lemma:

Lemma 3.1. *Let $w = t$ and $W = 1/\eta$ in the window-smoothed gradient $\hat{\nabla} F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) = (1/W) \sum_{i=0}^{w-1} \nu^i \hat{\nabla} f_{t-i}(\mathbf{x}_{t-i}, \mathbf{y}_{t-i}; \mathcal{B}_{t-i})$, where $\mathcal{B}_t := \{\xi_1, \ldots, \xi_b\}$ is drawn i.i.d. from $\mathcal{D}_f$. Then,*

$$\hat{\nabla} F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) = \sum_{j=1}^{t} \eta(1-\eta)^{t-j} \hat{\nabla} f_j(\mathbf{x}_j, \mathbf{y}_j; \mathcal{B}_j).$$

Furthermore, we have $\hat{\nabla} F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) = \hat{\mathbf{d}}_t^{\mathbf{x}}$ with $\hat{\mathbf{d}}_t^{\mathbf{x}} = \eta \hat{\nabla} f_t(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) + (1 - \eta)\hat{\mathbf{d}}_{t-1}^{\mathbf{x}}$, and $\hat{\mathbf{d}}_1 = (1/W)\hat{\nabla} f_1(\mathbf{x}_1, \mathbf{y}_1; \mathcal{B}_1)$ for all $t \geq 2$.

As shown in Lemma 3.1, for a specific choice of w and W, the time-smoothed gradient forms a recursive momentum-type search direction. However, achieving sublinear regret in stochastic OBO requires large-window smoothing ($w = o(T)$). To address this, we propose the following search direction:

$$\mathbf{d}_t^{\mathbf{x}} = \eta \nabla f_t(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) + (1 - \eta)\mathbf{d}_{t-1}^{\mathbf{x}} \tag{6a}$$

$$+ (1 - \eta)(\nabla f_t(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) - \nabla f_t(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \mathcal{B}_t)). \tag{6b}$$

This direction is used for updating $\mathbf{x}$, with similar updates for $\mathbf{y}$ and $\mathbf{v}$, as discussed below and detailed in Algorithm 1.

The quadratic optimization formulation of (4b) in (5) leads to single-loop frameworks such as Dagréou et al. (2022). Inspired by this, we present Simultaneous Online Gradient Descent (SOGD) for constrained OBO, outlined in Algorithm 1. SOGD evolves the follower's decision (inner) variable, the linear system solution, and the leader's decision (outer) variable simultaneously at each step for given batches $\mathcal{B} := \{\xi_1, \ldots, \xi_b\}$ and $\bar{\mathcal{B}} := \{\zeta_1, \ldots, \zeta_{\bar{b}}\}$, which are drawn i.i.d. from unknown distributions $\mathcal{D}_f$ and $\mathcal{D}_g$ with batch sizes b and $\bar{b}$. Computing directions in **S1.** of Algorithm 1 does not require $\nabla^2_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)$ and $\nabla^2_{\mathbf{xy}} g_t(\mathbf{x}_t, \mathbf{y}_t)$, only their product with a vector, at the same cost as computing a gradient. Technically, it utilizes an auxiliary variable $\mathbf{v}_t$ and concurrently updates $\mathbf{y}_t$, $\mathbf{v}_t$, and $\mathbf{x}_t$ at each local iteration t. Moreover, **S2.** of Algorithm 1 introduces an auxiliary projection $\Pi_{\mathcal{Z}_p}$ on the ball $\mathcal{Z}_p$ defined as follows:

$$\Pi_{\mathcal{Z}_p}(\mathbf{v}) := \min\left\{1, \frac{p}{\|\mathbf{v}\|}\right\}\mathbf{v}, \tag{8}$$

where $\mathcal{Z}_p := \{\mathbf{v} \in \mathbb{R}^{d_2} \mid \|\mathbf{v}\| \le p\}$.

Unlike OAGD (Tarzanagh et al., 2024), which updates $\mathbf{x}$ and $\mathbf{y}$ in separate loops, SOGD updates both simultaneously. Compared to SOBOW (Lin et al., 2024), which uses multiple CG updates, our method employs a single OGD to update the inner solution, linear system, and outer variable.

Assumption 3.2. $g_t(\mathbf{x}, \mathbf{y})$ is twice continuously differentiable and μ_g-strongly convex in $\mathbf{y}$ for all $\mathbf{x} \in \mathcal{X}, t \in [T]$.

Assumption 3.3. Let $\mathbf{z} = [\mathbf{x}; \mathbf{y}]$ and $\mathbf{z}' = [\mathbf{x}'; \mathbf{y}']$, where $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $\mathbf{y}, \mathbf{y}' \in \mathbb{R}^{d_2}$. For any $\mathbf{z}, \mathbf{z}'$, and $t \in [T]$:

B1. $\exists\, \ell_{f,0} \in \mathbb{R}_+$ s.t. $\|f_t(\mathbf{z}; \xi) - f_t(\mathbf{z}'; \xi)\| \le \ell_{f,0}\|\mathbf{z} - \mathbf{z}'\|$;
B2. $\exists\, \ell_{f,1} \in \mathbb{R}_+$ s.t. $\|\nabla f_t(\mathbf{z}; \xi) - \nabla f_t(\mathbf{z}'; \xi)\| \le \ell_{f,1}\|\mathbf{z} - \mathbf{z}'\|$;
B3. $\exists\, \ell_{g,1} \in \mathbb{R}_+$ s.t. $\|\nabla g_t(\mathbf{z}; \zeta) - \nabla g_t(\mathbf{z}'; \zeta)\| \le \ell_{g,1}\|\mathbf{z} - \mathbf{z}'\|$;
B4. $\exists\, \ell_{g,2} \in \mathbb{R}_+$ s.t. $\|\nabla^2 g_t(\mathbf{z}; \zeta) - \nabla^2 g_t(\mathbf{z}'; \zeta)\| \le \ell_{g,2}\|\mathbf{z} - \mathbf{z}'\|$.

Assumption 3.4. For any $t \in [T]$, $|f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))| \le M$ for some finite constant $M \in \mathbb{R}_{++}$ and any $\mathbf{x} \in \mathcal{X}$.

Assumption 3.5. There exist constants $\sigma_{g_\mathbf{y}}, \sigma_{g_{\mathbf{yy}}}, \sigma_{g_{\mathbf{xy}}}, \sigma_{f_\mathbf{y}}, \sigma_{f_\mathbf{y}}$ such that, for all $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$:

C1. $\mathbb{E}\|\nabla_\mathbf{y} g_t(\mathbf{z}; \zeta) - \nabla_\mathbf{y} g_t(\mathbf{z})\|^2 \le \sigma^2_{g_\mathbf{y}}$,
C2. $\mathbb{E}\|\nabla^2_\mathbf{y} g_t(\mathbf{z}; \zeta) - \nabla^2_\mathbf{y} g_t(\mathbf{z})\|^2 \le \sigma^2_{g_{\mathbf{yy}}}$,
C3. $\mathbb{E}\|\nabla^2_{\mathbf{xy}} g_t(\mathbf{z}; \zeta) - \nabla^2_{\mathbf{xy}} g_t(\mathbf{z})\|^2 \le \sigma^2_{g_{\mathbf{xy}}}$,
C4. $\mathbb{E}\|\nabla_\mathbf{y} f_t(\mathbf{z}; \xi) - \nabla_\mathbf{y} f_t(\mathbf{z})\|^2 \le \sigma^2_{f_\mathbf{y}}$,
C5. $\mathbb{E}\|\nabla_\mathbf{x} f_t(\mathbf{z}; \xi) - \nabla_\mathbf{x} f_t(\mathbf{z})\|^2 \le \sigma^2_{f_\mathbf{x}}$.

Throughout this paper, we define

$$\sigma^2 := \sigma^2_{g_\mathbf{y}} + \sigma^2_{g_{\mathbf{yy}}} + \sigma^2_{f_\mathbf{y}} + \sigma^2_{g_{\mathbf{xy}}} + \sigma^2_{f_\mathbf{x}}. \tag{9}$$

Assumptions 3.2 and 3.3 are widely used in both BO (Chen et al., 2021; Ji et al., 2021) and OBO (Tarzanagh et al., 2024), and many bilevel machine learning problems satisfy it (Franceschi et al., 2018). Further, Assumption 3.4 is widely used in the study of non-convex online optimization (Hazan et al., 2017; Lin et al., 2024). Assumption 3.5 assumes that we have access to an unbiased stochastic gradient, Hessian and Jacobian with bounded variance, which is standard in the literature (Chen et al., 2021).

Achieving sublinear dynamic regret is generally impossible due to arbitrary fluctuations in time-varying functions (Besbes et al., 2015). Existing analyses (Tarzanagh et al., 2024; Lin et al., 2024) bound regret by imposing regularity constraints on the comparator sequence. To achieve sublinear regret, we introduce the following regularities:

• **Path-length (of order p)** and **function variation**: Tarzanagh et al. (2024) defines the following metrics for bilevel sequences:

$$\begin{aligned} H_{p,T} &:= \sum_{t=2}^{T} \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{y}^*_{t-1}(\mathbf{x}) - \mathbf{y}^*_t(\mathbf{x})\|^p, \\ V_T &:= \sum_{t=2}^{T} \sup_{\mathbf{x}\in\mathcal{X}} |f_{t-1}(\mathbf{x}, \mathbf{y}^*_{t-1}(\mathbf{x})) - f_t(\mathbf{x}, \mathbf{y}^*_t(\mathbf{x}))|. \end{aligned} \tag{10}$$

Path-length $H_{p,T}$ measures changes in the follower's costs, while V_T captures the smoothness of the leader's objective. We use path-length for the follower and function variation for the leader, as the follower's objective is strongly convex (see Assumption 3.2), while the leader's is nonconvex.

• **Inner and Outer Gradient Variations**: Another regularity is the sequential difference between the individual gradients of the upper-level loss function:

$$\begin{aligned} D_{\mathbf{x},T} &:= \sum_{t=2}^{T} \sup_{\mathbf{x},\mathbf{y}} \|\nabla_\mathbf{x} f_{t-1}(\mathbf{x}, \mathbf{y}) - \nabla_\mathbf{x} f_t(\mathbf{x}, \mathbf{y})\|^2, \\ D_{\mathbf{y},T} &:= \sum_{t=2}^{T} \sup_{\mathbf{x},\mathbf{y}} \|\nabla_\mathbf{y} f_{t-1}(\mathbf{x}, \mathbf{y}) - \nabla_\mathbf{y} f_t(\mathbf{x}, \mathbf{y})\|^2. \end{aligned} \tag{11}$$

As in Huang et al.; Hallak et al. (2021), $D_{\mathbf{x},T}$ and $D_{\mathbf{y},T}$ measure the gradient drift of f_t relative to f_{t-1} for $\mathbf{x}$ and $\mathbf{y}$, respectively. We further define deviations in the gradient, Hessian, and Jacobian of the lower-level objective as:

$$\begin{aligned} G_{\mathbf{y},T} &:= \sum_{t=2}^{T} \|\nabla_\mathbf{y} g_{t-1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_\mathbf{y} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2, \\ G_{\mathbf{yy},T} &:= \sum_{t=2}^{T} \|\nabla^2_\mathbf{y} g_{t-1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla^2_\mathbf{y} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2, \\ G_{\mathbf{xy},T} &:= \sum_{t=2}^{T} \|\nabla^2_{\mathbf{xy}} g_{t-1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla^2_{\mathbf{xy}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2. \end{aligned} \tag{12}$$

We introduce the following notations for simplicity:

$$\Delta_T := E_1 + V_T, \quad \Psi_T := H_{2,T} + G_T + D_T, \tag{13}$$

where $(V_T, H_{p,T})$ are defined in (10), and

$$\begin{aligned} E_1 &:= \|\mathbf{y}_1 - \mathbf{y}^*_1(\mathbf{x}_1)\|^2 + \|\mathbf{v}_1 - \mathbf{v}^*_1(\mathbf{x}_1)\|^2, \\ G_T &:= G_{\mathbf{y},T} + G_{\mathbf{yy},T} + G_{\mathbf{xy},T}, \\ D_T &:= D_{\mathbf{x},T} + D_{\mathbf{y},T}. \end{aligned} \tag{14}$$

By accounting for both D_T and G_T, we can represent the variations in the environments of OBO.

Theorem 3.6. *Let $\{(f_t, g_t)\}_{t=1}^T$ be the sequence of functions presented to Algorithm 1, satisfying Assumptions 3.2-3.5. For all $t \in [T]$, let*

$$\alpha_t = \frac{1}{(c+t)^{1/3}}, \quad \beta_t = c_\beta \alpha_t, \quad \delta_t = c_\delta \alpha_t, \quad b = \bar{b} = 1,$$
$$\gamma_{t+1} = c_\gamma \alpha_t^2, \quad \eta_{t+1} = c_\eta \alpha_t^2, \quad \lambda_{t+1} = c_\lambda \alpha_t^2. \tag{15}$$

Here, c, c_β, c_δ, c_γ, c_η, and c_λ are specified in (104). Algorithm 1 guarantees:

$$\text{BL-Reg}_T \leq \mathcal{O}\left(T^{1/3}(\sigma^2 + \Delta_T) + T^{2/3}\Psi_T\right), \tag{16}$$

where σ and (Δ_T, Ψ_T) are defined in (9) and (13).

Theorem 3.6 bounds the regret of Algorithm 1 without window-smoothing, based on the regularities in (14). We note that the average dynamic regret $\text{BL-Reg}_T/T \leq \mathcal{O}(T^{-2/3}(\sigma^2 + \Delta_T) + T^{-1/3}\Psi_T)$ remains sublinear under suitable conditions on Δ_T and Ψ_T.

Remark 3.7 (**Stochastic Regret Guarantee for OBO and OSO with $w = 1$**). The additional terms in (6b) improve the average regret dependence on variance, achieving a $T^{-2/3}\sigma^2$ bound, better than the $T^{-1/2}\sigma^2$ bound for stochastic OBO (Bohne et al., 2024). This also provides the first regret bound without window-smoothing, unlike (Bohne et al., 2024; Tarzanagh et al., 2024; Lin et al., 2024; Huang et al., 2023). For OSO, our approach improves the $T^{-1/2}\sigma^2$ dependence from (Hallak et al., 2021).

4. OBO with Zeroth Order Oracles

Black-box optimization arises in machine learning when explicit gradients are unavailable (Chen et al., 2017). We study ZO-type OBO algorithms with limited access to the leader's and follower's objective values. Let $\mathbf{s} \in \mathbb{R}^{d_1}$ and $\mathbf{r} \in \mathbb{R}^{d_2}$ be vectors uniformly generated from the unit balls B_1 and B_2, respectively. Given positive smoothing parameters $\boldsymbol{\rho} = (\rho_\mathbf{s}, \rho_\mathbf{r})$, we use the Gaussian smoothing function (Nesterov & Spokoiny, 2017) to define the OBO objectives:

$$f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) = \mathop{\mathbb{E}}_{(\mathbf{s},\mathbf{r})}\left[f_t(\mathbf{x} + \rho_\mathbf{s}\mathbf{s}, \hat{\mathbf{y}}_t^*(\mathbf{x}) + \rho_\mathbf{r}\mathbf{r}; \xi)\right], \tag{17}$$

where

$$\hat{\mathbf{y}}_t^*(\mathbf{x}) \in \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^{d_2}} \Big\{ g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) := \mathop{\mathbb{E}}_{(\mathbf{s},\mathbf{r})}\left[g_t(\mathbf{x} + \rho_\mathbf{s}\mathbf{s}, \mathbf{y} + \rho_\mathbf{r}\mathbf{r}; \zeta)\right] \Big\}. \tag{18}$$

Using (17), we provide methodology to approximate each term in (7) using ZO oracles. Specifically, following Shamir (2017), we estimate the gradient of a function $h : \mathbb{R}^d \to \mathbb{R}$, querying at $\mathbf{x} - \lambda\mathbf{s}$ and $\mathbf{x} + \lambda\mathbf{s}$, yielding an estimator $(d/2\lambda)\left(h(\mathbf{x} + \lambda\mathbf{s}) - h(\mathbf{x} - \lambda\mathbf{s})\right)\mathbf{s}$. Using this strategy, the finite-difference estimation of $\nabla g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})$, denoted as $\hat{\nabla} g_t(\mathbf{x}, \mathbf{y})$, is constructed for given smoothing parameters $\boldsymbol{\rho} = (\rho_\mathbf{s}, \rho_\mathbf{r})$, and batches $\mathcal{B} := \{\xi_1, \ldots, \xi_b\}$ and $\bar{\mathcal{B}} := \{\zeta_1, \ldots, \zeta_{\bar{b}}\}$, drawn i.i.d. from $\mathcal{D}_f$ and $\mathcal{D}_g$, as:

$$\hat{\nabla}_\mathbf{y} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathcal{B}}) := \frac{d_2}{2\bar{b}\rho_\mathbf{r}} \sum_{i=1}^{\bar{b}} \left(g_t(\mathbf{x}, \mathbf{y} + \rho_\mathbf{r}\mathbf{r}_i; \zeta_i) - g_t(\mathbf{x}, \mathbf{y} - \rho_\mathbf{r}\mathbf{r}_i; \zeta_i)\right)\mathbf{r}_i, \tag{19a}$$

$$\hat{\nabla}_\mathbf{x} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathcal{B}}) := \frac{d_1}{2\bar{b}\rho_\mathbf{s}} \sum_{i=1}^{\bar{b}} (g_t(\mathbf{x} + \rho_\mathbf{s}\mathbf{s}_i, \mathbf{y}; \zeta_i) - g_t(\mathbf{x} - \rho_\mathbf{s}\mathbf{s}_i, \mathbf{y}; \zeta_i))\mathbf{s}_i. \tag{19b}$$

Similarly, we estimate $\nabla_\mathbf{y} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}; \mathcal{B})$ and $\nabla_\mathbf{x} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}; \mathcal{B})$, respectively, by

$$\hat{\nabla}_\mathbf{y} f_t(\mathbf{x}, \mathbf{y}; \mathcal{B}) := \frac{d_2}{2b\rho_\mathbf{r}} \sum_{i=1}^{b} (f_t(\mathbf{x}, \mathbf{y} + \rho_\mathbf{r}\mathbf{r}_i; \xi_i) - f_t(\mathbf{x}, \mathbf{y} - \rho_\mathbf{r}\mathbf{r}_i; \xi_i))\mathbf{r}_i, \tag{20a}$$

$$\hat{\nabla}_\mathbf{x} f_t(\mathbf{x}, \mathbf{y}; \mathcal{B}) := \frac{d_1}{2b\rho_\mathbf{s}} \sum_{i=1}^{b} (f_t(\mathbf{x} + \rho_\mathbf{s}\mathbf{s}_i, \mathbf{y}; \xi_i) - f_t(\mathbf{x} - \rho_\mathbf{s}\mathbf{s}_i, \mathbf{y}; \xi_i))\mathbf{s}_i. \tag{20b}$$

Further, given a smoothing parameter $\rho_\mathbf{v} > 0$, we can approximate the Hessian-vector product $\nabla_\mathbf{y}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})\mathbf{v}$ and the Jacobian-vector product $\nabla_{\mathbf{xy}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})\mathbf{v}$ as the finite difference between two gradients, respectively, as

$$\hat{\nabla}_\mathbf{y}^2 g_t(\mathbf{x}, \mathbf{y}; \bar{\mathcal{B}}) := \frac{1}{2\bar{b}\rho_\mathbf{v}} \sum_{i=1}^{\bar{b}} (\hat{\nabla}_\mathbf{y} g_t(\mathbf{x}, \mathbf{y} + \rho_\mathbf{v}\mathbf{v}; \zeta_i) - \hat{\nabla}_\mathbf{y} g_t(\mathbf{x}, \mathbf{y} - \rho_\mathbf{v}\mathbf{v}; \zeta_i)), \tag{21a}$$

$$\hat{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}, \mathbf{y}; \bar{\mathcal{B}}) := \frac{1}{2\bar{b}\rho_\mathbf{v}} \sum_{i=1}^{\bar{b}} (\hat{\nabla}_\mathbf{x} g_t(\mathbf{x}, \mathbf{y} + \rho_\mathbf{v}\mathbf{v}; \zeta_i) - \hat{\nabla}_\mathbf{x} g_t(\mathbf{x}, \mathbf{y} - \rho_\mathbf{v}\mathbf{v}; \zeta_i)). \tag{21b}$$

Using (19)–(21), the first-order terms in (7) are approximated as $\hat{\mathbf{d}}_t^\mathbf{y}$, $\hat{\mathbf{d}}_t^\mathbf{v}$, and $\hat{\mathbf{d}}_t^\mathbf{x}$ in (22). The approximations in (21a) and (21b) introduce errors in the hypergradient, which must be controlled. (21) depends on the dimension of $\mathbf{y}$, as in ZO optimization (Nesterov & Spokoiny, 2017; Shamir, 2017). The projection $\Pi_{\mathcal{Z}_p}$ in (8) bounds $\mathbf{v}$, controlling variance in $\mathbf{v}$ and $\mathbf{x}$ updates for convergence.

Assumption 4.1. There exist constants $\hat{\sigma}_{g_\mathbf{y}}, \hat{\sigma}_{g_\mathbf{x}}, \hat{\sigma}_{f_\mathbf{y}}, \hat{\sigma}_{f_\mathbf{x}}$ such that, for all $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$, the following holds:

D1. $\mathbb{E}\|\hat{\nabla}_\mathbf{y} g_t(\mathbf{z}; \zeta) - \nabla_\mathbf{y} g_{t,\boldsymbol{\rho}}(\mathbf{z})\|^2 \leq \hat{\sigma}_{g_\mathbf{y}}^2$,

D2. $\mathbb{E}\|\hat{\nabla}_\mathbf{x} g_t(\mathbf{z}; \zeta) - \nabla_\mathbf{x} g_{t,\boldsymbol{\rho}}(\mathbf{z})\|^2 \leq \hat{\sigma}_{g_\mathbf{x}}^2$,

D3. $\mathbb{E}\|\hat{\nabla}_\mathbf{y} f_t(\mathbf{z}; \xi) - \nabla_\mathbf{y} f_{t,\boldsymbol{\rho}}(\mathbf{z})\|^2 \leq \hat{\sigma}_{f_\mathbf{y}}^2$,

D4. $\mathbb{E}\|\hat{\nabla}_\mathbf{x} f_t(\mathbf{z}; \xi) - \nabla_\mathbf{x} f_{t,\boldsymbol{\rho}}(\mathbf{z})\|^2 \leq \hat{\sigma}_{f_\mathbf{x}}^2$.

Assumption 4.1 is analogous to the upper bound on the variance of stochastic partial gradients discussed in Luo et al. (2020); Wang et al. (2020). We simplify the notation by introducing the following shorthand.

$$\hat{\sigma}^2 := \hat{\sigma}_{g_\mathbf{y}}^2 + \hat{\sigma}_{g_\mathbf{x}}^2 + \hat{\sigma}_{f_\mathbf{y}}^2 + \hat{\sigma}_{f_\mathbf{x}}^2. \tag{23}$$

Next, we establish a regret bound for ZO-SOGD. Similar to

Algorithm 2 ZO-SOGD

Require: In addition to parameters in SOGD, choose $\rho_{\mathbf{v}}, \rho_{\mathbf{r}}, \rho_{\mathbf{s}}, \in \mathbb{R}_{++}$.

For $t = 1$ **to** T **do**:

S1. Draw samples $\mathcal{B}_t$ and $\bar{\mathcal{B}}_t$ with batch sizes b and $\bar{b}$. Using (19)–(21), get ZO search directions $\hat{\mathbf{d}}_t^{\mathbf{y}}, \hat{\mathbf{d}}_t^{\mathbf{v}}, \hat{\mathbf{d}}_t^{\mathbf{x}}$:

$$\mathbf{d}_t^{\mathbf{y}}\left(\mathbf{z}_t; \bar{\mathcal{B}}_t\right) = \hat{\nabla}_{\mathbf{y}} g_t(\mathbf{z}_t; \bar{\mathcal{B}}_t), \tag{22a}$$
$$\hat{\mathbf{d}}_t^{\mathbf{y}} = \mathbf{d}_t^{\mathbf{y}}(\mathbf{z}_t; \bar{\mathcal{B}}_t) + (1-\gamma_t)(\hat{\mathbf{d}}_{t-1}^{\mathbf{y}} - \mathbf{d}_t^{\mathbf{y}}(\mathbf{z}_{t-1}; \bar{\mathcal{B}}_t)),$$
$$\mathbf{d}_t^{\mathbf{vv}}\left(\mathbf{z}_t; \mathcal{B}_t\right) = \hat{\nabla}_{\mathbf{y}} f_t\left(\mathbf{z}_t; \mathcal{B}_t\right) + \hat{\nabla}^2_{\mathbf{y}} g_t\left(\mathbf{z}_t; \bar{\mathcal{B}}_t\right), \tag{22b}$$
$$\hat{\mathbf{d}}_t^{\mathbf{v}} = \mathbf{d}_t^{\mathbf{vv}}(\mathbf{z}_t; \mathcal{B}_t) + (1-\lambda_t)(\hat{\mathbf{d}}_{t-1}^{\mathbf{v}} - \mathbf{d}_t^{\mathbf{vv}}(\mathbf{z}_{t-1}; \mathcal{B}_t)),$$
$$\mathbf{d}_t^{\mathbf{xy}}\left(\mathbf{z}_t; \mathcal{B}_t\right) = \hat{\nabla}_{\mathbf{x}} f_t\left(\mathbf{z}_t; \mathcal{B}_t\right) + \hat{\nabla}^2_{\mathbf{xy}} g_t\left(\mathbf{z}_t; \bar{\mathcal{B}}_t\right), \tag{22c}$$
$$\hat{\mathbf{d}}_t^{\mathbf{x}} = \mathbf{d}_t^{\mathbf{xy}}(\mathbf{z}_t; \mathcal{B}_t) + (1-\eta_t)(\hat{\mathbf{d}}_{t-1}^{\mathbf{x}} - \mathbf{d}_t^{\mathbf{xy}}(\mathbf{z}_{t-1}; \mathcal{B}_t)),$$

S2. Update inner, system, and outer solutions:

$$\mathbf{y}_{t+1} = \mathbf{y}_t - \beta_t \hat{\mathbf{d}}_t^{\mathbf{y}}, \quad \mathbf{v}_{t+1} = \Pi_{\mathcal{Z}_p}\left[\mathbf{v}_t - \delta_t \hat{\mathbf{d}}_t^{\mathbf{v}}\right],$$
$$\mathbf{x}_{t+1} = \Pi_{\mathcal{X}}\left[\mathbf{x}_t - \alpha_t \hat{\mathbf{d}}_t^{\mathbf{x}}\right].$$

previous results, we introduce regularity conditions for the smoothed functions in (17) and (18).

Inner and Outer Perturbed Gradient Variations: We define the gradient variations at the perturbed point as follows:

$$G_{\mathbf{v},T} := \sum_{t=2}^{T}(\chi_{1t}+\chi_{2t}), \quad G_{\mathbf{x},T} := \sum_{t=2}^{T}(\chi_{3t}+\chi_{4t}). \tag{24}$$

where $\mathbf{z}_t^+ := (\mathbf{x}_{t-1}, \mathbf{y}_{t-1}+\rho_{\mathbf{v}}\mathbf{v}_{t-1})$, $\mathbf{z}_t^- := (\mathbf{x}_{t-1}, \mathbf{y}_{t-1}-\rho_{\mathbf{v}}\mathbf{v}_{t-1})$, and

$$\chi_{1t} := \|\nabla_{\mathbf{y}} g_t(\mathbf{z}_t^+) - \nabla_{\mathbf{y}} g_{t-1}(\mathbf{z}_t^+)\|^2,$$
$$\chi_{2t} := \|\nabla_{\mathbf{y}} g_t(\mathbf{z}_t^-) - \nabla_{\mathbf{y}} g_{t-1}(\mathbf{z}_t^-)\|^2,$$
$$\chi_{3t} := \|\nabla_{\mathbf{x}} g_t(\mathbf{z}_t^+) - \nabla_{\mathbf{x}} g_{t-1}(\mathbf{z}_t^+)\|^2,$$
$$\chi_{4t} := \|\nabla_{\mathbf{x}} g_t(\mathbf{z}_t^-) - \nabla_{\mathbf{x}} g_{t-1}(\mathbf{z}_t^-)\|^2.$$

Further, for simplicity of notation, we define

$$\begin{aligned} \hat{\Delta}_T &:= E_1 + V_T + D_T + G_{\mathbf{y},T}, \\ \hat{\Psi}_T &:= H_{2,T} + G_{\mathbf{v},T} + G_{\mathbf{x},T}, \end{aligned} \tag{25}$$

where $(V_T, H_{p,T})$ and (E_1, D_T) are defined in (10), and (14), repectively. Moreover, $G_{\mathbf{y},T}$ and $(G_{\mathbf{v},T}, G_{\mathbf{x},T})$, are defined in (12) and (24), respectively.

Theorem 4.2. *Let $\{(f_t, g_t)\}_{t=1}^T$ be the sequence of functions presented to Algorithm 2, satisfying Assumptions 3.2-3.4 and 4.1. For all $t \in [T]$, let*

$$\alpha_t = \frac{1}{(d_1+d_2)^{3/4}(c+t)^{1/3}}, \quad \beta_t = c_\beta \alpha_t, \quad \delta_t = c_\delta \alpha_t,$$
$$\gamma_{t+1} = c_\gamma \alpha_t, \quad \eta_{t+1} = c_\eta \alpha_t, \quad \lambda_{t+1} = c_\lambda \alpha_t,$$
$$\rho_{\mathbf{v}}^2 = c_{\mathbf{v}} \alpha_t, \quad \rho_{\mathbf{r}}^2 = \frac{1}{d_2^2 T}, \quad \rho_{\mathbf{s}}^2 = \frac{1}{d_1^2 T},$$
$$b = \frac{T^{1/3}}{(d_1+d_2)^{3/2}}, \quad \bar{b} = \frac{T^{2/3}}{(d_1+d_2)^{3/4}}, \tag{26}$$

where c, c_β, c_δ, c_γ, c_η, $c_{\mathbf{v}}$, and c_λ are specified in (232). Let $p = \ell_{f,0}/\mu_g$ for the set $\mathcal{Z}_p$ defined in (8). Then, Algorithm 2 guarantees:

$$\text{BL-Reg}_T \leq \mathcal{O}\Big((d_1+d_2)^{3/4} T^{1/3}\left(\hat{\sigma}^2 + \hat{\Delta}_T\right) + (d_1+d_2)^{3/2} T^{2/3} \hat{\Psi}_T\Big).$$

where $\hat{\sigma}^2$ and $(\hat{\Delta}_T, \hat{\Psi}_T)$ are defined in (23) and (25).

Theorem 4.2 bounds the regret of Algorithm 2 without window-smoothing, based on the regularities in (25). We note that the average dynamic regret $\text{BL-Reg}_T/T \leq \mathcal{O}((d_1+d_2)^{3/4} T^{-2/3}\left(\hat{\sigma}^2 + \hat{\Delta}_T\right) + (d_1+d_2)^{3/2} T^{-1/3} \hat{\Psi}_T)$ remains sublinear under suitable conditions on $\hat{\Delta}_T$ and $\hat{\Psi}_T$.

Remark 4.3 (**Regret Guarantee for Zeroth Order OBO**). Theorem 4.2 provides the first regret guarantee for OBO with access only to noisy function evaluations of the leader and follower. The dimensional dependence $\mathcal{O}(d_1+d_2)$ in Theorem 4.2 aligns with optimal results for simpler offline min-max problems (Huang et al., 2022). The bound also depends on the sample sizes $b, \bar{b}$ and smoothing parameters $\rho_{\mathbf{v}}, \rho_{\mathbf{r}}, \rho_{\mathbf{s}}$ at each iteration.

Remark 4.4 (**Improved Regret for OSO**). Our dynamic regret for single-level non-stationary optimization is $\mathcal{O}((d_1+d_2)^{3/4} T^{-2/3}(\hat{\sigma}^2 + E_1 + V_T + D_T))$, improving the result in Roy et al. (2022), which is $\mathcal{O}(T^{-1/2}\sigma^2\sqrt{d})$. Roy et al. (2022) proposed a zeroth-order stochastic gradient descent algorithm for unconstrained, non-convex, time-varying objective functions, achieving a regret bound of $\mathcal{O}(T^{-1/2}\sigma^2\sqrt{dW_T})$ using a two-point gradient estimator, where W_T bounds the nonstationarity. Additionally, Guan et al. (2023a) showed that the local regret for standard online stochastic gradient descent with the standard two-point gradient estimator (Agarwal et al., 2010) is $\mathcal{O}(T^{-1/2} d\sqrt{V_T})$.

5. Experimental Results

In this section, we provide experimental results on bilevel optimization-based black-box attacks on deep neural networks and parametric loss tuning for imbalanced data.

5.1. Bilevel Optimization-Based Black-Box Attacks

Deep neural network classifiers are vulnerable to adversarial examples—images subtly modified to mislead the classifier. These examples can deceive classifiers even without knowledge of the model, as seen in black-box adversarial attacks (BBAA) (Chen et al., 2017; Liu et al., 2018b; Chen et al.,

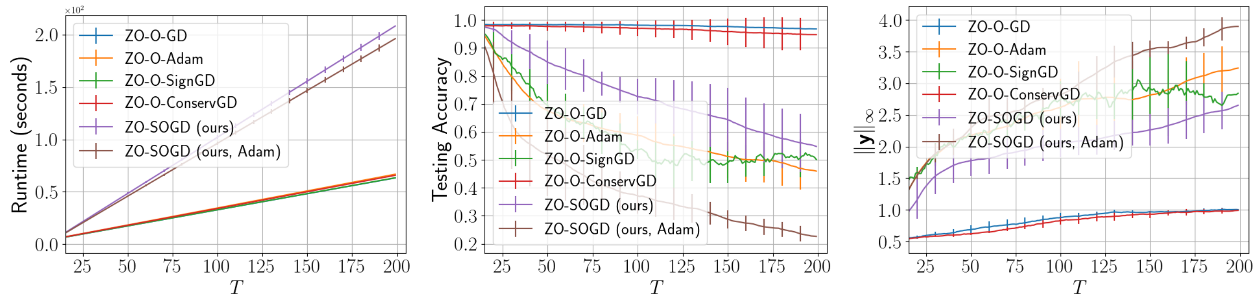

Figure 2. Performance comparison (mean±std) of optimizers including ZO-O-GD, ZO-O-Adam, ZO-O-SignSGD, ZO-O-ConservSGD, ZO-SOGD, and ZO-SOGD (Adam) on **online adversarial attack** for MNIST data across five runs.

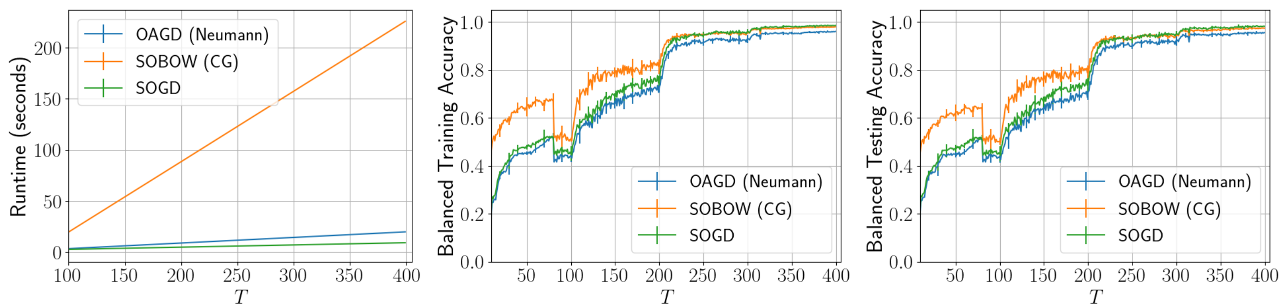

Figure 3. Performance comparison (mean±std) on **imbalanced loss tuning with distribution shift** for MNIST data across five runs between OGD (Zinkevich, 2003), OAGD (Tarzanagh et al., 2024), SOBOW (Lin et al., 2024), and our SOGD.

2019).

We first review the ZO single-level optimization for BBAA (Chen et al., 2017). Let $(\mathbf{a}, b)$ denote a legitimate image $\mathbf{a} \in \mathbb{R}^d$ with true label $b \in \{1, 2, \ldots, J\}$, where J is the total number of classes. Define $\mathbf{a}' = \mathbf{a} + \mathbf{y}$ as an adversarial example, with $\mathbf{y}$ as the adversarial perturbation. Let $\mathcal{Y} := [-5, 5]^d \subset \mathbb{R}^d$, and $\ell : \mathbb{R}^d \to \mathbb{R}$ denote the black-box attack loss. The goal of BBAA (Chen et al., 2017) is to design $\mathbf{y}$ for images $\{\mathbf{a}_i\}_{i=1}^m$ by solving:

$$\min_{\mathbf{y} \in \mathcal{Y}} \quad \frac{1}{m} \sum_{i=1}^{m} \ell(\mathbf{a}_i + \mathbf{y}) + \lambda \|\mathbf{y}\|^2. \tag{27}$$

Here, $\lambda > 0$ is a hyperparameter balancing attack loss minimization and ℓ_2 regularization.

To adapt (27) to our OBO, consider OBO for supervised learning: at each timestep t, new samples $(\mathbf{a}_t, b_t) \in \mathcal{D}_t := \{\mathcal{D}_t^{\text{val}}, \mathcal{D}_t^{\text{tr}}\}$ are received, where $\mathbf{a}_t \in \mathbb{R}^{d_2}$ is the feature vector (image) and $b_t \in \mathbb{R}$ is the corresponding target. Note that the correct decision can change abruptly. We consider an S-stage scenario where $(\mathbf{x}_s^*, \mathbf{y}_s^*(\mathbf{x}_s^*))$ represents the best decisions for the s-th stage, for all $s \in [S]$.

$$\begin{aligned} \mathbf{x}_s^* \in \operatorname*{argmin}_{\mathbf{x} \in \mathcal{X}} & \sum_{t=1}^{T_s} f\left(\mathbf{y}_s^*(\mathbf{x}); \mathcal{D}_t^{\text{val}}\right) \\ \text{s.t.} \quad \mathbf{y}_s^*(\mathbf{x}) \in \operatorname*{argmin}_{\mathbf{y} \in \mathcal{Y}} & \sum_{t=1}^{T_s} g\left(\mathbf{x}, \mathbf{y}; \mathcal{D}_t^{\text{tr}}\right), \end{aligned} \tag{28}$$

where

$$g(\mathbf{x}_t, \mathbf{y}_t; \mathcal{D}_t^{\text{tr}}) = \frac{1}{|\mathcal{D}_t^{\text{tr}}|} \sum_{i \in \mathcal{D}_t^{\text{tr}}} \ell(\mathbf{a}_t^{(i)} + \mathbf{y}_t) + \frac{1}{2} \sum_{\iota=1}^{p} e^{[\mathbf{x}_t]_\iota} [\mathbf{y}_t]_\iota^2, \tag{29a}$$

and

$$f(\mathbf{y}_t(\mathbf{x}_t); \mathcal{D}_t^{\text{val}}) = \frac{1}{|\mathcal{D}_t^{\text{val}}|} \sum_{i \in \mathcal{D}_t^{\text{val}}} \ell(\mathbf{a}_t^{(i)} + \mathbf{y}_t). \tag{29b}$$

Here, $\{\mathbf{a}_t^{(i)}\}_{i \in \mathcal{D}_t^{\text{tr}}}$ and $\{\mathbf{a}_t^{(i)}\}_{i \in \mathcal{D}_t^{\text{val}}}$ are batches of training and validation samples at timestep t; $\mathbf{a}_t^{(i)}$ is the ith sample in that batch; and $[\mathbf{x}_t]_\iota$ and $[\mathbf{y}_t]_\iota$ denote the ιth component of $\mathbf{x}_t$ and $\mathbf{y}_t$, respectively.

We normalize the pixel values to $\mathcal{Y}$. For an untargeted attack, the loss in (29) is $\ell(\mathbf{a}_t') = \max\{Z(\mathbf{a}_t')_{b_t} - \max_{j \neq b_t} Z(\mathbf{a}_t')_j, -\kappa\}$, where $Z(\mathbf{a}_t')_j$ is the prediction score for class j given input $\mathbf{a}_t' = \mathbf{a}_t + \mathbf{y}_t$, and $\kappa > 0$ controls the confidence gap. In our experiments, we set $\kappa = 0$.

Eq. (28) introduces the first OBO formulation of BBAA. Using a vector $\mathbf{x} \in \mathbb{R}_+^d$ for hyperparameters instead of $\lambda \in \mathbb{R}_{++}$ in (27) enables finer control over model components, enhancing performance for complex models and heterogeneous data (Lorraine et al., 2020). For a fair comparison with single-level BBAA, we replace λ with a fixed vector multiplied by each component of $\mathbf{y}$ in (27).

We compare our ZO-SOGD and ZO-SOGD (Adam) with the following competing methods in the online setting:

ZO-O-GD: A single-level method that updates $\mathbf{y}_t$ with a fixed $\mathbf{x}$ at each timestep using ZO gradient descent (Nesterov & Spokoiny, 2017).

ZO-O-Adam: A single-level method that updates $\mathbf{y}_t$ with a fixed $\mathbf{x}$ at each timestep using ZO Adam (Kingma & Ba, 2014; Chen et al., 2019).

ZO-O-SignSGD: A single-level method that updates $\mathbf{y}_t$ with a fixed $\mathbf{x}$ at each timestep using ZO SignSGD (Bernstein et al., 2018).

ZO-O-ConservSGD: A single-level method that updates $\mathbf{y}_t$ with a fixed $\mathbf{x}$ at each timestep using ZO Conservative SGD (Cutkosky & Boahen, 2019).

Note that ZO-SOGD (Adam) is a variant of our algorithm with an adaptive stepsize, similar to that of (Kingma & Ba, 2014).

We evaluated the proposed algorithms based on runtime, test accuracy on perturbed samples, and the infinity norm of $\mathbf{y}_t$. Figure 2 compares the methods. The left panel shows that ZO-SOGD has similar runtime to single-level baselines, despite outer-level optimization on $\mathbf{x}$. The middle panel shows that all methods' accuracy decreases as the adversarial attack $\mathbf{y}$ strengthens, with ZO-SOGD outperforming ZO-O-GD and ZO-O-ConservGD, and ZO-SOGD (Adam) outperforming ZO-O-Adam and all baselines. The right panel shows that the increasing infinity norm of $\mathbf{y}_t$ over time for all methods, which reduces accuracy. However, the perturbations remain unnoticeable with a max $\mathbf{y}_t$ no larger than 4, demonstrating that ZO-SOGD achieves effective attacks with better performance than other methods.

5.2. Parametric Loss Tuning for Imbalanced Data

Imbalanced datasets are common in modern machine learning, causing challenges in generalization and fairness due to underrepresented classes and sensitive attributes. Deep NNs often overfit, seeming accurate and fair during training but performing poorly during testing. A common solution is designing a parametric training loss that balances accuracy and fairness while preventing overfitting (Li et al., 2021).

We consider an optimization problem similar to (28). For a new sample $(\mathbf{a}_t, b_t)$, the follower and leader incur a parametric and balanced cross-entropy loss, respectively:

$$g(\mathbf{x}_t, \mathbf{y}_t; \mathcal{D}_t^{\text{tr}}) = -\log \frac{e^{\gamma_{b_t}[\mathbf{y}_t(\mathbf{a}_t)]_{b_t} + \Delta_{b_t}}}{\sum_{j=1}^J e^{\gamma_j [\mathbf{y}_t(\mathbf{a}_t)]_j + \Delta_j}}, \quad \text{and}$$

$$f(\mathbf{y}_t(\mathbf{x}_t); \mathcal{D}_t^{\text{val}}) = -u_{b_t} \log \frac{e^{[\mathbf{y}_t(\mathbf{a}_t)]_{b_t}}}{\sum_{j=1}^J e^{[\mathbf{y}_t(\mathbf{a}_t)]_j}}. \tag{30}$$

Here, $\mathbf{x}_t := (\Delta_j, \gamma_j)_{j=1}^J$ represents the logits adjustments, with j indexing the J classes, and u_j is the reciprocal of the proportion of samples from the j-th class to the total number of samples (Li et al., 2021).

To clarify the notation in (30): $\mathbf{y}_t(\mathbf{x}_t)$ denotes the follower $\mathbf{y}_t$ conditioned on the leader $\mathbf{x}_t$, while $[\mathbf{y}_t(\mathbf{a}_t)]_{b_t}$ represents the predicted logit for class b_t on sample $\mathbf{a}_t$. The backbone model for $\mathbf{y}_t$ is a 4-layer CNN, leading to a nonconvex bilevel objective.

We compare SOGD with the following methods:

OAGD (Tarzanagh et al., 2024): A state-of-the-art static online bilevel gradient descent method using the Neumann series for hypergradient approximation.

SOBOW (Lin et al., 2024): A dynamic online bilevel gradient descent method using conjugate gradients (CG) for hypergradient approximation.

We conducted experiments on the MNIST (LeCun et al., 2010). We used a batch size of 64 per timestep. We evaluated cumulative runtime, balanced accuracy, and test accuracy, where balanced accuracy is the class-specific average accuracy:

$$\frac{1}{J} \sum_{j=1}^J \mathbb{P}_{\mathbf{a}_t \sim \mathcal{D}_j} \left[\operatorname{argmax}_i([\mathbf{y}_t(\mathbf{a}_t)]_i) = j\right],$$

with $\mathcal{D}_j$ denoting the distribution over samples of class j (Li et al., 2021). Learning rates were tuned as $\beta_t = \delta_t = \beta \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$ and $\alpha_t = \alpha \in \{0.0001, 0.0005, 0.001, 0.005, 0.01\}$ for all $t \in [T]$. The parameters $\gamma_t, \lambda_t, \eta_t$ were tuned as $\gamma_t = \lambda_t = \eta_t = \gamma \in \{0.9, 0.99, 0.999\}$. The Neumann series iterations in OAGD and CG iterations in SOBOW were set to 5.

We evaluated performance over 400 timesteps in four 100-timestep phases, transitioning from an imbalanced (0.4^i) to a balanced (0.8^i) distribution for each class ($i = 0, 1, \dots, 9$). Figure 3 (left) shows SOBOW's longer runtime due to CG complexity, while SOGD is the fastest with simultaneous updates. Figures 3 (middle, right) show accuracy gains as balance increases, with SOGD achieving competitive accuracy.

6. Conclusion

We introduced a novel online bilevel optimization (OBO) framework that overcomes the limitations of existing algorithms, which often rely on extensive oracle information and incur high computational costs. Our approach uses limited feedback and zeroth-order updates for efficient hypergradient estimation and simultaneous updates of decision variables, achieving sublinear bilevel regret without window smoothing. Experiments on online parametric loss tuning and black-box adversarial attacks confirm its effectiveness.

Impact Statements

This paper develops methods to advance online learning. While our work has societal implications, none require specific emphasis here.

References

Agarwal, A., Dekel, O., and Xiao, L. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Colt*, pp. 28–40. Citeseer, 2010.

Agarwal, N., Gonen, A., and Hazan, E. Learning in non-convex games with an optimization oracle. In *Conference on Learning Theory*, pp. 18–29. PMLR, 2019.

Aghasi, A. and Ghadimi, S. Fully zeroth-order bilevel programming via gaussian smoothing. *arXiv preprint arXiv:2404.00158*, 2024.

Allen-Zhu, Z. and Li, Y. Neon2: Finding local minima via first-order oracles. *Advances in Neural Information Processing Systems*, 31, 2018.

Bach, F. and Perchet, V. Highly-smooth zero-th order online optimization. In *Conference on Learning Theory*, pp. 257–283. PMLR, 2016.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. Signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018.

Besbes, O., Gur, Y., and Zeevi, A. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Bohne, J., Rosenberg, D., Kazantsev, G., and Polak, P. Online nonconvex bilevel optimization with bregman divergences. *arXiv preprint arXiv:2409.10470*, 2024.

Bracken, J. and McGill, J. T. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.

Bubeck, S., Stoltz, G., Szepesvári, C., and Munos, R. Online optimization in x-armed bandits. *Advances in Neural Information Processing Systems*, 21, 2008.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pp. 15–26, 2017.

Chen, T., Sun, Y., and Yin, W. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34, 2021.

Chen, X., Liu, S., Xu, K., Li, X., Lin, X., Hong, M., and Cox, D. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *Advances in neural information processing systems*, 32, 2019.

Crockett, C., Fessler, J. A., et al. Bilevel methods for image reconstruction. *Foundations and Trends® in Signal Processing*, 15(2-3):121–289, 2022.

Cutkosky, A. and Boahen, K. Anytime online-to-batch conversions and the conservative algorithm. *Advances in Neural Information Processing Systems*, 32, 2019.

Dagréou, M., Ablin, P., Vaiter, S., and Moreau, T. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *arXiv preprint arXiv:2201.13409*, 2022.

Dempe, S. *Foundations of bilevel programming*. Springer Science & Business Media, 2002.

Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Finn, C., Rajeswaran, A., Kakade, S., and Levine, S. Online meta-learning. In *International Conference on Machine Learning*, pp. 1920–1930. PMLR, 2019.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

Franceschi, L., Donini, M., Frasconi, P., and Pontil, M. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pp. 1165–1173. PMLR, 2017.

Franceschi, L., Frasconi, P., Salzo, S., Grazzi, R., and Pontil, M. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pp. 1568–1577. PMLR, 2018.

Gao, X., Li, X., and Zhang, S. Online learning with non-convex losses and non-stationary regret. In *International Conference on Artificial Intelligence and Statistics*, pp. 235–243. PMLR, 2018.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Ghadimi, S. and Wang, M. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Ghadimi, S., Lan, G., and Zhang, H. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1-2):267–305, 2016.

Goel, G., Lin, Y., Sun, H., and Wierman, A. Beyond online balanced descent: An optimal algorithm for smoothed online optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Guan, Z., Zhou, Y., and Liang, Y. On the hardness of online nonconvex optimization with single oracle feedback. In *The Twelfth International Conference on Learning Representations*, 2023a.

Guan, Z., Zhou, Y., and Liang, Y. Online nonconvex optimization with limited instantaneous oracle feedback. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 3328–3355. PMLR, 2023b.

Hallak, N., Mertikopoulos, P., and Cevher, V. Regret minimization in stochastic non-convex learning via a proximal-gradient approach. In *International Conference on Machine Learning*, pp. 4008–4017. PMLR, 2021.

Hansen, P., Jaumard, B., and Savard, G. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on scientific and Statistical Computing*, 13(5):1194–1217, 1992.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016a. URL http://ocobook.cs.princeton.edu/OCObook.pdf.

Hazan, E. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016b.

Hazan, E., Agarwal, A., and Kale, S. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

Hazan, E., Singh, K., and Zhang, C. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*, pp. 1433–1441. PMLR, 2017.

Héliou, A., Martin, M., Mertikopoulos, P., and Rahier, T. Online non-convex optimization with imperfect feedback. *Advances in Neural Information Processing Systems*, 33: 17224–17235, 2020.

Héliou, A., Martin, M., Mertikopoulos, P., and Rahier, T. Zeroth-order non-convex learning via hierarchical dual averaging. In *International Conference on Machine Learning*, pp. 4192–4202. PMLR, 2021.

Huang, F., Gao, S., Pei, J., and Huang, H. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70, 2022.

Huang, Y., Cheng, Y., Liang, Y., and Huang, L. Online min-max problems with non-convexity and non-stationarity. *Transactions on Machine Learning Research*.

Huang, Y., Cheng, Y., Liang, Y., and Huang, L. Online min-max problems with non-convexity and non-stationarity. *Transactions on Machine Learning Research*, 2023.

Ji, K., Wang, Z., Zhou, Y., and Liang, Y. Improved zeroth-order variance reduced algorithms and analysis for non-convex optimization. In *International conference on machine learning*, pp. 3100–3109. PMLR, 2019.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pp. 4882–4892. PMLR, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014.

Kleinberg, R., Slivkins, A., and Upfal, E. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 681–690, 2008.

Krichene, W., Balandat, M., Tomlin, C., and Bayen, A. The hedge algorithm on a continuum. In *International Conference on Machine Learning*, pp. 824–832. PMLR, 2015.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Li, M., Zhang, X., Thrampoulidis, C., Chen, J., and Oymak, S. Autobalance: Optimized loss functions for imbalanced data. *Advances in Neural Information Processing Systems*, 34:3163–3177, 2021.

Lin, S., Sow, D., Ji, K., Liang, Y., and Shroff, N. Non-convex bilevel optimization with time-varying objective functions. *Advances in Neural Information Processing Systems*, 36, 2024.

Liu, H., Simonyan, K., and Yang, Y. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018a.

Liu, S., Chen, J., Chen, P.-Y., and Hero, A. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In *International Conference on Artificial Intelligence and Statistics*, pp. 288–297. PMLR, 2018b.

Lorraine, J., Vicol, P., and Duvenaud, D. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pp. 1540–1552. PMLR, 2020.

Luo, L., Ye, H., Huang, Z., and Zhang, T. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33:20566–20577, 2020.

Lv, Y., Hu, T., Wang, G., and Wan, Z. A penalty function method based on kuhn–tucker condition for solving linear bilevel programming. *Applied Mathematics and Computation*, 188(1):808–813, 2007.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103:127–152, 2005.

Nesterov, Y. and Spokoiny, V. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Roy, A., Balasubramanian, K., Ghadimi, S., and Mohapatra, P. Stochastic zeroth-order optimization under nonstationarity and nonconvexity. *Journal of Machine Learning Research*, 23(64):1–47, 2022.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

Shamir, O. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

Sow, D., Ji, K., and Liang, Y. On the convergence theory for hessian-free bilevel algorithms. *Advances in Neural Information Processing Systems*, 35:4136–4149, 2022.

Stackelberg, H. v. Theory of the market economy. *Oxford University Press*, 1952.

Stadie, B., Zhang, L., and Ba, J. Learning intrinsic rewards as a bi-level optimization problem. In *Conference on Uncertainty in Artificial Intelligence*, pp. 111–120. PMLR, 2020.

Suggala, A. S. and Netrapalli, P. Online non-convex learning: Following the perturbed leader is optimal. In *Algorithmic Learning Theory*, pp. 845–861. PMLR, 2020.

Tarzanagh, D. A., Nazari, P., Hou, B., Shen, L., and Balzano, L. Online bilevel optimization: Regret analysis of online alternating gradient methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 2854–2862. PMLR, 2024.

Wang, Z., Balasubramanian, K., Ma, S., and Razaviyayn, M. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*, 2020.

Zhang, Y., Zhou, Y., Ji, K., and Zavlanos, M. M. Boosting one-point derivative-free online optimization via residual feedback. *arXiv preprint arXiv:2010.07378*, 2020.

Zhou, W., Li, Y., Yang, Y., Wang, H., and Hospedales, T. Online meta-critic learning for off-policy actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17662–17673, 2020.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pp. 928–936, 2003.

A. Related Work

BO was introduced in game theory by (Stackelberg, 1952) and modeled mathematically in (Bracken & McGill, 1973). Initial works (Hansen et al., 1992; Lv et al., 2007) reduced it to single-level optimization. Recently, gradient-based approaches have gained popularity for their simplicity and efficacy (Franceschi et al., 2017; Ghadimi & Wang, 2018; Ji et al., 2021; Chen et al., 2021), though they assume offline objectives.

OBO was initiated by Tarzanagh et al. (2024), proposing the OAGD method with regret bounds. (Huang et al., 2023) developed algorithms for online minimax optimization, special cases of OBO with local regret guarantees. (Lin et al., 2024) introduced SOBOW, a single-loop optimizer using window-smoothed functions and multiple CGs for nonconvex-strongly-convex cases. Unlike these works, we propose using *projected gradient* as a more general performance measure for constrained objectives, focusing on the original functions and their regret; See Table 1 for a comparison.

Single-Level Regret Minimization. Single-level online optimization predominantly focuses on convex problems, either with static or dynamic convex regret minimization (Zinkevich, 2003; Hazan, 2016a; Shalev-Shwartz et al., 2011). Non-convex online optimization (Hazan et al., 2017; Guan et al., 2023b;a) poses greater challenges than its convex counterparts (Shalev-Shwartz et al., 2011; Zinkevich, 2003; Hazan et al., 2007; Besbes et al., 2015). Notable contributions in this field include adversarial multi-armed bandit algorithms (Bubeck et al., 2008; Héliou et al., 2020; 2021; Krichene et al., 2015) and the Follow-the-Perturbed-Leader approach (Agarwal et al., 2019; Kleinberg et al., 2008; Suggala & Netrapalli, 2020). Hazan et al. (Hazan et al., 2017) introduced window-smoothed local regret for gradient averaging in non-convex models, which Hallak et al. (Hallak et al., 2021) extended to non-smooth, non-convex problems. Inspired by their work, we employ local regret for Online Bandit Optimization (OBO) without window-smoothing.

Zeroth-Order Optimization. Single-Level ZO Optimization has been widely studied in both offline (Ghadimi & Lan, 2013; Duchi et al., 2015; Agarwal et al., 2010; Nesterov & Spokoiny, 2017) and online settings (Liu et al., 2018b; Guan et al., 2023a;b; Zhang et al., 2020; Bach & Perchet, 2016). We next review closely related work. Liu et al. (Liu et al., 2018b) proposed ZOO-ADMM, a gradient-free online optimization algorithm utilizing ADMM. Guan et al. (Guan et al., 2023b) studied online non-convex optimization with limited oracle feedback. Research on online non-convex optimization with bandit feedback includes work by Heliou et al. (Héliou et al., 2020), which established bounds on global static and dynamic regret using dual averaging, further refined in (Héliou et al., 2021). Gao et al. (Gao et al., 2018) extended these ideas to ZO algorithms. Flaxman et al. (Flaxman et al., 2004) provided algorithms for bandit online optimization of convex functions using ZO gradient approximation. Our work closely relates to (Sow et al., 2022), which proposes a Hessian-free method approximating the Jacobian matrix using a ZO method based on finite differences of gradients. In contrast, our method uses function oracles to approximate both the Hessian and gradients and is derivative-free. We also point out the recent work (Aghasi & Ghadimi, 2024) on ZO stochastic algorithms for solving bilevel problems when neither the upper/lower objective values nor their unbiased gradient estimates are available. Their approach, limited to the *offline* setting, does not include numerical results, thus leaving its practical efficiency unclear.

B. Additional Preliminaries and Notations

B.1. Preliminary Lemmas

We first provide several useful lemmas for the main proofs.

Definition B.1 (**Projected gradient** (Ghadimi et al., 2016))**.** Let $\mathcal{X} \subset \mathbb{R}^{d_1}$ be a closed convex set. Then, the projected gradient for any $\alpha_t > 0$ and $\mathbf{p} \in \mathbb{R}^{d_1}$ is defined as

$$\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x};\mathbf{p}) := \frac{1}{\alpha_t}\left(\mathbf{x}-\mathbf{x}^+\right), \tag{31a}$$

where

$$\mathbf{x}^+ = \Pi_{\mathcal{X}}\left(\mathbf{x}-\alpha_t\mathbf{p}\right), \tag{31b}$$

and $\Pi_{\mathcal{X}}[\cdot]$ denotes the orthogonal projection operator onto set $\mathcal{X}$.

Lemma B.2. Goel et al. (2019, Lemma 13) *If $f : \mathcal{X} \to \mathbb{R}$ is a μ_f-strongly convex function with respect to some norm $\|\cdot\|$, and $\mathbf{x}^*$ is the minimizer of f (i.e. $\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$), then we have $\forall\, \mathbf{x} \in \mathcal{X}$,*

$$\frac{\mu_f}{2}\|\mathbf{x}-\mathbf{x}^*\|^2 \le f(\mathbf{x}) - f(\mathbf{x}^*) \le \frac{1}{2\mu_f}\|\nabla f(\mathbf{x})\|^2.$$

Lemma B.3. *Suppose $f(\mathbf{x})$ is L-smooth, and $\mathbf{x}^* \in argmin_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})$. Then, we can upper bound the magnitude of the* *gradient at any given point $\mathbf{x} \in \mathbb{R}^d$ in terms of the objective sub optimality at $\mathbf{x}$, as follows:*

$$\frac{1}{2L}\|\nabla f(\mathbf{x})\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{L}{2}\|\mathbf{x} - \mathbf{x}^*\|^2. \tag{32}$$

Lemma B.4. *For any set of vectors $\{\mathbf{x}_i\}_{i=1}^m$ with $\mathbf{x}_i \in \mathbb{R}^d$, we have*

$$\left\|\sum_{i=1}^m \mathbf{x}_i\right\|^2 \leq m\sum_{i=1}^m \|\mathbf{x}_i\|^2.$$

Lemma B.5. *For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the following holds for any $c > 0$:*

$$\|\mathbf{x} + \mathbf{y}\|^2 \leq (1 + c)\|\mathbf{x}\|^2 + \left(1 + \frac{1}{c}\right)\|\mathbf{y}\|^2, \text{ and} \tag{33}$$

$$\|\mathbf{x} - \mathbf{y}\|^2 \geq (1 - c)\|\mathbf{x} - \mathbf{z}\|^2 + \left(1 - \frac{1}{c}\right)\|\mathbf{z} - \mathbf{y}\|^2. \tag{34}$$

We provide a set of auxiliary lemmas that will be used in establishing the proofs for the main theorems.

Lemma B.6. Ghadimi et al. (2016, Proposition 1) *Let $\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p})$ be defined in Definition B.1. Then, for any $\mathbf{p}_1$ and $\mathbf{p}_2$ in $\mathbb{R}^d$, we*

$$\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p}_1) - \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p}_2)\| \leq \|\mathbf{p}_1 - \mathbf{p}_2\|.$$

Lemma B.7. Hazan et al. (2017, Proposition 2.4) *Let $\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p})$ be the projected gradient as per Definition B.1. For any $\mathbf{x}, \mathbf{p}_1, \mathbf{p}_2 \in \mathbb{R}^d$ and $\alpha_t > 0$ it holds that*

$$\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p}_1 + \mathbf{p}_2)\| \leq \|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p}_1)\| + \|\mathbf{p}_2\|.$$

Lemma B.8. *Let $\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p})$ be as given in Definition B.1. Then, for any $\mathbf{p} \in \mathbb{R}^d$ and $\alpha_t > 0$, we have*

$$\langle \mathbf{p}, \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p})\rangle \geq \|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p})\|^2.$$

Proof. By the definition of $\mathbf{x}^+$, the optimality condition of (31b) is

$$\left\langle \mathbf{p} + \frac{1}{\alpha_t}(\mathbf{x}^+ - \mathbf{x}), \mathbf{z} - \mathbf{x}^+\right\rangle \geq 0, \quad \forall \mathbf{z} \in \mathcal{X}.$$

Letting $\mathbf{z} = \mathbf{x}$, we obtain

$$\langle \mathbf{p}, \mathbf{x} - \mathbf{x}^+\rangle \geq \frac{1}{\alpha_t}\langle \mathbf{x} - \mathbf{x}^+, \mathbf{x} - \mathbf{x}^+\rangle,$$

which can be rearranged to

$$\begin{aligned}\langle \mathbf{p}, \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p})\rangle = \frac{1}{\alpha_t}\langle \mathbf{p}, \mathbf{x} - \mathbf{x}^+\rangle &\geq \frac{1}{\alpha_t^2}\langle \mathbf{x} - \mathbf{x}^+, \mathbf{x} - \mathbf{x}^+\rangle \\ &= \|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}; \mathbf{p})\|^2.\end{aligned}$$

□

B.2. Examples

Theorem 3.6 achieves sublinear bilevel regret when the variations V_T and $H_{2,T}$ are both $o(T)$. Below, we provide some examples of online optimization in both single-level and bilevel settings to illustrate when this occurs.

Example B.9. Consider function $f_t(\mathbf{x}) = \|\mathbf{A}_t\mathbf{x} - \mathbf{b}_t\|^2$, where $\mathbf{A}_t = [1, 0; 0, 1 + \frac{1}{t}]$, $\mathbf{x} = \mathbf{b}_t = (1, 1)$. Then, $V_T :=$

$\sum_{t=2}^{T} \max_{\mathbf{x}} |f_t(\mathbf{x}) - f_{t-1}(\mathbf{x})| = \sum_{t=2}^{T} | \left(\frac{1}{t}\right)^2 - \left(\frac{1}{t-1}\right)^2 |$. By $a^2 - b^2 = (a-b)(a+b)$, we have

$$
\begin{aligned}
V_T &= \sum_{t=2}^{T} | \left(\frac{1}{t} - \frac{1}{t-1}\right) - \left(\frac{1}{t} + \frac{1}{t-1}\right) | \\
&= \sum_{t=2}^{T} | \left(\frac{t-1-t}{t(t-1)}\right) - \left(\frac{1}{t} + \frac{1}{t-1}\right) | \\
&= \sum_{t=2}^{T} | \left(-\frac{1}{t(t-1)}\right) - \left(\frac{1}{t} + \frac{1}{t-1}\right) | \\
&= \sum_{t=2}^{T} |\frac{1}{t(t-1)}||\frac{t-1+t}{t(t-1)}| \\
&= \sum_{t=2}^{T} |\frac{2}{t(t-1)^2}|.
\end{aligned}
$$

Then, $V_T \leq \sum_{t=2}^{T} \frac{2}{t^3} \approx \int_2^T \frac{2}{t^3} dt = \frac{1}{4} - \frac{1}{T^2}$. As $T \to \infty$, V_T becomes bounded and approaches a constant value, indicating that V_T grows slower than T itself.

Example B.10. Let $f_t(\mathbf{x}) = (-\frac{1}{T}, 0, 0, 0)$ if t is even, and $f_t(\mathbf{x}) = (0, -\frac{1}{T}, 0, 0)$ if t is odd. Then, $V_T = \sum_{t=2}^{T} \max_{\mathbf{x}} |f_t(\mathbf{x}) - f_{t-1}(\mathbf{x})| = \mathcal{O}(1)$.

Example B.11. Let $x \in \mathcal{X} = [-1, 1] \subset \mathbb{R}$, $y \in \mathbb{R}$, and consider a sequence of quadratic cost functions

$$
\begin{aligned}
f_t(x, y) &= \frac{1}{2}\left(x + 2a_t^{(1)}\right)^2 + \frac{1}{2}\left(y - a_t^{(2)}\right)^2, \\
g_t(x, y) &= \frac{1}{2}y^2 - \left(x - a_t^{(2)}\right)y,
\end{aligned}
$$

where $a_t^{(1)} = 1/t$ and $a_t^{(2)} = 1/\sqrt{t}$ for all $t \in [T]$.

We have

$$
y_t^*(x) = x - a_t^{(2)}.
$$

We have

$$
\begin{aligned}
& f_t(x, y_t^*(x)) - f_{t-1}(x, y_{t-1}^*(x)) \\
&= \frac{1}{2}\left[\left(x + 2a_t^{(1)}\right)^2 - \left(x + 2a_{t-1}^{(1)}\right)^2\right] + \frac{1}{2}\left[\left(y_t^*(x) - a_t^{(2)}\right)^2 - \left(y_{t-1}^*(x) - a_{t-1}^{(2)}\right)^2\right] \\
&= \frac{1}{2}\left[\left(x^2 + 4xa_t^{(1)} + 4(a_t^{(1)})^2\right) - \left(x^2 + 4xa_{t-1}^{(1)} + 4(a_{t-1}^{(1)})^2\right)\right] \\
&+ \frac{1}{2}\left[\left((x - a_t^{(2)})^2 - 2(x - a_t^{(2)})a_t^{(2)} + (a_t^{(2)})^2\right) - \left((x - a_{t-1}^{(2)})^2 - 2(x - a_{t-1}^{(2)})a_{t-1}^{(2)} + (a_{t-1}^{(2)})^2\right)\right] \\
&= 2x\left(a_t^{(1)} - a_{t-1}^{(1)} - a_t^{(2)} + a_{t-1}^{(2)}\right) + 2\left((a_t^{(1)})^2 - (a_{t-1}^{(1)})^2 + (a_t^{(2)})^2 - (a_{t-1}^{(2)})^2\right).
\end{aligned}
$$

Taking the maximum over x and using $x \in [-1, 1]$:

$$
\begin{aligned}
\sup_x |f_t(x, y_t^*(x)) - f_{t-1}(x, y_{t-1}^*(x))| = 2\left|a_t^{(1)} - a_{t-1}^{(1)}\right| + 2\left|-a_t^{(2)} + a_{t-1}^{(2)}\right| \\
+ 2\left|(a_t^{(1)})^2 - (a_{t-1}^{(1)})^2\right| + 2\left|(a_t^{(2)})^2 - (a_{t-1}^{(2)})^2\right|.
\end{aligned}
$$

Since $a_t^{(1)} = 1/t$ and $a_t^{(2)} = 1/\sqrt{t}$ for all $t \in [T]$, then we have

$$|a_t^{(1)} - a_{t-1}^{(1)}| \approx \frac{1}{t^2}, \quad |a_t^{(2)} - a_{t-1}^{(2)}| \approx \frac{1}{2t^{3/2}},$$
$$|(a_t^{(1)})^2 - (a_{t-1}^{(1)})^2| \approx \frac{1}{t^3}, \quad |(a_t^{(2)})^2 - (a_{t-1}^{(2)})^2| \approx \frac{1}{t^2}.$$

Then, we get

$$V_T := \sum_{t=2}^{T} \sup_x |f_t(x, y_t^*(x)) - f_{t-1}(x, y_{t-1}^*(x))| = \sum_{t=2}^{T} \left(\frac{2}{t^2} + \frac{1}{2t^{3/2}} + \frac{1}{t^3} \right).$$

The series $\sum_{t=2}^{T} \left(\frac{2}{t^2} + \frac{1}{2t^{3/2}} + \frac{1}{t^3} \right)$ converges, implying $V_T = \mathcal{O}(1)$. Moreover, we have

$$\begin{aligned} H_{2,T} &= \sum_{t=2}^{T} \sup_x \|y_t^*(x) - y_{t-1}^*(x)\|^2 = \sum_{t=2}^{T} \sup_x \|x - a_t^{(2)} - x + a_{t-1}^{(2)}\|^2 \\ &= \sum_{t=2}^{T} |-a_t^{(2)} + a_{t-1}^{(2)}|^2 = \sum_{t=2}^{T} |a_t^{(2)} - a_{t-1}^{(2)}|^2 \approx \sum_{t=2}^{T} \frac{1}{4t^3}, \end{aligned}$$

which implies $H_{2,T} = \mathcal{O}(1)$.

To achieve $V_T = o(T)$, the changes in the cost functions $f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))$ and $\mathbf{y}_t^*(\mathbf{x})$ should decay to zero faster than $\mathcal{O}(1/t)$. For example, if the coefficients in the functions change as $\mathcal{O}(1/t^a)$ with $a > 1$, then the cumulative sum over T will be $o(T)$. When $f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))$ and $\mathbf{y}_t^*(\mathbf{x})$ decay as $\mathcal{O}(1/\sqrt{t})$, then the total variation grows at most as $\mathcal{O}(\sqrt{T})$.

C. Proof of Regret Bounds for Simultaneous Online Gradient Descent (SOGD)

Proof Roadmap. We introduce Lemma C.2, which quantifies the error between the approximated direction of the momentum-based gradient estimator, $\mathbf{d}_t^{\mathbf{y}}$, and the true direction, $\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)$, at each iteration. To bound the error of the lower-level variable, we provide Lemma C.4, which captures the gap $\|\mathbf{y}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\|^2$ and incorporates the error introduced in Lemma C.2. Moreover, we provide Lemma C.5, which quantifies the error between the approximated direction of the momentum-based gradient estimator, $\mathbf{d}_t^{\mathbf{v}}$, and the true direction, $\nabla_{\mathbf{y}}^2 g_t(\mathbf{z}_t) \mathbf{v}_t + \nabla_{\mathbf{y}} f_t(\mathbf{z}_t)$, at each iteration. To bound the error of the system solution, we provide Lemma C.8, which captures the gap $\|\mathbf{v}_{t+1} - \mathbf{v}_t^*(\mathbf{x}_t)\|^2$ and incorporates the error introduced in Lemma C.5. Moreover, we provide Lemma C.9, which quantifies the error between the approximated direction of the momentum-based hypergradient estimator, $\mathbf{d}_t^{\mathbf{x}}$, and the true direction, $\nabla_{\mathbf{x}} f_t(\mathbf{z}_t) + \nabla_{\mathbf{xy}}^2 g_t(\mathbf{z}_t) \mathbf{v}_t$, at each iteration. We also present Lemma C.11, which provides an upper bound for the projection mapping and relates to the three errors discussed in Lemmas C.4, C.8, and C.9. Finally, by combining these lemmas and appropriately setting the parameters, we achieve the desired result.

C.1. Proof of Lemma 3.1

Proof. SOBOW (Lin et al., 2024) has estimated the hypergradient as the weighted average of previous ones over a sliding window of size w for a given $\mathcal{B}_t := \{\xi_1, \ldots, \xi_b\}$ drawn i.i.d. from the distribution $\mathcal{D}_f$, as follows:

$$\hat{\nabla} F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) = \frac{1}{W} \sum_{i=0}^{w-1} \nu^i \hat{\nabla} f_{t-i}(\mathbf{x}_{t-i}, \mathbf{y}_{t-i}; \mathcal{B}_{t-i}),$$

with $W = \sum_{i=0}^{w-1} \nu^i, \ \nu \in (0,1)$. Let $\nu = 1 - \eta$ for $\eta \in (0,1)$.

Then, the above equality is equivalent to

$$\hat{\nabla} F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) = \frac{1}{W} \sum_{j=t-w+1}^{t} (1-\eta)^{t-j} \hat{\nabla} f_j(\mathbf{x}_j, \mathbf{y}_j; \mathcal{B}_j), \tag{35}$$

with $W = \sum_{j=t-w+1}^{t} (1-\eta)^{t-j}$.

Let $\hat{\mathbf{d}}_t^{\mathbf{x}} := \hat{\nabla} F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t)$. Then (35) is equivalent to

$$\hat{\mathbf{d}}_t^{\mathbf{x}} = \frac{1}{W}\hat{\nabla} f_t(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) + (1-\eta)\hat{\mathbf{d}}_{t-1}^{\mathbf{x}} - \frac{(1-\eta)^w}{W}\hat{\nabla} f_{t-w}(\mathbf{x}_{t-w}, \mathbf{y}_{t-w}; \mathcal{B}_{t-w}), \tag{36}$$

with $f_i(\cdot) = 0$ for all $i \leq 0$.

If $w = t$ and $W = \frac{1}{\eta}$, then, we have

$$\hat{\mathbf{d}}_t^{\mathbf{x}} = \eta \hat{\nabla} f_t(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) + (1-\eta)\hat{\mathbf{d}}_{t-1}^{\mathbf{x}}.$$

□

C.2. Bounds on the Inner Decision Variable

We first provide a lemma that characterizes the Lipschitz continuity of approximate gradients, inner, and system solutions.

Lemma C.1. *Under Assumptions 3.2 and 3.3, for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and the search directions $\{\mathbf{d}_t^{\mathbf{x}}\}_{t=1}^T$ and $\{\mathbf{d}_t^{\mathbf{v}}\}_{t=1}^T$ generated by Algorithm 1, we have*

$$\|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\|^2 \leq M_f^2 \left(\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 + \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2\right), \tag{37a}$$

$$\|\mathbf{d}_t^{\mathbf{v}}\|^2 \leq M_{\mathbf{v}}^2 \left(\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 + \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2\right), \tag{37b}$$

$$\|\nabla f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) - \nabla f_t(\mathbf{x}', \mathbf{y}_t^*(\mathbf{x}'))\| \leq L_f \|\mathbf{x} - \mathbf{x}'\|, \tag{37c}$$

$$\|\mathbf{y}_t^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x}')\| \leq L_{\mathbf{y}} \|\mathbf{x} - \mathbf{x}'\|, \tag{37d}$$

$$\|\mathbf{v}_t^*(\mathbf{x}) - \mathbf{v}_t^*(\mathbf{x}')\| \leq L_{\mathbf{v}} \|\mathbf{x} - \mathbf{x}'\|, \tag{37e}$$

where M_f, $M_{\mathbf{v}}$, and $(L_{\mathbf{y}}, L_{\mathbf{v}}, L_f)$ are defined in (40), (41), and (42), respectively.

Proof. We first show (37a).

Using Assumptions 3.2 and 3.3, we have $\nabla^2_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \succeq \mu_g$, and

$$\|\mathbf{v}_t^*(\mathbf{x}_t)\| = \|\left(\nabla^2_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)^{-1} \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| \leq \frac{\ell_{f,0}}{\mu_g}. \tag{38}$$

Observe that

$$\begin{aligned}
\|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| &\leq \|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| \\
&\quad + \|\mathbf{v}_t \nabla^2_{\mathbf{xy}} g_t(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{v}_t^*(\mathbf{x}_t)\nabla^2_{\mathbf{xy}} g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| \\
&\leq \|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| \\
&\quad + \|\nabla^2_{\mathbf{xy}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\| \\
&\quad + \|\mathbf{v}_t^*(\mathbf{x}_t)\|\|\nabla^2_{\mathbf{xy}} g_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla^2_{\mathbf{xy}} g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| \\
&\leq \left(\ell_{f,1} + \frac{\ell_{g,2}\ell_{f,0}}{\mu_g}\right)\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\| + \ell_{g,1}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\| \\
&\leq M_f^2 \left(\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\| + \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|\right),
\end{aligned} \tag{39}$$

where

$$M_f := \sqrt{2}\max\left\{\ell_{f,1} + \frac{\ell_{g,2}\ell_{f,0}}{\mu_g}, \ell_{g,1}\right\}, \tag{40}$$

the third inequality is by Assumption 3.3, and the last inequality follows from (38).

Next, we establish (37b).

Since $\mathbf{d}_t^{\mathbf{v}*} := \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) + \nabla_{\mathbf{y}}^2 g_t\left(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)\right) \mathbf{v}_t^*(\mathbf{x}_t) = 0$, we have

$$\begin{aligned}
\|\mathbf{d}_t^{\mathbf{v}}\| &= \|\mathbf{d}_t^{\mathbf{v}} - \mathbf{d}_t^{\mathbf{v}*}\| \\
&= \|\mathbf{v}_t \nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) + \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t) \\
&\quad - \left(\mathbf{v}_t^*(\mathbf{x}_t) \nabla_{\mathbf{y}}^2 g_t\left(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)\right) + \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right) \| \\
&\leq \| \left(\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right) \mathbf{v}_t^*(\mathbf{x}_t)\| \\
&\quad + \|\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) \left(\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\right) \| \\
&\quad + \|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\|.
\end{aligned}$$

Then, from Assumption 3.3 and (38), we have

$$\begin{aligned}
\|\mathbf{d}_t^{\mathbf{v}}\| &\leq \ell_{g,2}\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|\|\mathbf{v}_t^*(\mathbf{x}_t)\| + \ell_{g,1}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\| + \ell_{f,1}\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\| \\
&\leq \left(\frac{\ell_{g,2}\ell_{f,0}}{\mu_g} + \ell_{f,1}\right) \|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\| + \ell_{g,1}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\| \\
&\leq M_{\mathbf{v}} \left(\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\| + \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|\right),
\end{aligned}$$

where

$$M_{\mathbf{v}} := \sqrt{2} \max \left\{ \frac{\ell_{g,2}\ell_{f,0}}{\mu_g} + \ell_{f,1}, \ell_{g,1} \right\}. \tag{41}$$

The proofs of Eqs. (37c)-(37e) follow from Tarzanagh et al. (2024, Lemma 17) by setting

$$\begin{aligned}
L_{\mathbf{y}} &:= \frac{\ell_{g,1}}{\mu_g}, \\
L_{\mathbf{v}} &:= \ell_{f,1} + \frac{\ell_{g,1}\ell_{f,1}}{\mu_g} + \frac{\ell_{f,0}}{\mu_g}\left(\ell_{g,2} + \frac{\ell_{g,1}\ell_{g,2}}{\mu_g}\right), \\
L_f &:= \ell_{f,1} + \frac{\ell_{g,1}(\ell_{f,1} + M_f)}{\mu_g} + \frac{\ell_{f,0}}{\mu_g}\left(\ell_{g,2} + \frac{\ell_{g,1}\ell_{g,2}}{\mu_g}\right),
\end{aligned} \tag{42}$$

where the other constants are defined in Assumption 3.3. □

Lemma C.2. *Suppose Assumptions 3.5, B3. and C1. hold. Let $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ be generated according to Algorithm 1. For e_t^g defined as*

$$e_t^g := \mathbf{d}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t), \tag{43}$$

we have:

$$\begin{aligned}
\mathbb{E}\|e_{t+1}^g\|^2 &\leq (1-\gamma_{t+1})^2(1 + 48\ell_{g,1}^2\beta_t^2)\mathbb{E}\|e_t^g\|^2 + 2\gamma_{t+1}^2\frac{\sigma_{g_{\mathbf{y}}}^2}{\bar{b}} + 24(1-\gamma_{t+1})^2\ell_{g,1}^2\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&\quad + 6(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1})\|^2 \\
&\quad + 48(1-\gamma_{t+1})^2\ell_{g,1}^2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2.
\end{aligned} \tag{44}$$

Proof. From Algorithm 1, we have

$$\mathbf{d}_{t+1}^{\mathbf{y}} = \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) + (1-\gamma_{t+1})(\mathbf{d}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})).$$

Then, we have

$$\begin{aligned}
\mathbb{E}\|e_{t+1}^g\|^2 &= \mathbb{E}\|\mathbf{d}_{t+1}^{\mathbf{y}} - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1})\|^2 \\
&= \mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) + (1-\gamma_{t+1})(\mathbf{d}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1})\|^2 \\
&= \mathbb{E}\|(1-\gamma_{t+1})e_t^g + (\nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1})) \\
&\quad - (1-\gamma_{t+1})\left(\nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\right) - \nabla_{\mathbf{y}} g_t(\mathbf{z}_t)\|^2,
\end{aligned}$$

which implies that

$$\begin{aligned}\mathbb{E}\|e^g_{t+1}\|^2 &= (1-\gamma_{t+1})^2\mathbb{E}\|e^g_t\|^2 + \mathbb{E}\|(\nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}))\\
&- (1-\gamma_{t+1})\left(\nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\right) - \nabla_{\mathbf{y}} g_t(\mathbf{z}_t)\|^2\\
&\le (1-\gamma_{t+1})^2\mathbb{E}\|e^g_t\|^2 + 2\gamma_{t+1}^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1})\|^2\\
&+ 2(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\\
&- \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1}) + \nabla_{\mathbf{y}} g_t(\mathbf{z}_t)\|^2\\
&\le (1-\gamma_{t+1})^2\mathbb{E}\|e^g_t\|^2 + 2\gamma_{t+1}^2\frac{\sigma^2_{g_{\mathbf{y}}}}{\bar{b}}\\
&+ 2(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\\
&- \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1}) + \nabla_{\mathbf{y}} g_t(\mathbf{z}_t)\|^2,\end{aligned}$$

where the second inequality follows from Cauchy–Schwartz inequality and Assumption 3.5. Moreover, from Cauchy–Schwartz inequality, we have

$$\begin{aligned}\mathbb{E}\|e^g_{t+1}\|^2 &\le (1-\gamma_{t+1})^2\mathbb{E}\|e^g_t\|^2 + 2\gamma_{t+1}^2\frac{\sigma^2_{g_{\mathbf{y}}}}{\bar{b}}\\
&+ 6(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{z}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{z}_{t+1})\|^2\\
&+ 6(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1})\|^2\\
&+ 6(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\|^2.\end{aligned}$$

From Assumption B3., we have

$$\begin{aligned}&\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} g_t(\mathbf{z}_t)\|^2\\
&\le 2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_{t+1},\mathbf{y}_t)\|^2 + 2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_{t+1},\mathbf{y}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\\
&\le 2\ell_{g,1}^2\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\ell_{g,1}^2\mathbb{E}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2\\
&= 2\ell_{g,1}^2\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\ell_{g,1}^2\beta_t^2\mathbb{E}\|\mathbf{d}^{\mathbf{y}}_t\|^2,\end{aligned}$$

and

$$\begin{aligned}&\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\|^2\\
&\le 2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_t;\bar{\mathcal{B}}_{t+1})\|^2\\
&+ 2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_t;\bar{\mathcal{B}}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t;\bar{\mathcal{B}}_{t+1})\|^2\\
&\le 2\ell_{g,1}^2\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\ell_{g,1}^2\mathbb{E}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2\\
&= 2\ell_{g,1}^2\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\ell_{g,1}^2\beta_t^2\mathbb{E}\|\mathbf{d}^{\mathbf{y}}_t\|^2.\end{aligned}$$

From the two inequalities above, we have

$$\begin{aligned}\mathbb{E}\|e^g_{t+1}\|^2 &\le (1-\gamma_{t+1})^2\mathbb{E}\|e^g_t\|^2 + 2\gamma_{t+1}^2\frac{\sigma^2_{g_{\mathbf{y}}}}{\bar{b}}\\
&+ 6(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1})\|^2\\
&+ 24(1-\gamma_{t+1})^2\ell_{g,1}^2\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \beta_t^2\mathbb{E}\|\mathbf{d}^{\mathbf{y}}_t\|^2\right).\end{aligned}$$

Since $e_t^g := \mathbf{d}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)$, we have

$$\begin{aligned}
\mathbb{E}\|e_{t+1}^g\|^2 &\leq (1-\gamma_{t+1})^2\mathbb{E}\|e_t^g\|^2 + 2\gamma_{t+1}^2\frac{\sigma_{g_{\mathbf{y}}}^2}{\bar{b}} + 24(1-\gamma_{t+1})^2\ell_{g,1}^2\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \\
&+ 6(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1})\|^2 \\
&+ 48(1-\gamma_{t+1})^2\ell_{g,1}^2\beta_t^2\mathbb{E}\|e_t^g\|^2 + 48(1-\gamma_{t+1})^2\ell_{g,1}^2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2 \\
&\leq (1-\gamma_{t+1})^2(1+48\ell_{g,1}^2\beta_t^2)\mathbb{E}\|e_t^g\|^2 + 2\gamma_{t+1}^2\frac{\sigma_{g_{\mathbf{y}}}^2}{\bar{b}} + 24(1-\gamma_{t+1})^2\ell_{g,1}^2\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \\
&+ 6(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1})\|^2 \\
&+ 48(1-\gamma_{t+1})^2\ell_{g,1}^2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2.
\end{aligned}$$

□

Lemma C.3. *Suppose Assumptions 3.2, and B3. hold. Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$ generated by Algorithm 1, we have*

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] &\leq (1+a)\left(1 - 2\beta_t\frac{\mu_g\ell_{g,1}}{\mu_g+\ell_{g,1}}\right)\mathbb{E}\left[\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] \\
&+ \left(-(1+a)\left(\frac{2\beta_t}{\mu_g+\ell_{g,1}} - \beta_t^2\right)\right)\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right] \\
&+ (1+\frac{1}{a})\beta_t^2\mathbb{E}\left[\|e_t^g\|^2\right],
\end{aligned}$$

where e_t^g defined in (43) and $a > 0$ is a constant.

Proof. From Lemma B.5, we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] &= \mathbb{E}\left[\|\mathbf{y}_t - \beta_t\mathbf{d}_t^{\mathbf{y}} - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] \\
&\leq (1+a)\mathbb{E}\left[\|\mathbf{y}_t - \beta_t\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] \\
&+ (1+\frac{1}{a})\beta_t^2\mathbb{E}\left[\|\mathbf{d}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right].
\end{aligned} \tag{45}$$

Next, we will bound the first term on the RHS of (45).
We have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_t - \beta_t\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] &= \mathbb{E}\left[\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] + \beta_t^2\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right] \\
&- 2\beta_t\mathbb{E}\left[\langle\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t), \mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\rangle\right] \\
&\leq \left(1 - 2\beta_t\frac{\mu_g\ell_{g,1}}{\mu_g+\ell_{g,1}}\right)\mathbb{E}\left[\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] \\
&- \left(\frac{2\beta_t}{\mu_g+\ell_{g,1}} - \beta_t^2\right)\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right],
\end{aligned} \tag{46}$$

where the inequality results from the strong convexity of g_t by Assumption 3.2, which implies

$$\langle\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t), \mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\rangle \geq \frac{\mu_g\ell_{g,1}}{\mu_g+\ell_{g,1}}\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 + \frac{1}{\mu_g+\ell_{g,1}}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2.$$

Substituting (46) into (45), gives the desired result.

□

To simplify the notation in the analysis, we introduce the definitions

$$\theta_t^{\mathbf{y}} := \|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2, \quad \text{and} \quad \theta_t^{\mathbf{v}} := \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2. \tag{47}$$

Lemma C.4. *Suppose Assumptions 3.2, and B2., B3. hold. Let $\theta_t^{\mathbf{y}}$ be defined as in (47). Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$ generated by Algorithm 1, the following bound is guaranteed:*

$$
\begin{aligned}
&\sum_{t=1}^{T} \left(\mathbb{E}[\theta_{t+1}^{\mathbf{y}}] - \mathbb{E}[\theta_t^{\mathbf{y}}]\right) \\
&\leq -\frac{L_{\mu_g}}{2} \sum_{t=1}^{T} \beta_t \mathbb{E}[\theta_t^{\mathbf{y}}] + \frac{2}{L_{\mu_g}} \sum_{t=1}^{T} \beta_t \mathbb{E}\left[\|e_t^g\|^2\right] + \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}} \sum_{t=1}^{T} \frac{1}{\beta_t} \mathbb{E}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\
&+ \frac{4}{L_{\mu_g}} \sum_{t=2}^{T} \frac{1}{\beta_t} \sup_{\mathbf{x}\in\mathcal{X}} \mathbb{E}\|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2 + \sum_{t=1}^{T} \left(-\frac{2\beta_t}{\mu_g + \ell_{g,1}} + \beta_t^2\right) \mathbb{E}\left[\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2\right],
\end{aligned} \tag{48}
$$

where $L_{\mu_g} = \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}$, $L_{\mathbf{y}} = \frac{\ell_{g,1}}{\mu_g}$ is defined as in (42); $H_{2,T}$ is defined in (10). Moreover, e_t^g is defined in (43).

Proof. From Lemma B.5, we have for any $\acute{c} > 0$

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_{t+1} - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\|^2\right] &= \mathbb{E}\left[\|\mathbf{y}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t) + \mathbf{y}_t^*(\mathbf{x}_t) - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\|^2\right] \\
&\leq (1+\acute{c})\, \mathbb{E}\left[\|\mathbf{y}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] \\
&+ \left(1 + \frac{1}{\acute{c}}\right) \mathbb{E}\left[\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right].
\end{aligned} \tag{49}
$$

From Lemma C.3, we have for any $a > 0$

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] \leq (1+a)&\left(1 - 2\beta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right) \mathbb{E}\left[\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] \\
&+ \left(-(1+a)\left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right) \mathbb{E}\left[\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2\right] \\
&+ \left(1 + \frac{1}{a}\right) \beta_t^2 \mathbb{E}\left[\|e_t^g\|^2\right].
\end{aligned} \tag{50}
$$

Substituting (50) into (49), we get

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{y}_{t+1} - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\|^2\right] \\
&\leq (1+\acute{c})(1+a)\left(1 - 2\beta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right) \mathbb{E}\left[\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right] \\
&+ \left(-(1+\acute{c})(1+a)\left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right) \mathbb{E}\left[\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2\right] \\
&+ (1+\acute{c})(1 + \frac{1}{a})\beta_t^2 \mathbb{E}\left[\|e_t^g\|^2\right] \\
&+ \left(1 + \frac{1}{\acute{c}}\right) \mathbb{E}\left[\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2\right].
\end{aligned} \tag{51}
$$

Choose $\acute{c} = \frac{\beta_t L_{\mu_g}/2}{1 - \beta_t L_{\mu_g}}$ and $a = \frac{\beta_t L_{\mu_g}}{1 - 2\beta_t L_{\mu_g}}$. Then, the following equations and inequalities are satisfied.

$$
\begin{aligned}
&(1+\acute{c})(1+a)\left(1 - 2\beta_t L_{\mu_g}\right) = 1 - \frac{\beta_t L_{\mu_g}}{2}, \\
&(1+a)\left(1 - 2\beta_t L_{\mu_g}\right) = 1 - \beta_t L_{\mu_g}, \\
&(1+\acute{c})\left(1 - \beta_t L_{\mu_g}\right) = 1 - \frac{\beta_t L_{\mu_g}}{2}, \\
&1 + \frac{1}{a} \leq \frac{1}{\beta_t L_{\mu_g}}, \quad 1 + \frac{1}{\acute{c}} \leq \frac{2}{\beta_t L_{\mu_g}},
\end{aligned} \tag{52}
$$

where $L_{\mu_g}=\frac{\mu_g\ell_{g,1}}{\mu_g+\ell_{g,1}}$. Based on (51) and (52), we get

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{y}_{t+1}-\mathbf{y}^*_{t+1}(\mathbf{x}_{t+1})\|^2\right]-\mathbb{E}\left[\|\mathbf{y}_t-\mathbf{y}^*_t(\mathbf{x}_t)\|^2\right]\\
&\le -\frac{\beta_t L_{\mu_g}}{2}\mathbb{E}\left[\|\mathbf{y}_t-\mathbf{y}^*_t(\mathbf{x}_t)\|^2\right]+\left(-\left(\frac{2\beta_t}{\mu_g+\ell_{g,1}}-\beta_t^2\right)\right)\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right]\\
&+\frac{2}{\beta_t L_{\mu_g}}\beta_t^2\mathbb{E}\left[\|e^g_t\|^2\right]+\frac{2}{\beta_t L_{\mu_g}}\mathbb{E}\left[\|\mathbf{y}^*_{t+1}(\mathbf{x}_{t+1})-\mathbf{y}^*_t(\mathbf{x}_t)\|^2\right]. \qquad (53)
\end{aligned}$$

Next, we upper-bound the last term of the above inequality.

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{y}^*_{t+1}(\mathbf{x}_{t+1})-\mathbf{y}^*_t(\mathbf{x}_t)\|^2\right]\\
&\le 2\left(\mathbb{E}\left[\|\mathbf{y}^*_{t+1}(\mathbf{x}_{t+1})-\mathbf{y}^*_{t+1}(\mathbf{x}_t)\|^2\right]+\mathbb{E}\left[\|\mathbf{y}^*_{t+1}(\mathbf{x}_t)-\mathbf{y}^*_t(\mathbf{x}_t)\|^2\right]\right)\\
&\le 2\left(L^2_{\mathbf{y}}\mathbb{E}\left[\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2+\|\mathbf{y}^*_{t+1}(\mathbf{x}_t)-\mathbf{y}^*_t(\mathbf{x}_t)\|^2\right]\right), \qquad (54)
\end{aligned}$$

where the second inequality is by Lemma D.2.

Substituting (54) into (53) and summing over $t\in[T]$, give the desired result.

□

C.3. Bounds on the Linear System Solution

Lemma C.5. *Suppose Assumptions B2., B3., B4., C2. and C4. hold. Let $\{(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\}_{t=1}^T$ be generated according to Algorithm 1. For $e^{\mathbf{v}}_{t+1}$ defined as*

$$e^{\mathbf{v}}_t:=\mathbf{d}^{\mathbf{v}}_t-\nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t),\quad \text{where}\quad \nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t):=\nabla^2_{\mathbf{y}} g_t\left(\mathbf{x}_t,\mathbf{y}_t\right)\mathbf{v}_t+\nabla_{\mathbf{y}} f_t(\mathbf{x}_t,\mathbf{y}_t). \qquad (55)$$

we have:

$$\begin{aligned}
\mathbb{E}\|e^{\mathbf{v}}_{t+1}\|^2 &\le (1-\lambda_{t+1})^2(1+72\ell^2_{g,1}\delta_t^2)\mathbb{E}\|e^{\mathbf{v}}_t\|^2+4\lambda^2_{t+1}(\frac{\sigma^2_{g_{\mathbf{y}\mathbf{y}}}}{\bar{b}}p^2+\frac{\sigma^2_{f_{\mathbf{y}}}}{b})\\
&+12p^2(1-\lambda_{t+1})^2\mathbb{E}\|\nabla^2_{\mathbf{y}} g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)-\nabla^2_{\mathbf{y}} g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\|^2\\
&+12(1-\lambda_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1})-\nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2\\
&+72(1-\lambda_{t+1})^2(\ell^2_{g,2}p^2+\ell^2_{f,1})\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2+2\beta_t^2\mathbb{E}\|e^g_t\|^2+2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right)\\
&+72(1-\lambda_{t+1})^2\ell^4_{g,1}\delta_t^2\theta^{\mathbf{v}}_t, \qquad (56)
\end{aligned}$$

for all $t\in[T]$ and $\theta^{\mathbf{v}}_t$ is defined in (47).

Proof. Note that

$$e^{\mathbf{v}}_{t+1}:=\mathbf{d}^{\mathbf{v}}_{t+1}-\nabla P_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1},\mathbf{v}_{t+1}),$$

where

$$\nabla P_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1},\mathbf{v}_{t+1}):=\nabla^2_{\mathbf{y}} g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\mathbf{v}_{t+1}+\nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}).$$

From Algorithm 1, we have

$$\mathbf{d}^{\mathbf{v}}_{t+1}=\mathbf{d}^{\mathbf{v}\mathbf{v}}_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\mathcal{B}_{t+1}\right)+(1-\lambda_{t+1})(\mathbf{d}^{\mathbf{v}}_t-\mathbf{d}^{\mathbf{v}\mathbf{v}}_{t+1}(\mathbf{x}_t,\mathbf{y}_t;\mathcal{B}_{t+1})).$$

Let $\mathbf{u}=[\mathbf{x};\mathbf{y};\mathbf{v}]$. Then, we have

$$\begin{aligned}
\mathbb{E}\|e^{\mathbf{v}}_{t+1}\|^2&=\mathbb{E}\|\mathbf{d}^{\mathbf{v}}_{t+1}-\nabla P_{t+1}(\mathbf{u}_{t+1})\|^2\\
&=\mathbb{E}\|\nabla P_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1})+(1-\lambda_{t+1})(\mathbf{d}^{\mathbf{v}}_t-\nabla P_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1}))-\nabla P_{t+1}(\mathbf{u}_{t+1})\|^2\\
&=\mathbb{E}\|(1-\lambda_{t+1})e^{\mathbf{v}}_t+\nabla P_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1})-\nabla P_{t+1}(\mathbf{u}_{t+1})\\
&\quad-(1-\lambda_{t+1})\left(\nabla P_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1})-\nabla P_t(\mathbf{u}_t)\right)\|^2,
\end{aligned}$$

which implies that

$$
\begin{aligned}
&\mathbb{E}\|e^{\mathbf{v}}_{t+1}\|^2 \\
&= (1-\lambda_{t+1})^2\mathbb{E}\|e^{\mathbf{v}}_t\|^2 + \mathbb{E}\|\lambda_{t+1}\left(\nabla P_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_{t+1})\right) \\
&- (1-\lambda_{t+1})\left(\nabla P_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) + \nabla P_{t+1}(\mathbf{u}_{t+1}) - \nabla P_t(\mathbf{u}_t)\right)\|^2 \\
&\leq (1-\lambda_{t+1})^2\mathbb{E}\|e^{\mathbf{v}}_t\|^2 + 2\lambda_{t+1}^2\mathbb{E}\|\nabla P_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_{t+1})\|^2 \\
&+ 2(1-\lambda_{t+1})^2\mathbb{E}\|\nabla P_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1}) + \nabla P_t(\mathbf{u}_t)\|^2,
\end{aligned}
$$

where the inequality follows from Cauchy–Schwartz inequality.
For the first term, from Assumptions C2. and C4., we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla P_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_{t+1})\|^2 \\
&= \mathbb{E}\|\left(\nabla^2_{\mathbf{y}} g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}\right) - \nabla^2_{\mathbf{y}} g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\right)\mathbf{v}_{t+1} \\
&+ \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\mathcal{B}_{t+1}) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 \\
&\leq 2\mathbb{E}\|\left(\nabla^2_{\mathbf{y}} g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}\right) - \nabla^2_{\mathbf{y}} g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\right)\mathbf{v}_{t+1}\|^2 \\
&+ 2\mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\mathcal{B}_{t+1}) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 \\
&\leq 2(\frac{\sigma^2_{g_{\mathbf{yy}}}}{\bar{b}}p^2 + \frac{\sigma^2_{f_{\mathbf{y}}}}{b}),
\end{aligned}
$$

where the last inequality follows from (8).

Then, from the above inequality and $\|a+b+c\|^2 \leq 3(\|a\|^2+\|b\|^2+\|c\|^2)$, we have

$$
\begin{aligned}
\mathbb{E}\|e^{\mathbf{v}}_{t+1}\|^2 \leq\ & (1-\lambda_{t+1})^2\mathbb{E}\|e^{\mathbf{v}}_t\|^2 + 4\lambda_{t+1}^2(\frac{\sigma^2_{g_{\mathbf{yy}}}}{\bar{b}}p^2 + \frac{\sigma^2_{f_{\mathbf{y}}}}{b}) \\
&+ 6(1-\lambda_{t+1})^2\mathbb{E}\|\nabla P_t(\mathbf{u}_t) - \nabla P_t(\mathbf{u}_{t+1})\|^2 \\
&+ 6(1-\lambda_{t+1})^2\mathbb{E}\|\nabla P_t(\mathbf{u}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_{t+1})\|^2 \\
&+ 6(1-\lambda_{t+1})^2\mathbb{E}\|\nabla P_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1})\|^2.
\end{aligned} \tag{57}
$$

Moreover, from $\|a+b+c\|^2 \leq 3(\|a\|^2+\|b\|^2+\|c\|^2)$, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla P_t(\mathbf{u}_{t+1}) - \nabla P_t(\mathbf{u}_t)\|^2 \\
&\leq 3\mathbb{E}\|\nabla P_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1},\mathbf{v}_{t+1}) - \nabla P_t(\mathbf{x}_t,\mathbf{y}_{t+1},\mathbf{v}_{t+1})\|^2 \\
&+ 3\mathbb{E}\|\nabla P_t(\mathbf{x}_t,\mathbf{y}_{t+1},\mathbf{v}_{t+1}) - \nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_{t+1})\|^2 \\
&+ 3\mathbb{E}\|\nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_{t+1}) - \nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\|^2 \\
&\leq 3\mathbb{E}\|(\nabla^2_{\mathbf{y}} g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right) - \nabla^2_{\mathbf{y}} g_t\left(\mathbf{x}_t,\mathbf{y}_{t+1}\right))\mathbf{v}_{t+1} + \nabla_{\mathbf{y}} f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t,\mathbf{y}_{t+1})\|^2 \\
&+ 3\mathbb{E}\|(\nabla^2_{\mathbf{y}} g_t\left(\mathbf{x}_t,\mathbf{y}_{t+1}\right) - \nabla^2_{\mathbf{y}} g_t\left(\mathbf{x}_t,\mathbf{y}_t\right))\mathbf{v}_{t+1} + \nabla_{\mathbf{y}} f_t(\mathbf{x}_t,\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t,\mathbf{y}_t)\|^2 \\
&+ 3\mathbb{E}\|\nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_{t+1}) - \nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\|^2 \\
&\leq 6(\ell_{g,2}^2\mathbb{E}\|\mathbf{v}_{t+1}\|^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \mathbb{E}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2\right) + 3\ell_{g,1}^2\mathbb{E}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 \\
&\leq 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \beta_t^2\mathbb{E}\|\mathbf{d}^{\mathbf{y}}_t\|^2\right) + 3\ell_{g,1}^2\delta_t^2\mathbb{E}\|\mathbf{d}^{\mathbf{v}}_t\|^2 \\
&\leq 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\beta_t^2\mathbb{E}\|e^g_t\|^2 + 2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right) \\
&+ 6\ell_{g,1}^2\delta_t^2(\mathbb{E}\|e^{\mathbf{v}}_t\|^2 + \mathbb{E}\|\nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\|^2) \\
&\leq 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\beta_t^2\mathbb{E}\|e^g_t\|^2 + 2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right) \\
&+ 6\ell_{g,1}^2\delta_t^2\left(\mathbb{E}\|e^{\mathbf{v}}_t\|^2 + \ell_{g,1}^2\mathbb{E}\|\mathbf{v}_t - \mathbf{v}^*_t(\mathbf{x}_t)\|^2\right),
\end{aligned} \tag{58}
$$

where the third inequality follows from Assumptions B2., B3. and B4.; the last inequality follows from (62).

Similarly, we have

$$\begin{aligned}&\mathbb{E}\|\nabla P_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1})-\nabla P_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1})\|^2\\&\leq 6(\ell_{g,2}^2p^2+\ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2+2\beta_t^2\mathbb{E}\|e_t^g\|^2+2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right)\\&+6\ell_{g,1}^2\delta_t^2\left(\mathbb{E}\|e_t^{\mathbf{v}}\|^2+\ell_{g,1}^2\mathbb{E}\|\mathbf{v}_t-\mathbf{v}_t^*(\mathbf{x}_t)\|^2\right).\end{aligned}\tag{59}$$

Substituting (59) and (58) into (57), we have

$$\begin{aligned}\mathbb{E}\|e_{t+1}^{\mathbf{v}}\|^2&\leq(1-\lambda_{t+1})^2(1+72\ell_{g,1}^2\delta_t^2)\mathbb{E}\|e_t^{\mathbf{v}}\|^2+4\lambda_{t+1}^2(\frac{\sigma_{g_{\mathbf{yy}}}^2}{b}p^2+\frac{\sigma_{f_{\mathbf{y}}}^2}{b})\\&+6(1-\lambda_{t+1})^2\mathbb{E}\|\nabla P_t(\mathbf{u}_{t+1})-\nabla P_{t+1}(\mathbf{u}_{t+1})\|^2\\&+72(1-\lambda_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2+2\beta_t^2\mathbb{E}\|e_t^g\|^2+2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right)\\&+72(1-\lambda_{t+1})^2\ell_{g,1}^4\delta_t^2\mathbb{E}\|\mathbf{v}_t-\mathbf{v}_t^*(\mathbf{x}_t)\|^2.\end{aligned}$$

From $\|a+b\|^2\leq 2\|a\|^2+2\|b\|^2$ and (8), we have

$$\begin{aligned}\mathbb{E}\|\nabla P_t(\mathbf{u}_{t+1})-\nabla P_{t+1}(\mathbf{u}_{t+1})\|^2&=\mathbb{E}\|\nabla_{\mathbf{y}}^2g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\mathbf{v}_{t+1}-\nabla_{\mathbf{y}}^2g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\mathbf{v}_{t+1}\\&+\nabla_{\mathbf{y}}f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1})-\nabla_{\mathbf{y}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2\\&\leq 2\mathbb{E}\|\left(\nabla_{\mathbf{y}}^2g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)-\nabla_{\mathbf{y}}^2g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\right)\mathbf{v}_{t+1}\|^2\\&+2\mathbb{E}\|\nabla_{\mathbf{y}}f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1})-\nabla_{\mathbf{y}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2\\&\leq 2\mathbb{E}\|\nabla_{\mathbf{y}}^2g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)-\nabla_{\mathbf{y}}^2g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\|^2p^2\\&+2\mathbb{E}\|\nabla_{\mathbf{y}}f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1})-\nabla_{\mathbf{y}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2.\end{aligned}$$

This completes the proof. □

As demonstrated in Lemma C.5, the gradient estimation error $e_{t+1}^{\mathbf{v}}$ for the linear system consists of four key components: (1) an iteratively refined error term $(1-\lambda_{t+1})^2(1+72\ell_{g,1}^2\delta_t^2)\mathbb{E}\|e_t^{\mathbf{v}}\|^2$, which depends on the stepsize δ_t; (2) the error arising from the variation in the Hessian of the lower-level objectiv; (3) the error resulting from the variation in the gradient of the upper-level objective, and (4) an approximation error term of order $\mathcal{O}(\delta_t^2\theta_t^{\mathbf{v}})$ associated with solving the linear system.

Lemma C.6. *Suppose Assumptions 3.2 and 3.3 hold. Then, for the sequence $\{(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 1, we have*

$$\begin{aligned}\mathbb{E}\|\mathbf{v}_{t+1}-\mathbf{v}_t^*(\mathbf{x}_t)\|^2\leq(1+\acute{c})&\left(1-2\delta_t\frac{(\ell_{g,1}+\ell_{g,1}^3)\mu_g}{\mu_g+\ell_{g,1}}+\delta_t^2\ell_{g,1}^2\right)\mathbb{E}\|\mathbf{v}_t-\mathbf{v}_t^*(\mathbf{x}_t)\|^2\\&+(1+\frac{1}{\acute{c}})\delta_t^2\mathbb{E}\|e_t^{\mathbf{v}}\|^2,\end{aligned}$$

where $e_t^{\mathbf{v}}$ defined in (55) and for any $\acute{c}>0$.

Proof. From the update rules in Algorithm 1, we have the following:

$$\begin{aligned}\mathbb{E}\|\mathbf{v}_{t+1}-\mathbf{v}_t^*(\mathbf{x}_t)\|^2&=\mathbb{E}\|\mathbf{v}_t-\delta_t\mathbf{d}_t^{\mathbf{v}}-\mathbf{v}_t^*(\mathbf{x}_t)\|^2\\&\leq(1+\acute{c})\mathbb{E}\|\mathbf{v}_t-\delta_t\nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)-\mathbf{v}_t^*(\mathbf{x}_t)\|^2\\&+(1+\frac{1}{\acute{c}})\delta_t^2\mathbb{E}\|\mathbf{d}_t^{\mathbf{v}}-\nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\|^2,\end{aligned}\tag{60}$$

where $\nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t):=\nabla_{\mathbf{y}}^2g_t\left(\mathbf{x}_t,\mathbf{y}_t\right)\mathbf{v}_t+\nabla_{\mathbf{y}}f_t(\mathbf{x}_t,\mathbf{y}_t)$.

For the first term of the above eq. (60), we have

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{v}_t - \delta_t \nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t) - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \\
&= \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 - 2\delta_t \mathbb{E}\langle \mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t), \nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\rangle + \delta_t^2 \mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2 \\
&\leq \left(1 - 2\delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right) \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 - (2\delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}} - \delta_t^2) \mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2 \\
&\leq \left(1 - 2\delta_t \frac{(\ell_{g,1} + \ell_{g,1}^3)\mu_g}{\mu_g + \ell_{g,1}} + \delta_t^2 \ell_{g,1}^2\right) \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2,
\end{aligned} \tag{61}
$$

where the first inequality follows from the strong convexity of P_t function (in eq. (4b)) that

$$
\mathbb{E}\langle \mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t), \nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\rangle \geq \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}} \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 + \frac{1}{\mu_g + \ell_{g,1}} \mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2.
$$

The second inequality is derived from the following inequality.

$$
\begin{aligned}
\mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2 &= \mathbb{E}\|\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)\, \mathbf{v}_t + \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&= \mathbb{E}\|\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)\, (\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t))\|^2 \leq \ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2.
\end{aligned} \tag{62}
$$

Combining (60) and (61), we get the desired result. □

Lemma C.7. *Suppose Assumptions 3.2 and 3.3 hold. Then, we have*

$$
\left\|\mathbf{v}_t^*(\mathbf{x}_t) - \mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2 \leq 2\frac{\nu^2}{\mu_g^2} \left(\left\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right),
$$

where $\nu := \ell_{f,1} + \frac{\ell_{g,2}\ell_{f,0}}{\mu_g}$, *and* $\mathbf{v}_t^*(\mathbf{x})$ *is a solution of Subproblem* (4b).

Proof. Based on (4b), we have that

$$
\begin{aligned}
&\left\|\mathbf{v}_t^*(\mathbf{x}_t) - \mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2 \\
&= \|\left(\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)^{-1} \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \\
&- \left(\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\right)^{-1} \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\|^2 \\
&\leq 2\left\|\left(\left(\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)^{-1} - \left(\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\right)^{-1}\right) \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right\|^2 \qquad (63a) \\
&+ 2\left\|\left(\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\right)^{-1} \left(\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\right)\right\|^2. \qquad (63b)
\end{aligned}
$$

In the following steps, we bound the terms (63a) and (63b), respectively.

For (63a), we have:

$$
\begin{aligned}
&\left\|\left(\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)^{-1} - \left(\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\right)^{-1}\right\|^2 \\
&= \|\left(\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)^{-1} (\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})) \\
&- \nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \left(\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\right)^{-1}\|^2 \\
&\leq \frac{1}{\mu_g^2} \left\|\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - \nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\right\|^2 \\
&\leq \frac{\ell_{g,2}}{\mu_g^2} \left\|(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - (\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\right\|^2 \\
&\leq \frac{\ell_{g,2}}{\mu_g^2} \left(\|\mathbf{y}_t^*(\mathbf{x}_t) - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\|^2 + \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2\right),
\end{aligned} \tag{64}
$$

where the first equality holds since for any invertible matrix $\mathbf{A}$ and $\mathbf{B}$ we have $\|\mathbf{A}^{-1} - \mathbf{B}^{-1}\| = \|\mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}\|$, and the second inequality is obtained from Assumption 3.3.

Thus, from (64) and Assumption 3.3, we get

$$\text{(63a)} \leq \frac{\ell_{f,0}\ell_{g,2}}{\mu_g^2}\left(\left\|\mathbf{y}_t^*(\mathbf{x}_t) - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2 + \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2\right). \tag{65}$$

For (63b), we have

$$\begin{aligned}
\text{(63b)} &\leq \frac{1}{\mu_g}\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\|^2 \\
&\leq \frac{\ell_{f,1}}{\mu_g}\|(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - (\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\|^2 \\
&\leq \frac{\ell_{f,1}}{\mu_g}\left(\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right).
\end{aligned} \tag{66}$$

Combining (65) and (66), we have

$$\left\|\mathbf{v}_t^*(\mathbf{x}_t) - \mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2 \leq \frac{1}{\mu_g}\left(\frac{\ell_{f,0}\ell_{g,2}}{\mu_g} + \ell_{f,1}\right)\left(\left\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2\right).$$

By raising both sides of the above inequality to the power 2 and using $(a+b)^2 \leq 2a^2 + 2b^2$, we complete the proof. □

Lemma C.8. *Suppose Assumptions 3.2 and 3.3 hold. Let $\theta_t^{\mathbf{v}}$ be defined in (47). Then, for any positive choice of step sizes as*

$$\delta_t \leq \frac{L_{\mu_g}}{\ell_{g,1}^2}, \quad \text{where} \quad L_{\mu_g} = \frac{(\ell_{g,1} + \ell_{g,1}^3)\mu_g}{(\mu_g + \ell_{g,1})},$$

for all $t \in [T]$, the sequence $\{\mathbf{v}_t\}_{t=1}^T$ generated by Algorithm 1 satisfy

$$\begin{aligned}
\sum_{t=1}^{T}\left(\mathbb{E}[\theta_{t+1}^{\mathbf{v}}] - \mathbb{E}[\theta_t^{\mathbf{v}}]\right) \leq &-\frac{\delta_t L_{\mu_g}}{4}\sum_{t=1}^{T}\mathbb{E}[\theta_t^{\mathbf{v}}] + \frac{4}{L_{\mu_g}}\delta_t\sum_{t=1}^{T}\mathbb{E}\|e_t^{\mathbf{v}}\|^2 \\
&+ \frac{16\nu^2}{L_{\mu_g}\mu_g^2\delta_t}\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2 \\
&+ \frac{8\nu^2}{L_{\mu_g}\mu_g^2\delta_t}(1 + 2L_{\mathbf{y}}^2)\sum_{t=1}^{T}\mathbb{E}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|^2,
\end{aligned} \tag{67}$$

where $e_t^{\mathbf{v}}$ is defined in (55).

Proof. By Lemma B.5, for any $a > 0$, we have

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{v}_{t+1} - \mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2 &= \mathbb{E}\left\|\mathbf{v}_{t+1} - \mathbf{v}_t^*(\mathbf{x}_t) + \mathbf{v}_t^*(\mathbf{x}_t) - \mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2 \\
&\leq (1+a)\,\mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \\
&+ \left(1 + \frac{1}{a}\right)\mathbb{E}\left\|\mathbf{v}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{v}_t^*(\mathbf{x}_t)\right\|^2.
\end{aligned} \tag{68}$$

From Lemma C.6, we have for any $\acute{c} > 0$:

$$\begin{aligned}
\mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \leq (1+\acute{c})&\left(1 - 2\delta_t\frac{(\ell_{g,1} + \ell_{g,1}^3)\mu_g}{\mu_g + \ell_{g,1}} + \delta_t^2\ell_{g,1}^2\right)\mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \\
&+ (1 + \frac{1}{\acute{c}})\delta_t^2\mathbb{E}\|e_t^{\mathbf{v}}\|^2.
\end{aligned} \tag{69}$$

Substituting (69) into (68), we get

$$\mathbb{E}\left\|\mathbf{v}_{t+1}-\mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2 \le (1+a)\,(1+\acute{c})\left(1-2\delta_t\frac{(\ell_{g,1}+\ell_{g,1}^3)\mu_g}{\mu_g+\ell_{g,1}}+\delta_t^2\ell_{g,1}^2\right)\mathbb{E}\|\mathbf{v}_t-\mathbf{v}_t^*(\mathbf{x}_t)\|^2$$
$$+(1+a)\,(1+\frac{1}{\acute{c}})\delta_t^2\mathbb{E}\|e_t^{\mathbf{v}}\|^2$$
$$+\left(1+\frac{1}{a}\right)\mathbb{E}\left\|\mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})-\mathbf{v}_t^*(\mathbf{x}_t)\right\|^2. \tag{70}$$

In the following, we provide a bound for the third term on the right-hand side of (70). To this end, we have from Lemma C.7:

$$\mathbb{E}\left\|\mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})-\mathbf{v}_t^*(\mathbf{x}_t)\right\|^2 \le 2\frac{\nu^2}{\mu_g^2}\left(\mathbb{E}\left\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2+\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right)$$
$$\le 2\frac{\nu^2}{\mu_g^2}\left(2\mathbb{E}\left\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})-\mathbf{y}_{t+1}^*(\mathbf{x}_t)\right\|^2\right.$$
$$\left.+2\mathbb{E}\left\|\mathbf{y}_{t+1}^*(\mathbf{x}_t)-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2+\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|^2\right)$$
$$\le 2\frac{\nu^2}{\mu_g^2}\left((1+2L_{\mathbf{y}}^2)\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|^2+2\mathbb{E}\left\|\mathbf{y}_{t+1}^*(\mathbf{x}_t)-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right),$$

where the last inequality follows from Lemma C.1.

Combining this result with (70) gives

$$\mathbb{E}\left\|\mathbf{v}_{t+1}-\mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2 \le (1+a)\,(1+\acute{c})\left(1-2\delta_t\frac{(\ell_{g,1}+\ell_{g,1}^3)\mu_g}{\mu_g+\ell_{g,1}}+\delta_t^2\ell_{g,1}^2\right)\mathbb{E}\|\mathbf{v}_t-\mathbf{v}_t^*(\mathbf{x}_t)\|^2$$
$$+(1+a)\,(1+\frac{1}{\acute{c}})\delta_t^2\mathbb{E}\|e_t^{\mathbf{v}}\|^2+4\left(1+\frac{1}{a}\right)\frac{\nu^2}{\mu_g^2}\mathbb{E}\left\|\mathbf{y}_{t+1}^*(\mathbf{x}_t)-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2$$
$$+2\left(1+\frac{1}{a}\right)\frac{\nu^2}{\mu_g^2}(1+2L_{\mathbf{y}}^2)\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|^2. \tag{71}$$

Let $L_{\mu_g}:=\frac{(\ell_{g,1}+\ell_{g,1}^3)\mu_g}{\mu_g+\ell_{g,1}}$, then we have

$$1-2\delta_t\frac{(\ell_{g,1}+\ell_{g,1}^3)\mu_g}{\mu_g+\ell_{g,1}}+\delta_t^2\ell_{g,1}^2 = 1-2\delta_tL_{\mu_g}+\delta_t^2\ell_{g,1}^2$$
$$\le 1-\delta_tL_{\mu_g}, \tag{72}$$

where the last inequality follows from $\delta_t\le\frac{L_{\mu_g}}{\ell_{g,1}^2}$.

Choose $a=\frac{\delta_tL_{\mu_g}/4}{1-\frac{\delta_tL_{\mu_g}}{2}}$ and $\acute{c}=\frac{\delta_tL_{\mu_g}/2}{1-\delta_tL_{\mu_g}}$. Then, from (72), we have

$$\begin{aligned}&(1+a)\,(1+\acute{c})\left(1-2\delta_t\frac{(\ell_{g,1}+\ell_{g,1}^3)\mu_g}{\mu_g+\ell_{g,1}}+\delta_t^2\ell_{g,1}^2\right)\\&\le(1+a)\,(1+\acute{c})\left(1-\delta_tL_{\mu_g}\right)=1-\frac{\delta_tL_{\mu_g}}{4},\\&(1+a)\left(1+\frac{1}{\acute{c}}\right)\le\frac{4}{\delta_tL_{\mu_g}},\\&1+\frac{1}{\acute{c}}\le\frac{2}{\delta_tL_{\mu_g}},\quad 1+\frac{1}{a}\le\frac{4}{\delta_tL_{\mu_g}},\end{aligned} \tag{73}$$

where $L_{\mu_g}:=\frac{(\ell_{g,1}+\ell_{g,1}^3)\mu_g}{\mu_g+\ell_{g,1}}$.

Thus, from (71) and (73) we have

$$\begin{aligned}\mathbb{E}\left\|\mathbf{v}_{t+1}-\mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2 \leq &\left(1-\frac{\delta_t L_{\mu_g}}{4}\right)\mathbb{E}\|\mathbf{v}_t-\mathbf{v}_t^*(\mathbf{x}_t)\|^2\\ &+\frac{4}{L_{\mu_g}}\delta_t\mathbb{E}\|e_t^{\mathbf{v}}\|^2+\frac{16\nu^2}{L_{\mu_g}\mu_g^2\delta_t}\mathbb{E}\left\|\mathbf{y}_{t+1}^*(\mathbf{x}_t)-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\ &+\frac{8\nu^2}{L_{\mu_g}\mu_g^2\delta_t}(1+2L_{\mathbf{y}}^2)\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|^2.\end{aligned}$$

Rearranging the terms and summing from $t=1$ to T, gives the desired result. □

C.4. Bounds on the Outer Objective and its Projected Gradient

Lemma C.9. *Suppose Assumptions B2., B3., C3. and C5. hold. Let $\{(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\}_{t=1}^T$ be generated according to Algorithm 1. For e_t^f defined as*

$$e_t^f:=\mathbf{d}_t^{\mathbf{x}}-\tilde{\mathbf{d}}_t\left(\mathbf{z}_t\right),\quad \text{where}\quad \tilde{\mathbf{d}}_t\left(\mathbf{z}_t\right)=\nabla_{\mathbf{x}}f_t(\mathbf{z}_t)+\nabla_{\mathbf{xy}}^2 g_t\left(\mathbf{z}_t\right)\mathbf{v}_t, \tag{74}$$

we have:

$$\begin{aligned}\mathbb{E}\|e_{t+1}^f\|^2 \leq &(1-\eta_{t+1})^2\mathbb{E}\|e_t^f\|^2+4\eta_{t+1}^2\left(\frac{\sigma_{g_{\mathbf{xy}}}^2}{\bar{b}}p^2+\frac{\sigma_{f_{\mathbf{x}}}^2}{b}\right)\\ &+12p^2(1-\eta_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{xy}}^2 g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)-\nabla_{\mathbf{xy}}^2 g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\|^2\\ &+12(1-\eta_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{x}}f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1})-\nabla_{\mathbf{x}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2\\ &+72(1-\eta_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2+2\beta_t^2\mathbb{E}\|e_t^g\|^2+2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right)\\ &+72\ell_{g,1}^2(1-\eta_{t+1})^2\delta_t^2\mathbb{E}\|e_t^{\mathbf{v}}\|^2+72(1-\eta_{t+1})^2\ell_{g,1}^4\delta_t^2\mathbb{E}[\theta_t^{\mathbf{v}}],\end{aligned} \tag{75}$$

for all $t\in[T]$, and $\theta_t^{\mathbf{v}}$ are defined in (47).

Proof. Note that

$$e_{t+1}^f=\mathbf{d}_{t+1}^{\mathbf{x}}-\tilde{\mathbf{d}}_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1},\mathbf{v}_{t+1}\right),$$

where

$$\tilde{\mathbf{d}}_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1},\mathbf{v}_{t+1}\right)=\nabla_{\mathbf{x}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})+\nabla_{\mathbf{xy}}^2 g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\mathbf{v}_{t+1}. \tag{76}$$

From Algorithm 1, we have

$$\mathbf{d}_{t+1}^{\mathbf{x}}=\mathbf{d}_{t+1}^{\mathbf{xx}}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\mathcal{B}_{t+1}\right)+(1-\eta_{t+1})(\mathbf{d}_{t+1}^{\mathbf{x}}-\mathbf{d}_{t+1}^{\mathbf{xx}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\mathcal{B}_{t+1})),$$

where $\mathbf{d}_{t+1}^{\mathbf{xx}}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\mathcal{B}_{t+1}\right)=\nabla_{\mathbf{x}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\mathcal{B}_{t+1})+\nabla_{\mathbf{xy}}^2 g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\mathcal{B}_{t+1}\right)\mathbf{v}_{t+1}$.
Let $\mathbf{u}=[\mathbf{x};\mathbf{y};\mathbf{v}]$. Then, we have

$$\begin{aligned}\mathbb{E}\|e_{t+1}^f\|^2&=\mathbb{E}\|\mathbf{d}_{t+1}^{\mathbf{x}}-\tilde{\mathbf{d}}_{t+1}\left(\mathbf{u}_{t+1}\right)\|^2\\ &=\mathbb{E}\|\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1})+(1-\eta_{t+1})(\mathbf{d}_t^{\mathbf{x}}-\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1}))-\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})\|^2\\ &=\mathbb{E}\|(1-\eta_{t+1})e_t^f+\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1})-\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})\\ &\quad-(1-\eta_{t+1})(\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1})-\tilde{\mathbf{d}}_t(\mathbf{u}_t))\|^2,\end{aligned}$$

which implies that

$$\begin{aligned}
&\mathbb{E}\|e_{t+1}^{f}\|^2 \\
&= (1-\eta_{t+1})^2\mathbb{E}\|e_t^f\|^2 + \mathbb{E}\|\eta_{t+1}(\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})) \\
&- (1-\eta_{t+1})(\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) + \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1}) - \tilde{\mathbf{d}}_t(\mathbf{u}_t))\|^2 \\
&\leq (1-\eta_{t+1})^2\mathbb{E}\|e_t^f\|^2 + 2\eta_{t+1}^2\mathbb{E}\|\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})\|^2 \\
&+ 2(1-\eta_{t+1})^2\mathbb{E}\|\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1}) + \tilde{\mathbf{d}}_t(\mathbf{u}_t)\|^2,
\end{aligned} \tag{77}$$

where the inequality follows from $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$.
Let us bound the second term in the right-hand side of (77). Based on (76), we have

$$\begin{aligned}
&\mathbb{E}\|\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})\|^2 \\
&= \mathbb{E}\|\left(\nabla^2_{\mathbf{xy}}g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}\right) - \nabla^2_{\mathbf{xy}}g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\right)\mathbf{v}_{t+1} \\
&+ \nabla_{\mathbf{x}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\mathcal{B}_{t+1}) - \nabla_{\mathbf{x}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 \\
&\leq 2\mathbb{E}\|\left(\nabla^2_{\mathbf{xy}}g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}\right) - \nabla^2_{\mathbf{xy}}g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\right)\mathbf{v}_{t+1}\|^2 \\
&+ 2\mathbb{E}\|\nabla_{\mathbf{x}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\mathcal{B}_{t+1}) - \nabla_{\mathbf{x}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 \\
&\leq 2(\frac{\sigma^2_{g_{\mathbf{xy}}}}{\bar{b}}p^2 + \frac{\sigma^2_{f_{\mathbf{x}}}}{b}),
\end{aligned}$$

where the first inequality is by and $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$; the second inequality follows from Assumptions C3., C5. and (8).
Substituting the above inequality into (77) and using $\|a+b+c\|^2 \leq 3(\|a\|^2+\|b\|^2+\|c\|^2)$, we obtain

$$\begin{aligned}
\mathbb{E}\|e_{t+1}^{f}\|^2 \leq\ & (1-\eta_{t+1})^2\mathbb{E}\|e_t^f\|^2 + 4\lambda_{t+1}^2(\frac{\sigma^2_{g_{\mathbf{xy}}}}{\bar{b}}p^2 + \frac{\sigma^2_{f_{\mathbf{x}}}}{b}) \\
&+ 6(1-\eta_{t+1})^2\mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{u}_t) - \tilde{\mathbf{d}}_t(\mathbf{u}_{t+1})\|^2 \\
&+ 6(1-\eta_{t+1})^2\mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{u}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})\|^2 \\
&+ 6(1-\eta_{t+1})^2\mathbb{E}\|\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1})\|^2.
\end{aligned} \tag{78}$$

Moreover, from $\|a+b+c\|^2 \leq 3(\|a\|^2+\|b\|^2+\|c\|^2)$, we have

$$\begin{aligned}
&\mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{u}_{t+1}) - \tilde{\mathbf{d}}_t(\mathbf{u}_t)\|^2 \\
&\leq 3\mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1},\mathbf{v}_{t+1}) - \tilde{\mathbf{d}}_t(\mathbf{x}_t,\mathbf{y}_{t+1},\mathbf{v}_{t+1})\|^2 \\
&+ 3\mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{x}_t,\mathbf{y}_{t+1},\mathbf{v}_{t+1}) - \tilde{\mathbf{d}}_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_{t+1})\|^2 \\
&+ 3\mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_{t+1}) - \tilde{\mathbf{d}}_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\|^2 \\
&\overset{(i)}{\leq} 3\mathbb{E}\|(\nabla^2_{\mathbf{xy}}g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right) - \nabla^2_{\mathbf{xy}}g_t\left(\mathbf{x}_t,\mathbf{y}_{t+1}\right))\mathbf{v}_{t+1} + \nabla_{\mathbf{x}}f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{x}}f_t(\mathbf{x}_t,\mathbf{y}_{t+1})\|^2 \\
&+ 3\mathbb{E}\|(\nabla^2_{\mathbf{xy}}g_t\left(\mathbf{x}_t,\mathbf{y}_{t+1}\right) - \nabla^2_{\mathbf{xy}}g_t\left(\mathbf{x}_t,\mathbf{y}_t\right))\mathbf{v}_{t+1} + \nabla_{\mathbf{x}}f_t(\mathbf{x}_t,\mathbf{y}_{t+1}) - \nabla_{\mathbf{x}}f_t(\mathbf{x}_t,\mathbf{y}_t)\|^2 \\
&+ 3\mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_{t+1}) - \tilde{\mathbf{d}}_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\|^2 \\
&\overset{(ii)}{\leq} 6(\ell_{g,2}^2\mathbb{E}\|\mathbf{v}_{t+1}\|^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \mathbb{E}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2\right) + 3\ell_{g,1}^2\mathbb{E}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 \\
&\overset{(iii)}{\leq} 6(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \beta_t^2\mathbb{E}\|\mathbf{d}_t^{\mathbf{y}}\|^2\right) + 3\ell_{g,1}^2\delta_t^2\mathbb{E}\|\mathbf{d}_t^{\mathbf{v}}\|^2 \\
&\overset{(iv)}{\leq} 6(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\beta_t^2\mathbb{E}\|e_t^g\|^2 + 2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right) \\
&+ 6\ell_{g,1}^2\delta_t^2(\mathbb{E}\|e_t^{\mathbf{v}}\|^2 + \mathbb{E}\|\nabla P_t(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\|^2) \\
&\overset{(vi)}{\leq} 6(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\beta_t^2\mathbb{E}\|e_t^g\|^2 + 2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right) \\
&+ 6\ell_{g,1}^2\delta_t^2\left(\mathbb{E}\|e_t^{\mathbf{v}}\|^2 + \ell_{g,1}^2\mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2\right),
\end{aligned} \tag{79}$$

where the (i) follows from (76); (ii) follows from Assumptions B2., B3. and B4.; (iii) follows from (8); (iv) follows from (43) and (55); (vi) follows from (62).
Similarly, we have

$$
\begin{aligned}
&\mathbb{E}\|\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1};\mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_t;\mathcal{B}_{t+1})\|^2 \\
&\le 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\beta_t^2\mathbb{E}\|e_t^g\|^2 + 2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right) \\
&+ 6\ell_{g,1}^2\delta_t^2\left(\mathbb{E}\|e_t^{\mathbf{v}}\|^2 + \ell_{g,1}^2\mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2\right). \qquad (80)
\end{aligned}
$$

Substituting (80) and (79) into (78), we have

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^f\|^2 \le\ & (1-\eta_{t+1})^2\mathbb{E}\|e_t^f\|^2 + 4\eta_{t+1}^2\left(\frac{\sigma_{g_{\mathbf{yy}}}^2}{\bar{b}}p^2 + \frac{\sigma_{f_{\mathbf{y}}}^2}{b}\right) \\
&+ 6(1-\eta_{t+1})^2\mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{u}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})\|^2 \\
&+ 72(1-\eta_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\beta_t^2\mathbb{E}\|e_t^g\|^2 + 2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right) \\
&+ 72\ell_{g,1}^2(1-\eta_{t+1})^2\delta_t^2\mathbb{E}\|e_t^{\mathbf{v}}\|^2 + 72(1-\eta_{t+1})^2\ell_{g,1}^4\delta_t^2\mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2.
\end{aligned}
$$

From $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and (8), we have

$$
\begin{aligned}
\mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{u}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})\|^2 &= \mathbb{E}\|\nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\mathbf{v}_{t+1} - \nabla_{\mathbf{xy}}^2 g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\mathbf{v}_{t+1} \\
&\quad + \nabla_{\mathbf{x}} f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 \\
&\le 2\mathbb{E}\|\left(\nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{xy}}^2 g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\right)\mathbf{v}_{t+1}\|^2 \\
&\quad + 2\mathbb{E}\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 \\
&\le 2\mathbb{E}\|\nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{xy}}^2 g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 p^2 \\
&\quad + 2\mathbb{E}\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2.
\end{aligned}
$$

This completes the proof. □

As demonstrated in Lemma C.9, the hypergradient estimator error e_{t+1}^f comprises five key components: (1) the term $(1-\eta_{t+1})^2\mathbb{E}\|e_t^f\|^2$, representing the per-iteration improvement achieved by the momentum-based update; (2) the error arising from the variation in the Jacobian of the lower-level objectiv; (3) the error caused by the variation in the gradient of the upper-level objective ; (4) the error term $\mathcal{O}(2\beta_t^2\mathbb{E}\|e_t^g\|^2 + 2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2)$, which is due to solving the lower-level problem; and (5) the error term $\mathcal{O}(\delta_t^2\mathbb{E}\|e_t^{\mathbf{v}}\|^2 + 72(1-\eta_{t+1})^2\ell_{g,1}^4\delta_t^2\theta_t^{\mathbf{v}})$, which is introduced by the one-step momentum update in solving the linear system problem.

Lemma C.10. *Let Assumption 3.4 holds. Then, for the sequence of functions $\{f_t\}_{t=1}^T$, we have*

$$
\sum_{t=1}^T \left(f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1},\mathbf{y}_t^*(\mathbf{x}_{t+1}))\right) \le 2M + V_T,
$$

where M is defined in Assumption 3.4; V_T is defined in (10).

Proof. Note that, we have

$$
\begin{aligned}
&\sum_{t=1}^T \left(f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1},\mathbf{y}_t^*(\mathbf{x}_{t+1}))\right) \\
&= f_1(\mathbf{x}_1,\mathbf{y}_1^*(\mathbf{x}_1)) - f_T(\mathbf{x}_{T+1},\mathbf{y}_T^*(\mathbf{x}_{T+1})) \\
&+ \sum_{t=2}^T \left(f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)) - f_{t-1}(\mathbf{x}_t,\mathbf{y}_{t-1}^*(\mathbf{x}_t))\right) \\
&\le 2M + V_T,
\end{aligned}
$$

where the inequality follows from Assumption 3.4. □

Lemma C.11. *Let $\{f_t\}_{t=1}^T$ denote the sequence of functions presented to Algorithm 1, satisfying Assumptions 3.2, 3.3 and 3.4. Let $\mathcal{P}_{\mathcal{X},\alpha_t}$ be defined as in Definition B.1. For any positive step size α_t such that $\alpha_t \le \frac{1}{L_f}$ for all $t \in [T]$, Algorithm 1 ensures the following bound:*

$$\sum_{t=1}^{T} \left(\alpha_t - L_f\alpha_t^2\right) \mathbb{E} \left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2$$
$$\le 8M + 4V_T + 2M_f^2 \sum_{t=1}^{T} \left(2\alpha_t - L_f\alpha_t^2\right) \left(\mathbb{E}[\theta_t^{\mathbf{y}}] + \mathbb{E}[\theta_t^{\mathbf{v}}]\right)$$
$$+ 2\sum_{t=1}^{T} \left(2\alpha_t - L_f\alpha_t^2\right) \mathbb{E}\left\|e_t^f\right\|^2. \tag{81}$$

Here, $\theta_t^{\mathbf{y}}$ and $\theta_t^{\mathbf{v}}$ are defined in (47); V_T is defined in (10), M is given in Assumption 3.4; and M_f is defined in (40).

Proof. It follows from Lemma C.1 that

$$\begin{aligned}
& f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1})) - f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \\
& \le \langle \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L_f}{2}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|^2 \\
& = -\alpha_t \langle \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)), \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}}) \rangle + \frac{L_f\alpha_t^2}{2}\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}})\right\|^2.
\end{aligned} \tag{82}$$

For the first term on the right hand side of (82), we have that

$$\begin{aligned}
& -\langle \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)), \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}}) \rangle \\
& = -\langle \mathbf{d}_t^{\mathbf{x}}, \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}}) \rangle - \langle \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - \mathbf{d}_t^{\mathbf{x}}, \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}}) \rangle \\
& \le -\frac{1}{2}\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}})\right\|^2 + \frac{1}{2}\left\|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right\|^2,
\end{aligned}$$

where the inequality follows from Lemma B.8.
Let $\tilde{\mathbf{d}}_t(\mathbf{z}_t) = \nabla_{\mathbf{x}} f_t(\mathbf{z}_t) + \nabla_{\mathbf{xy}}^2 g_t(\mathbf{z}_t)\mathbf{v}_t$. Then, from Lemma C.1, we have

$$\begin{aligned}
\left\|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right\|^2 &= \left\|\mathbf{d}_t^{\mathbf{x}} - \tilde{\mathbf{d}}_t(\mathbf{z}_t) + \tilde{\mathbf{d}}_t(\mathbf{z}_t) - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right\|^2 \\
&\le 2\left\|\mathbf{d}_t^{\mathbf{x}} - \tilde{\mathbf{d}}_t(\mathbf{z}_t)\right\|^2 + 2\left\|\tilde{\mathbf{d}}_t(\mathbf{z}_t) - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right\|^2 \\
&\le 2\left\|e_t^f\right\|^2 + 2\left\|\tilde{\mathbf{d}}_t(\mathbf{z}_t) - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right\|^2 \\
&\le 2\left\|e_t^f\right\|^2 + M_f^2\left(\theta_t^{\mathbf{y}} + \theta_t^{\mathbf{v}}\right),
\end{aligned} \tag{83}$$

where $e_t^f := \mathbf{d}_t^{\mathbf{x}} - \tilde{\mathbf{d}}_t(\mathbf{z}_t)$. This implies that

$$\begin{aligned}
& -\langle \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)), \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}}) \rangle \\
& \le -\frac{1}{2}\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}})\right\|^2 + 2\left\|e_t^f\right\|^2 + M_f^2\left(\theta_t^{\mathbf{y}} + \theta_t^{\mathbf{v}}\right),
\end{aligned} \tag{84}$$

Plugging the bound (84) into (82), we have that

$$\begin{aligned}
& f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1})) - f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \\
& \le \frac{(L_f\alpha_t^2 - \alpha_t)}{2}\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}})\right\|^2 + 2\alpha_t\left\|e_t^f\right\|^2 + M_f^2\left(\theta_t^{\mathbf{y}} + \theta_t^{\mathbf{v}}\right)\alpha_t,
\end{aligned}$$

which can be rearranged into

$$\begin{aligned}&(\alpha_t - L_f\alpha_t^2)\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\mathbf{d}_t^{\mathbf{x}})\right\|^2\\ &\leq 2f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1},\mathbf{y}_t^*(\mathbf{x}_{t+1})) + 4\alpha_t\left\|e_t^f\right\|^2 + 2M_f^2\left(\theta_t^{\mathbf{y}}+\theta_t^{\mathbf{v}}\right)\alpha_t. \end{aligned}\tag{85}$$

In addition, we have

$$\begin{aligned}&\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\right\|^2\\ &\leq 2\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\mathbf{d}_t^{\mathbf{x}}) - \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\right\|^2 + 2\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\mathbf{d}_t^{\mathbf{x}})\right\|^2\\ &\leq 2\left\|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right\|^2 + 2\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\mathbf{d}_t^{\mathbf{x}})\right\|^2\\ &\leq 4\left\|e_t^f\right\|^2 + 4M_f^2\left(\theta_t^{\mathbf{y}}+\theta_t^{\mathbf{v}}\right) + 4\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\mathbf{d}_t^{\mathbf{x}})\right\|^2, \end{aligned}\tag{86}$$

where the second inequaliy follows from non-expansiveness of the projection operator and the last inequality follows from (83).

Combining (85) and (86), we have

$$\begin{aligned}&\sum_{t=1}^{T}\left(\alpha_t - L_f\alpha_t^2\right)\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\right\|^2\\ &\leq 4\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1},\mathbf{y}_t^*(\mathbf{x}_{t+1}))\right)\\ &+ 2M_f^2\sum_{t=1}^{T}\left(2\alpha_t - L_f\alpha_t^2\right)\left(\theta_t^{\mathbf{y}}+\theta_t^{\mathbf{v}}\right) + 2\sum_{t=1}^{T}\left(2\alpha_t - L_f\alpha_t^2\right)\left\|e_t^f\right\|^2\\ &\leq 8M + 4V_T\\ &+ 2M_f^2\sum_{t=1}^{T}\left(2\alpha_t - L_f\alpha_t^2\right)\left(\theta_t^{\mathbf{y}}+\theta_t^{\mathbf{v}}\right) + 2\sum_{t=1}^{T}\left(2\alpha_t - L_f\alpha_t^2\right)\left\|e_t^f\right\|^2, \end{aligned}$$

where the second inequality is due to Lemma C.10. □

Lemma C.12. *Let Assumptions 3.3 and 3.4 hold. Let $\{\mathbf{x}_t\}_{t=1}^T$ be generated according to Algorithm 1. Then, we have*

$$\left\|\mathbf{x}_t - \mathbf{x}_{t+1}\right\|^2 \leq 2\alpha_t^2\left(\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\right\|^2 + M_f^2\left(\theta_t^{\mathbf{y}}+\theta_t^{\mathbf{v}}\right)\right),$$

where $\theta_t^{\mathbf{y}}$ and $\theta_t^{\mathbf{v}}$ are defined in (47).

Proof. From the update rule of Algorithm 1, we have

$$\begin{aligned}\left\|\mathbf{x}_t - \mathbf{x}_{t+1}\right\|^2 &= \alpha_t^2\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\mathbf{d}_t^{\mathbf{x}})\right\|^2\\ &\leq 2\alpha_t^2\Big(\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\right\|^2\\ &+ \left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\mathbf{d}_t^{\mathbf{x}}) - \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\right\|^2\Big)\\ &\leq 2\alpha_t^2\Big(\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\right\|^2\\ &+ \left\|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right\|^2\Big)\\ &\leq 2\alpha_t^2\left(\left\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\right\|^2 + M_f^2\left(\theta_t^{\mathbf{y}}+\theta_t^{\mathbf{v}}\right)\right), \end{aligned}\tag{87}$$

where the first inequality is by $(a+b)^2 \leq 2a^2+2b^2$; the second inequality follows from non-expansiveness of the projection operator; and the last inequality follows from Eq. (37a) in Lemma C.1. □

C.5. Proof of Theorem 3.6

Proof. **Bounding $\mathbb{E}\|e_t^f\|^2$ in (75)** . From (75), we have

$$
\begin{aligned}
&\frac{\mathbb{E}\|e_{t+1}^f\|^2}{\alpha_t}-\frac{\mathbb{E}\|e_t^f\|^2}{\alpha_{t-1}}\le\left(\frac{(1-\eta_{t+1})^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}\right)\mathbb{E}\|e_t^f\|^2+\frac{4\eta_{t+1}^2}{\alpha_t}(\frac{\sigma_{g_{\mathbf{xy}}}^2}{\bar{b}}p^2+\frac{\sigma_{f_{\mathbf{x}}}^2}{b})\\
&+\frac{12p^2}{\alpha_t}(1-\eta_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{xy}}^2 g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)-\nabla_{\mathbf{xy}}^2 g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\|^2\\
&+\frac{12}{\alpha_t}(1-\eta_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1})-\nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2\\
&+\frac{72}{\alpha_t}(1-\eta_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2+2\beta_t^2\mathbb{E}\|e_t^g\|^2+2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right)\\
&+\frac{72}{\alpha_t}\ell_{g,1}^2(1-\eta_{t+1})^2\delta_t^2\mathbb{E}\|e_t^{\mathbf{v}}\|^2+\frac{72}{\alpha_t}(1-\eta_{t+1})^2\ell_{g,1}^4\delta_t^2\mathbb{E}[\theta_t^{\mathbf{v}}]. \qquad (88)
\end{aligned}
$$

With respect to the coefficient of the first term on the right-hand side of equation (88), it is important to note that we have:

$$
\frac{(1-\eta_{t+1})^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}\le\frac{1}{\alpha_t}-\frac{\eta_{t+1}}{\alpha_t}-\frac{1}{\alpha_{t-1}}. \qquad (89)
$$

Using the definition of α_t in (15), we have

$$
\begin{aligned}
\frac{1}{\alpha_t}-\frac{1}{\alpha_{t-1}}&=(c+t)^{1/3}-(c+t-1)^{1/3}\overset{(i)}{\le}\frac{1}{3(c+t-1)^{2/3}}\overset{(ii)}{\le}\frac{1}{3(\frac{c}{2}+t)^{2/3}}\\
&=\frac{2^{2/3}}{3(c+2t)^{2/3}}\overset{(iii)}{\le}\frac{2^{2/3}}{3(c+t)^{2/3}}\overset{(iv)}{\le}\frac{2^{2/3}}{3}\alpha_t^2\overset{(vi)}{\le}\frac{\alpha_t}{6L_f}, \qquad (90)
\end{aligned}
$$

where the (i) follows from $(a+b)^{1/3}-a^{1/3}\le b/(3a^{2/3})$; (ii) follows from $c\ge 2$ in (104); (iii) follows from (15); (iv) follows from $\alpha_t\le 1/4L_f$ in (104).

Substituting (90) into (89) and using $\delta_t=c_\delta\alpha_t$ and $\eta_{t+1}=c_\eta\alpha_t^2$, we have

$$
\frac{(1-\eta_{t+1})^2}{\alpha_t}-\frac{1}{\alpha_{t-1}}\le\frac{\alpha_t}{6L_f}-\frac{\eta_{t+1}}{\alpha_t}=\frac{\alpha_t}{6L_f}-c_\eta\alpha_t\le-5\Omega\alpha_t, \qquad (91)
$$

where the inequalities follow from $c_\eta=\frac{1}{6L_f}+5\Omega$ with Ω in (103).

Then, substituting (91) into (88) yields

$$
\begin{aligned}
&\frac{1}{\Omega}\mathbb{E}\left(\frac{\|e_{t+1}^f\|^2}{\alpha_t}-\frac{\|e_t^f\|^2}{\alpha_{t-1}}\right)\le-5\alpha_t\mathbb{E}\|e_t^f\|^2+\frac{4\eta_{t+1}^2}{\Omega\alpha_t}(\frac{\sigma_{g_{\mathbf{xy}}}^2}{b}p^2+\frac{\sigma_{f_{\mathbf{x}}}^2}{\acute{b}})\\
&+\frac{12p^2}{\Omega\alpha_t}(1-\eta_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{xy}}^2 g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)-\nabla_{\mathbf{xy}}^2 g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\|^2\\
&+\frac{12}{\Omega\alpha_t}(1-\eta_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1})-\nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2\\
&+\frac{72}{\Omega\alpha_t}(1-\eta_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2+2\beta_t^2\mathbb{E}\|e_t^g\|^2+2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right)\\
&+\frac{72}{\Omega\alpha_t}\ell_{g,1}^2(1-\eta_{t+1})^2\delta_t^2\mathbb{E}\|e_t^{\mathbf{v}}\|^2+\frac{72}{\Omega\alpha_t}(1-\eta_{t+1})^2\ell_{g,1}^4\delta_t^2\mathbb{E}[\theta_t^{\mathbf{v}}]. \qquad (92)
\end{aligned}
$$

Bounding $\mathbb{E}\|e_t^g\|^2$ in (44) .
From (44), we have

$$
\begin{aligned}
\frac{\mathbb{E}\|e^g_{t+1}\|^2}{\alpha_t} - \frac{\mathbb{E}\|e^g_t\|^2}{\alpha_{t-1}} \leq & \left(\frac{1}{\alpha_t}(1-\gamma_{t+1})^2(1+48\ell_{g,1}^2\beta_t^2) - \frac{1}{\alpha_{t-1}}\right)\mathbb{E}\|e^g_t\|^2 \\
& + 2\frac{\gamma_{t+1}^2}{\alpha_t}\frac{\sigma_{g_{\mathbf{y}}}^2}{b} + \frac{24}{\alpha_t}(1-\gamma_{t+1})^2\ell_{g,1}^2\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \\
& + \frac{6}{\alpha_t}(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 \\
& + 48(1-\gamma_{t+1})^2\ell_{g,1}^2\frac{\beta_t^2}{\alpha_t}\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2. \qquad (93)
\end{aligned}
$$

Let us examine the coefficient of the first term on the right-hand side of Eq. (93). Specifically, for $\gamma_{t+1} = c_\gamma\alpha_t^2$ and $\beta_t = c_\beta\alpha_t$, we have:

$$
\begin{aligned}
\frac{1}{\alpha_t}(1-\gamma_{t+1})^2(1+48\ell_{g,1}^2\beta_t^2) - \frac{1}{\alpha_{t-1}} &\leq \frac{1}{\alpha_t}(1-\gamma_{t+1})(1+48\ell_{g,1}^2\beta_t^2) - \frac{1}{\alpha_{t-1}} \\
&= \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - \frac{\gamma_{t+1}}{\alpha_t} + \frac{1-\gamma_{t+1}}{\alpha_t}48\ell_{g,1}^2\beta_t^2 \\
&= \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - c_\gamma\alpha_t + (\frac{1}{\alpha_t} - c_\gamma\alpha_t)48\ell_{g,1}^2c_\beta^2\alpha_t^2 \\
&\leq \frac{\alpha_t}{6L_f} + 48\ell_{g,1}^2c_\beta^2\alpha_t - c_\gamma\alpha_t, \qquad (94)
\end{aligned}
$$

where the last inequality follows from (90).

Recalling Φ from (103) that we selected, we obtain

$$
c_\gamma = \frac{1}{6L_f} + 48\ell_{g,1}^2c_\beta^2 + \hbar\Phi, \quad \text{where} \quad \hbar := 25\frac{M_f^2}{L_{\mu_g}^2},
$$

which, when combined with Eq. (94), results in

$$
\frac{1}{\alpha_t}(1-\gamma_{t+1})^2(1+48\ell_{g,1}^2\beta_t^2) - \frac{1}{\alpha_{t-1}} \leq -\hbar\Phi\alpha_t. \qquad (95)
$$

Substituting eq. (95) into eq. (93) yields

$$
\begin{aligned}
\frac{1}{\Phi}\left(\frac{\mathbb{E}\|e^g_{t+1}\|^2}{\alpha_t} - \frac{\mathbb{E}\|e^g_t\|^2}{\alpha_{t-1}}\right) \leq & -\hbar\alpha_t\mathbb{E}\|e^g_t\|^2 \\
& + 2\frac{\gamma_{t+1}^2}{\Phi\alpha_t}\frac{\sigma_{g_{\mathbf{y}}}^2}{b} + \frac{24}{\Phi\alpha_t}(1-\gamma_{t+1})^2\ell_{g,1}^2\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 \\
& + \frac{6}{\Phi\alpha_t}(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 \\
& + 48(1-\gamma_{t+1})^2\ell_{g,1}^2\frac{\beta_t^2}{\Phi\alpha_t}\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2. \qquad (96)
\end{aligned}
$$

Bounding $\mathbb{E}\|e^{\mathbf{v}}_t\|^2$ in (56) .

From (56), we get

$$\begin{aligned}
&\frac{\mathbb{E}\|e^{\mathbf{v}}_{t+1}\|^2}{\alpha_t} - \frac{\mathbb{E}\|e^{\mathbf{v}}_{t}\|^2}{\alpha_{t-1}} \leq \left(\frac{1}{\alpha_t}(1-\lambda_{t+1})^2(1+72\ell_{g,1}^2\delta_t^2) - \frac{1}{\alpha_{t-1}}\right)\mathbb{E}\|e^{\mathbf{v}}_{t}\|^2 \\
&+ 4\frac{\lambda_{t+1}^2}{\alpha_t}(\frac{\sigma^2_{g_{\mathbf{yy}}}}{\bar{b}}p^2 + \frac{\sigma^2_{f_{\mathbf{y}}}}{b}) + \frac{12p^2}{\alpha_t}(1-\lambda_{t+1})^2\mathbb{E}\|\nabla^2_{\mathbf{y}}g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right) - \nabla^2_{\mathbf{y}}g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\|^2 \\
&+ \frac{12}{\alpha_t}(1-\lambda_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}}f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 \\
&+ \frac{72}{\alpha_t}(1-\lambda_{t+1})^2(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\beta_t^2\mathbb{E}\|e^g_t\|^2 + 2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right) \\
&+ \frac{72}{\alpha_t}(1-\lambda_{t+1})^2\ell_{g,1}^4\delta_t^2\mathbb{E}[\theta^{\mathbf{v}}_t].
\end{aligned} \tag{97}$$

Let us examine the coefficient of the first term on the right-hand side of equation (97). Specifically, for $\lambda_{t+1} = c_\lambda\alpha_t^2$ and $\delta_t = c_\delta\alpha_t$, we have:

$$\begin{aligned}
\frac{1}{\alpha_t}(1-\lambda_{t+1})^2(1+72\ell_{g,1}^2\delta_t^2) - \frac{1}{\alpha_{t-1}} &\leq \frac{1}{\alpha_t}(1-\lambda_{t+1})(1+72\ell_{g,1}^2\delta_t^2) - \frac{1}{\alpha_{t-1}} \\
&= \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - \frac{\lambda_{t+1}}{\alpha_t} + \frac{1-\lambda_{t+1}}{\alpha_t}72\ell_{g,1}^2\delta_t^2 \\
&= \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - c_\lambda\alpha_t + (\frac{1}{\alpha_t} - c_\lambda\alpha_t)72\ell_{g,1}^2c_\delta^2\alpha_t^2 \\
&\leq \frac{\alpha_t}{6L_f} + 72\ell_{g,1}^2c_\delta^2\alpha_t - c_\lambda\alpha_t,
\end{aligned} \tag{98}$$

where the last inequality follows from (90).

Recalling Ψ from (103) that we selected, we obtain

$$c_\lambda = \frac{1}{6L_f} + 72\ell_{g,1}^2c_\delta^2 + \jmath\Psi, \quad \text{where} \quad \jmath = 90\frac{M_f^2}{L_{\mu_g}^2},$$

which, when combined with Eq. (98), results in

$$\frac{1}{\alpha_t}(1-\lambda_{t+1})^2(1+72\ell_{g,1}^2\delta_t^2) - \frac{1}{\alpha_{t-1}} \leq -\jmath\Psi\alpha_t. \tag{99}$$

Substituting eq. (99) into eq. (97) yields

$$\begin{aligned}
&\frac{1}{\Psi}\left(\frac{\mathbb{E}\|e^{\mathbf{v}}_{t+1}\|^2}{\alpha_t} - \frac{\mathbb{E}\|e^{\mathbf{v}}_{t}\|^2}{\alpha_{t-1}}\right) \leq -\jmath\alpha_t\mathbb{E}\|e^{\mathbf{v}}_{t}\|^2 \\
&+ 4\frac{\lambda_{t+1}^2}{\Psi\alpha_t}(\frac{\sigma^2_{g_{\mathbf{yy}}}}{\bar{b}}p^2 + \frac{\sigma^2_{f_{\mathbf{y}}}}{b}) + \frac{12p^2}{\Psi\alpha_t}(1-\lambda_{t+1})^2\mathbb{E}\|\nabla^2_{\mathbf{y}}g_t\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right) - \nabla^2_{\mathbf{y}}g_{t+1}\left(\mathbf{x}_{t+1},\mathbf{y}_{t+1}\right)\|^2 \\
&+ \frac{12}{\Psi\alpha_t}(1-\lambda_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}}f_t(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}}f_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1})\|^2 \\
&+ \frac{72}{\Psi\alpha_t}(1-\lambda_{t+1})^2(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\left(\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\beta_t^2\mathbb{E}\|e^g_t\|^2 + 2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\right) \\
&+ \frac{72}{\Psi\alpha_t}(1-\lambda_{t+1})^2\ell_{g,1}^4\delta_t^2\mathbb{E}[\theta^{\mathbf{v}}_t].
\end{aligned} \tag{100}$$

Combining the outcomes . We recall from Lemma C.12 that we have

$$\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \leq 2\alpha_t^2\left(\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\|^2 + M_f^2\left(\theta^{\mathbf{y}}_t + \theta^{\mathbf{v}}_t\right)\right). \tag{101}$$

Let

$$\Lambda := \Gamma \sum_{t=1}^{T} \left(\mathbb{E}[\theta_{t+1}^{\mathbf{y}}] - \mathbb{E}[\theta_t^{\mathbf{y}}]\right) + \Upsilon \sum_{t=1}^{T} \left(\mathbb{E}[\theta_{t+1}^{\mathbf{v}}] - \mathbb{E}[\theta_t^{\mathbf{v}}]\right) + \frac{1}{\Phi} \sum_{t=1}^{T} \left(\frac{\mathbb{E}\|e_{t+1}^{g}\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^{g}\|^2}{\alpha_{t-1}}\right)$$
$$+ \frac{1}{\Psi} \sum_{t=1}^{T} \left(\frac{\mathbb{E}\|e_{t+1}^{\mathbf{v}}\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^{\mathbf{v}}\|^2}{\alpha_{t-1}}\right) + \frac{1}{\Omega} \sum_{t=1}^{T} \left(\frac{\mathbb{E}\|e_{t+1}^{f}\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^{f}\|^2}{\alpha_{t-1}}\right). \tag{102}$$

Here

$$\Gamma = \frac{11M_f^2}{L_{\mu_g} c_\beta}, \quad \Upsilon = \frac{22M_f^2}{L_{\mu_g} c_\delta}, \quad \Phi = 480\ell_{g,1}^2,$$
$$\Psi = \max\left\{144(\ell_{g,2}^2 p^2 + \ell_{f,1}^2)\left(10 + \frac{L_{\mu_g}^2 c_\beta^2}{11M_f^2}\right), \frac{288\ell_{g,1}^4}{M_f^2} c_\delta^2\right\}, \tag{103}$$
$$\Omega = \max\left\{144(\ell_{g,2}^2 p^2 + \ell_{f,1}^2)\left(10 + \frac{L_{\mu_g}^2 c_\beta^2}{11M_f^2}\right), \frac{288\ell_{g,1}^4}{M_f^2} c_\delta^2\right\}.$$

Here, from (15), we have

$$\begin{aligned}
c &\geq \max\left\{4L_f, c_\beta(\mu_g + \ell_{g,1}), 2\right\}, \\
c_\beta &= \sqrt{880\frac{L_{\mathbf{y}}^2 M_f^2}{L_{\mu_g}^2}}, \\
c_\delta &= \sqrt{3520\frac{\nu^2 M_f^2}{L_{\mu_g}^2 \mu_g^2}(1 + 2L_{\mathbf{y}}^2)}, \\
c_\gamma &= \frac{2}{3L_f} + 48\ell_{g,1}^2 c_\beta^2 + \hbar\Phi, \quad \text{where} \quad \hbar := 25\frac{M_f^2}{L_{\mu_g}^2}, \\
c_\eta &= \frac{2}{3L_f} + 5\Omega, \\
c_\lambda &= \frac{2}{3L_f} + 72\ell_{g,1}^2 c_\delta^2 + \jmath\Psi, \quad \text{where} \quad \jmath = 90\frac{M_f^2}{L_{\mu_g}^2}.
\end{aligned} \tag{104}$$

Using (100), (96), (92), (81), (67), and (48), along with (101) and the fact that α_t decreases with respect to t, we obtain:

$$\sum_{t=1}^{T} A(\alpha_t, \beta_t, \delta_t)\mathbb{E}\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2 + \Lambda$$
$$\leq 8M + 4V_T + \sum_{t=1}^{T} B(\alpha_t, \beta_t, \delta_t)\mathbb{E}[\theta_t^{\mathbf{v}}] + \sum_{t=1}^{T} C(\alpha_t, \beta_t)\mathbb{E}[\theta_t^{\mathbf{y}}] \tag{105a}$$
$$+ \sum_{t=1}^{T} D(\alpha_t)\mathbb{E}\|e_t^{f}\|^2 + \sum_{t=1}^{T} F(\beta_t, \delta_t)\mathbb{E}\|e_t^{g}\|^2 + \sum_{t=1}^{T} I(\alpha_t, \beta_t, \delta_t)\mathbb{E}\|e_t^{\mathbf{v}}\|^2 \tag{105b}$$
$$+ \sum_{t=1}^{T} L(\beta_t)\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \sum_{t=2}^{T} N(\beta_t, \delta_t) \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2 \tag{105c}$$
$$+ \frac{\sigma_{g_{\mathbf{y}}}^2}{b}\frac{2}{\Phi}\sum_{t=1}^{T}\frac{\gamma_{t+1}^2}{\alpha_t} + \frac{4}{\Psi}\left(\frac{\sigma_{g_{\mathbf{yy}}}^2}{b}p^2 + \frac{\sigma_{f_{\mathbf{y}}}^2}{b}\right)\sum_{t=1}^{T}\frac{\lambda_{t+1}^2}{\alpha_t} + \frac{4}{\Omega}\left(\frac{\sigma_{g_{\mathbf{xy}}}^2}{b}p^2 + \frac{\sigma_{f_{\mathbf{x}}}^2}{b}\right)\sum_{t=1}^{T}\frac{\eta_{t+1}^2}{\alpha_t} \tag{105d}$$
$$+ \frac{6}{\Phi\alpha_T}G_{\mathbf{y},T} + \frac{12p^2}{\Omega\alpha_T}G_{\mathbf{xy},T} + \frac{12p^2}{\Psi\alpha_T}G_{\mathbf{yy},T} + \frac{12\ell_{f,1}^2}{\Psi\alpha_T}D_{\mathbf{y},T} + \frac{12\ell_{f,1}^2}{\Omega\alpha_T}D_{\mathbf{x},T}. \tag{105e}$$

Here, M is defined in Assumption 3.4, V_T and $H_{2,T}$ are defined in (10). Moreover, $G_{\mathbf{y},T}$, $G_{\mathbf{xy},T}$, and $G_{\mathbf{yy},T}$ are defined in

(12). Let

$$\begin{aligned}
E(\beta_t,\delta_t) &:= \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}\beta_t}\Gamma + \frac{8\nu^2}{L_{\mu_g}\mu_g^2\delta_t}(1+2L_{\mathbf{y}}^2)\Upsilon + 72(1-\eta_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\frac{1}{\Omega\alpha_t} \\
&\quad + 24(1-\gamma_{t+1})^2\ell_{g,1}^2\frac{1}{\Phi\alpha_t} + 72(1-\lambda_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\frac{1}{\Psi\alpha_t}, \\
A(\alpha_t,\beta_t,\delta_t) &:= \alpha_t - (L_f + 2E(\beta_t,\delta_t))\,\alpha_t^2, \\
B(\alpha_t,\beta_t,\delta_t) &:= -\frac{L_{\mu_g}\Upsilon}{4}\delta_t + 4M_f^2\alpha_t - 2M_f^2L_f\alpha_t^2 + 2M_f^2E(\beta_t,\delta_t)\alpha_t^2 \\
&\quad + 72(1-\lambda_{t+1})^2\ell_{g,1}^4\delta_t^2\frac{1}{\Psi\alpha_t} + 72(1-\eta_{t+1})^2\ell_{g,1}^4\delta_t^2\frac{1}{\Omega\alpha_t}, \\
C(\alpha_t,\beta_t) &:= -\frac{L_{\mu_g}\Gamma}{2}\beta_t + 4M_f^2\alpha_t - 2L_fM_f^2\alpha_t^2 + 2M_f^2E(\beta_t,\delta_t)\alpha_t^2, \\
D(\alpha_t) &:= 2\left(2\alpha_t - L_f\alpha_t^2\right) - 5\alpha_t, \\
F(\alpha_t,\beta_t,\delta_t) &:= \frac{2\Gamma}{L_{\mu_g}}\beta_t - \hbar\alpha_t + 72(1-\lambda_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)2\frac{\beta_t^2}{\Psi\alpha_t} \\
&\quad + 72(1-\eta_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)2\frac{\beta_t^2}{\Omega\alpha_t}, \\
I(\alpha_t,\beta_t,\delta_t) &:= \frac{4\Upsilon}{L_{\mu_g}}\delta_t - \jmath\alpha_t + 72\ell_{g,1}^2(1-\eta_{t+1})^2\frac{\delta_t^2}{\Omega\alpha_t}.
\end{aligned} \tag{106}$$

Moreover, we have

$$\begin{aligned}
L(\beta_t) &:= -\frac{2\Gamma}{\mu_g+\ell_{g,1}}\beta_t + \Gamma\beta_t^2 + 48(1-\gamma_{t+1})^2\ell_{g,1}^2\frac{\beta_t^2}{\Phi\alpha_t} \\
&\quad + 72(1-\lambda_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)2\frac{\beta_t^2}{\Psi\alpha_t} + 72(1-\eta_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)2\frac{\beta_t^2}{\Omega\alpha_t}, \\
N(\beta_t,\delta_t) &:= \frac{4}{L_{\mu_g}\beta_t}\Gamma + \frac{16\nu^2}{L_{\mu_g}\mu_g^2\delta_t}\Upsilon.
\end{aligned} \tag{107}$$

Note that, we have

$$\begin{aligned}
E(\beta_t,\delta_t) &= \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}\beta_t}\Gamma + \frac{8\nu^2}{L_{\mu_g}\mu_g^2\delta_t}(1+2L_{\mathbf{y}}^2)\Upsilon + 72(1-\eta_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\frac{1}{\Omega\alpha_t} \\
&\quad + 24(1-\gamma_{t+1})^2\ell_{g,1}^2\frac{1}{\Phi\alpha_t} + 72(1-\lambda_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\frac{1}{\Psi\alpha_t},
\end{aligned}$$

which together with $\beta_t = c_\beta\alpha_t$, $\delta_t = c_\delta\alpha_t$, we have

$$\begin{aligned}
\alpha_t^2E(\beta_t,\delta_t) &= \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}}\Gamma\frac{\alpha_t^2}{\beta_t} + \frac{8\nu^2}{L_{\mu_g}\mu_g^2}(1+2L_{\mathbf{y}}^2)\Upsilon\frac{\alpha_t^2}{\delta_t} + 72(1-\eta_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\frac{\alpha_t}{\Omega} \\
&\quad + 24(1-\gamma_{t+1})^2\ell_{g,1}^2\frac{\alpha_t}{\Phi} + 72(1-\lambda_{t+1})^2(\ell_{g,2}^2p^2+\ell_{f,1}^2)\frac{\alpha_t}{\Psi} \\
&\le \frac{44L_{\mathbf{y}}^2}{L_{\mu_g}^2}M_f^2\frac{\alpha_t}{c_\beta^2} + \frac{176\nu^2}{L_{\mu_g}^2\mu_g^2}(1+2L_{\mathbf{y}}^2)M_f^2\frac{\alpha_t}{c_\delta^2} \\
&\quad + 24\ell_{g,1}^2\frac{\alpha_t}{\Phi} + 72(\ell_{g,2}^2p^2+\ell_{f,1}^2)(\frac{1}{\Omega}+\frac{1}{\Psi})\alpha_t \\
&\le \frac{\alpha_t}{4},
\end{aligned} \tag{108}$$

where the first inequality follows from $\Gamma = \frac{11M_f^2}{L_{\mu_g}c_\beta}$ and $\Upsilon = \frac{22M_f^2}{L_{\mu_g}c_\delta}$ in (103); the last inequality follows from $c_\beta \ge$

$\sqrt{880\frac{L_{\mathbf{y}}^2 M_f^2}{L_{\mu_g}^2}}$, $c_\delta \geq \sqrt{3520\frac{\nu^2 M_f^2}{L_{\mu_g}^2 \mu_g^2}(1+2L_{\mathbf{y}}^2)}$, in (104) and $\Phi \geq 480\ell_{g,1}^2$, and $\Omega, \Psi \geq 1440(\ell_{g,2}^2 p^2 + \ell_{f,1}^2)$ in (103).
Moreover, we have

$$
\begin{aligned}
A(\alpha_t, \beta_t, \delta_t) &:= \alpha_t - L_f \alpha_t^2 - 2E(\beta_t, \delta_t)\alpha_t^2 \\
&\geq \alpha_t - L_f \alpha_t^2 - \frac{\alpha_t}{2} \\
&\geq \frac{\alpha_t}{4},
\end{aligned} \tag{109}
$$

where the last inequality is by $\alpha_t \leq 1/4L_f$ in (104).
Bounding (105a) .
From (106), we have

$$
\begin{aligned}
B(\alpha_t, \beta_t, \delta_t) &= -\frac{L_{\mu_g}\Upsilon}{4}\delta_t + 4M_f^2\alpha_t - 2M_f^2 L_f \alpha_t^2 + 2M_f^2 E(\beta_t, \delta_t)\alpha_t^2 \\
&\quad + 72(1-\lambda_{t+1})^2 \ell_{g,1}^4 \delta_t^2 \frac{1}{\Psi\alpha_t} + 72(1-\eta_{t+1})^2 \ell_{g,1}^4 \delta_t^2 \frac{1}{\Omega\alpha_t} \\
&\leq -\frac{L_{\mu_g}\Upsilon}{4}\delta_t + 4M_f^2\alpha_t - 2M_f^2 L_f \alpha_t^2 + \frac{M_f^2}{2}\alpha_t + 72\ell_{g,1}^4(\frac{1}{\Psi} + \frac{1}{\Omega})\frac{\delta_t^2}{\alpha_t} \\
&= \left(-\frac{L_{\mu_g}}{4}\Upsilon c_\delta + \frac{9}{2}M_f^2 + 72\ell_{g,1}^4(\frac{1}{\Psi} + \frac{1}{\Omega})c_\delta^2\right)\alpha_t \\
&\leq -\frac{1}{2}M_f^2\alpha_t,
\end{aligned} \tag{110}
$$

where the first inequality follows from $\beta_t = c_\beta\alpha_t$, $\delta_t = c_\delta\alpha_t$, and (108); the second inequality is by $\Upsilon = \frac{22M_f^2}{L_{\mu_g}c_\delta}$, and $\Psi, \Omega \geq \frac{288\ell_{g,1}^4}{M_f^2}c_\delta^2$ in (103); the last inequality follows from in (103).
Moreover, from (106), and $\beta_t = c_\beta\alpha_t$, we have

$$
\begin{aligned}
C(\alpha_t, \beta_t) &= -\frac{L_{\mu_g}\Gamma}{2}\beta_t + 4M_f^2\alpha_t - 2L_f M_f^2\alpha_t^2 + 2M_f^2 E(\beta_t, \delta_t)\alpha_t^2 \\
&\leq -\frac{L_{\mu_g}}{2}\Gamma c_\beta\alpha_t + \frac{9}{2}M_f^2\alpha_t \\
&= -M_f^2\alpha_t,
\end{aligned} \tag{111}
$$

where the first inequality follows from (108); the last equality follows from $\Gamma = \frac{11M_f^2}{L_{\mu_g}c_\beta}$ in (103).
Thus, from (110) and (111), we get

$$
(105a) \leq \mathcal{O}\left(V_T\right). \tag{112}
$$

Bounding (105b) .
From (106), we also obtain

$$
D(\alpha_t) = 4\alpha_t - 2L_f\alpha_t^2 - 5\alpha_t \leq 0.
$$

From $\beta_t = c_\beta \alpha_t$, $\Gamma = \frac{11M_f^2}{L_{\mu_g} c_\beta}$ and (106), we obtain

$$\begin{aligned} F(\alpha_t, \beta_t, \delta_t) =& \frac{2\Gamma}{L_{\mu_g}} \beta_t - \hbar \alpha_t + 144(1-\lambda_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) \frac{\beta_t^2}{\Psi \alpha_t} \\ &+ 144(1-\eta_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) \frac{\beta_t^2}{\Omega \alpha_t} \\ \leq& \frac{22M_f^2}{L_{\mu_g}^2} \alpha_t - \hbar \alpha_t + 144(\ell_{g,2}^2 p^2 + \ell_{f,1}^2)(\frac{1}{\Psi} + \frac{1}{\Omega}) c_\beta^2 \alpha_t \\ \leq& 24 \frac{M_f^2}{L_{\mu_g}^2} \alpha_t - \hbar \alpha_t \\ =& -\frac{M_f^2}{L_{\mu_g}^2} \alpha_t, \end{aligned}$$

where the second inequality follows from $\Omega, \Psi \geq 144(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) \frac{L_{\mu_g}^2 c_\beta^2}{M_f^2}$ in (103); and the last equality is by $\hbar := 25 \frac{M_f^2}{L_{\mu_g}^2}$.

From $\delta_t = c_\delta \alpha_t$, we obtain

$$\begin{aligned} I(\alpha_t, \beta_t, \delta_t) =& \frac{4\Upsilon}{L_{\mu_g}} \delta_t - \jmath \alpha_t + 72\ell_{g,1}^2 (1-\eta_{t+1})^2 \frac{\delta_t^2}{\Omega \alpha_t} \\ \leq& \frac{4\Upsilon}{L_{\mu_g}} c_\delta \alpha_t - \jmath \alpha_t + 72\ell_{g,1}^2 \frac{c_\delta^2 \alpha_t}{\Omega} \\ \leq& \frac{89M_f^2}{L_{\mu_g}^2} \alpha_t - \jmath \alpha_t \\ =& -\frac{M_f^2}{L_{\mu_g}^2} \alpha_t, \end{aligned}$$

where the second inequality follows from $\Upsilon = \frac{22M_f^2}{L_{\mu_g} c_\delta}$ and $\Omega \geq \frac{72\ell_{g,1}^2 L_{\mu_g}^2}{M_f^2} c_\delta^2$; the last equality follows from $\jmath = 90 \frac{M_f^2}{L_{\mu_g}^2}$.

Thus, we get

$$(105b) \leq 0. \tag{113}$$

Bounding (105c) .
From $\beta_t = c_\beta \alpha_t$ and (107), we have

$$\begin{aligned} L(\beta_t) =& -\frac{2\Gamma \beta_t}{\mu_g + \ell_{g,1}} + \Gamma \beta_t^2 + 48(1-\gamma_{t+1})^2 \ell_{g,1}^2 \frac{\beta_t^2}{\Phi \alpha_t} \\ &+ 72(1-\lambda_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) 2 \frac{\beta_t^2}{\Psi \alpha_t} + 72(1-\eta_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) 2 \frac{\beta_t^2}{\Omega \alpha_t} \\ \leq& -\frac{2\Gamma c_\beta \alpha_t}{\mu_g + \ell_{g,1}} + \Gamma c_\beta^2 \alpha_t^2 + 48\ell_{g,1}^2 c_\beta^2 \frac{\alpha_t}{\Phi} + 144(\ell_{g,2}^2 p^2 + \ell_{f,1}^2)(\frac{1}{\Psi} + \frac{1}{\Omega}) c_\beta^2 \alpha_t \\ \leq& -\frac{2\Gamma c_\beta \alpha_t}{\mu_g + \ell_{g,1}} + \Gamma c_\beta^2 \alpha_t^2 + \frac{3\Gamma c_\beta \alpha_t}{4(\mu_g + \ell_{g,1})} \\ \leq& -\frac{\Gamma c_\beta \alpha_t}{4(\mu_g + \ell_{g,1})}, \end{aligned}$$

where the second inequality is by $\Phi \geq 192\ell_{g,1}^2 \frac{(\mu_g + \ell_{g,1})}{\Gamma} c_\beta$, and $\Omega, \Psi \geq 576(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) \frac{(\mu_g + \ell_{g,1})}{\Gamma} c_\beta$ in (103); the last inequality follows from $\alpha_t \leq \frac{1}{c_\beta (\mu_g + \ell_{g,1})}$ in (104).

From $\beta_t = c_\beta \alpha_t$, $\delta_t = c_\delta \alpha_t$ and (107), we obtain

$$\begin{aligned}
N(\beta_t, \delta_t) &= \frac{4}{L_{\mu_g}\beta_t}\Gamma + \frac{16\nu^2}{L_{\mu_g}\mu_g^2\delta_t}\Upsilon \\
&= \frac{4}{L_{\mu_g}c_\beta\alpha_t}\Gamma + \frac{16\nu^2}{L_{\mu_g}\mu_g^2 c_\delta\alpha_t}\Upsilon.
\end{aligned}$$

Thus, we get

$$\begin{aligned}
(105c) &= \sum_{t=1}^{T} L(\beta_t)\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \sum_{t=2}^{T} N(\beta_t, \delta_t) \sup_{\mathbf{x}\in\mathcal{X}} \|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2 \\
&\le \mathcal{O}\left(\frac{H_{2,T}}{\alpha_T}\right).
\end{aligned} \tag{114}$$

Bounding (105d) .
From $\eta_{t+1} = c_\eta \alpha_t^2$, $\gamma_{t+1} = c_\gamma \alpha_t^2$, $\lambda_{t+1} = c_\lambda \alpha_t^2$, we obtain

$$\begin{aligned}
(105d) &= \frac{\sigma_{g_{\mathbf{y}}}^2}{\bar{b}}\frac{2}{\Phi}\sum_{t=1}^{T}\frac{\gamma_{t+1}^2}{\alpha_t} + \frac{4}{\Psi}\left(\frac{\sigma_{g_{\mathbf{yy}}}^2}{\bar{b}}p^2 + \frac{\sigma_{f_{\mathbf{y}}}^2}{b}\right)\sum_{t=1}^{T}\frac{\lambda_{t+1}^2}{\alpha_t} + \frac{4}{\Omega}\left(\frac{\sigma_{g_{\mathbf{xy}}}^2}{\bar{b}}p^2 + \frac{\sigma_{f_{\mathbf{x}}}^2}{b}\right)\sum_{t=1}^{T}\frac{\eta_{t+1}^2}{\alpha_t} \\
&\le \mathcal{O}\left((\frac{\sigma_{g_{\mathbf{y}}}^2}{\bar{b}} + \frac{\sigma_{g_{\mathbf{yy}}}^2}{\bar{b}} + \frac{\sigma_{f_{\mathbf{y}}}^2}{b} + \frac{\sigma_{g_{\mathbf{xy}}}^2}{\bar{b}} + \frac{\sigma_{f_{\mathbf{x}}}^2}{b})\sum_{t=1}^{T}\alpha_t^3 \right).
\end{aligned} \tag{115}$$

Bounding (105e) .
We have

$$\begin{aligned}
(105e) &= \frac{6}{\Phi\alpha_T}G_{\mathbf{y},T} + \frac{12p^2}{\Omega\alpha_T}G_{\mathbf{xy},T} + \frac{12p^2}{\Psi\alpha_T}G_{\mathbf{yy},T} + \frac{12\ell_{f,1}^2}{\Psi\alpha_T}D_{\mathbf{y},T} + \frac{12\ell_{f,1}^2}{\Omega\alpha_T}D_{\mathbf{x},T} \\
&\le \mathcal{O}\left(\frac{1}{\alpha_T}(G_{\mathbf{y},T} + G_{\mathbf{xy},T} + G_{\mathbf{yy},T} + D_{\mathbf{y},T} + D_{\mathbf{x},T})\right).
\end{aligned} \tag{116}$$

Let

$$\begin{aligned}
&G_T := G_{\mathbf{y},T} + G_{\mathbf{xy},T} + G_{\mathbf{yy},T}, \\
&D_T := D_{\mathbf{y},T} + D_{\mathbf{x},T}, \\
&\sigma^2 := \sigma_{g_{\mathbf{y}}}^2 + \sigma_{g_{\mathbf{yy}}}^2 + \sigma_{f_{\mathbf{y}}}^2 + \sigma_{g_{\mathbf{xy}}}^2 + \sigma_{f_{\mathbf{x}}}^2, \\
&b = \bar{b} = 1.
\end{aligned}$$

By inequalities (109), (112), (113), (114), (115), (116), we have

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{\alpha_t}{2}\mathbb{E}\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2 + \Lambda \\
&\le \mathcal{O}\left(V_T + \frac{H_{2,T}}{\alpha_T} + \frac{\sigma^2}{b}\sum_{t=1}^{T}\alpha_t^3 + \frac{G_T}{\alpha_T} + \frac{D_T}{\alpha_T}\right).
\end{aligned} \tag{117}$$

From the definition of Λ in (102), we have

$$
\begin{aligned}
-\Lambda &= \Gamma \sum_{t=1}^{T} \left(\mathbb{E}[\theta_t^{\mathbf{y}}] - \mathbb{E}[\theta_{t+1}^{\mathbf{y}}]\right) + \Upsilon \sum_{t=1}^{T} \left(\mathbb{E}[\theta_t^{\mathbf{v}}] - \mathbb{E}[\theta_{t+1}^{\mathbf{v}}]\right) + \frac{1}{\Phi} \sum_{t=1}^{T} \left(\frac{\mathbb{E}\|e_t^g\|^2}{\alpha_{t-1}} - \frac{\mathbb{E}\|e_{t+1}^g\|^2}{\alpha_t}\right) \\
&\quad + \frac{1}{\Psi} \sum_{t=1}^{T} \left(\frac{\mathbb{E}\|e_t^{\mathbf{v}}\|^2}{\alpha_{t-1}} - \frac{\mathbb{E}\|e_{t+1}^{\mathbf{v}}\|^2}{\alpha_t}\right) + \frac{1}{\Omega} \sum_{t=1}^{T} \left(\frac{\mathbb{E}\|e_t^f\|^2}{\alpha_{t-1}} - \frac{\mathbb{E}\|e_{t+1}^f\|^2}{\alpha_t}\right) \\
&\leq \Gamma\theta_1^{\mathbf{y}} + \Upsilon\theta_1^{\mathbf{v}} + \frac{\sigma_{g_{\mathbf{y}}}^2}{\Phi\alpha_0} + \frac{\sigma_{g_{\mathbf{yy}}}^2 + \sigma_{f_{\mathbf{y}}}^2}{\Psi\alpha_0} + \frac{\sigma_{g_{\mathbf{xy}}}^2 + \sigma_{f_{\mathbf{x}}}^2}{\Omega\alpha_0}.
\end{aligned} \tag{118}
$$

Using (118), we get

$$
\begin{aligned}
&\sum_{t=1}^{T} \frac{\alpha_t}{2} \mathbb{E}\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2 \\
&\leq \mathcal{O}\left(V_T + \frac{H_{2,T}}{\alpha_T} + \frac{\sigma^2}{b}\sum_{t=1}^{T}\alpha_t^3 + \frac{G_T}{\alpha_T} + \frac{D_T}{\alpha_T} - \Lambda\right) \\
&\leq \mathcal{O}\left(V_T + \theta_1^{\mathbf{y}} + \theta_1^{\mathbf{v}} + \frac{\sigma^2}{b}\sum_{t=1}^{T}\alpha_t^3 + \frac{H_{2,T}}{\alpha_T} + \frac{G_T}{\alpha_T} + \frac{D_T}{\alpha_T} + \frac{\sigma^2}{\alpha_0}\right).
\end{aligned}
$$

Since $\alpha_t = 1/(c+t)^{1/3}$, we get

$$
\sum_{t=1}^{T} \alpha_t^3 = \sum_{t=1}^{T} \frac{1}{c+t} \leq \sum_{t=1}^{T} \frac{1}{1+t} \leq \log(T+1),
$$

which, combined with the fact that α_t decreases with respect to t and by multiplying both sides by $2/\alpha_T$, results in Thus, we have

$$
\begin{aligned}
\text{BL-Reg}_T &= \sum_{t=1}^{T} \mathbb{E}\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2 \\
&\leq \mathcal{O}\Big(\frac{1}{\alpha_T}(V_T + \|\mathbf{y}_1 - \mathbf{y}_1^*(\mathbf{x}_1)\|^2 + \|\mathbf{v}_1 - \mathbf{v}_1^*(\mathbf{x}_1)\|^2 + \sigma^2\log(T+1) + \frac{\sigma^2}{\alpha_0}) \\
&\quad + \frac{1}{\alpha_T^2}(H_{2,T} + G_T + D_T)\Big).
\end{aligned}
$$

This completes the proof. $\square$

D. Proof of Regret Bounds for Zeroth Order SOGD (ZO-SOGD)

Proof Roadmap. We provide Lemma D.9, which quantifies the error between the approximated direction of the momentum-based gradient estimator, $\hat{\mathbf{d}}_t^{\mathbf{y}}$ and the true direction, $\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)$, at each iteration. Lemma D.11 assesses the convergence of the iterative solutions $\{\mathbf{y}_t\}_{t=1}^T$, specifically the gap $\mathbb{E}\left[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right]$, while accounting for the error introduced in Lemma D.9. To establish Lemma D.15, which quantifies the error between the approximated direction of the momentum-based gradient estimator, $\hat{\mathbf{d}}_t^{\mathbf{v}}$, and the true direction, $\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) + \nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\,\mathbf{v}_t$, we need to present Lemma D.13. This lemma quantifies the error between $\hat{\mathbf{d}}_t^{\mathbf{v}}$ and $\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) + \frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t))$. Then, Lemma D.17 captures the error of the system solution of Problem (17), i.e., gap $\mathbb{E}\left[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2\right]$, based on these errors. To establish Lemma D.21, which quantifies the error between the approximated direction of the momentum-based hypergradient estimator, $\hat{\mathbf{d}}_t^{\mathbf{x}}$, and the true direction, $\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) + \nabla_{\mathbf{xy}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\,\mathbf{v}_t$, it is necessary to introduce Lemma D.19. This lemma quantifies the error between $\hat{\mathbf{d}}_t^{\mathbf{x}}$ and $\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) + \frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t))$. Then, Lemma D.22 bounds the projection mapping based on these errors. By combining these lemmas and setting parameters, we achieve the desired result.

D.1. Auxiliary Lemmas for Proof of Theorem 4.2

To solve the online bilevel problem in (17), we need to obtain the hyper-gradient of $f_{t,\boldsymbol{\rho}}$ in (17) at $(\mathbf{x},\mathbf{y})$ as

$$\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x},\hat{\mathbf{y}}_t^*(\mathbf{x})) := \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\hat{\mathbf{y}}_t^*(\mathbf{x})) + \nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\hat{\mathbf{y}}_t^*(\mathbf{x}))\,\mathbf{v}_t^*(\mathbf{x}), \quad \text{where}$$
$$\mathbf{v}_t^*(\mathbf{x}) := -[\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\hat{\mathbf{y}}_t^*(\mathbf{x}))]^{-1}\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\hat{\mathbf{y}}_t^*(\mathbf{x})).$$

Obtaining $\hat{\mathbf{y}}_t^*(\mathbf{x})$ in closed-form is usually a challenging task, so it is natural to use the following gradient surrogate. At any $(\mathbf{x},\mathbf{y})$, define:

$$\tilde{\nabla} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) := \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) + \nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})\,\hat{\mathbf{v}}_t^*(\mathbf{x}), \quad \text{where} \tag{119a}$$
$$\hat{\mathbf{v}}_t^*(\mathbf{x}) := -[\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})]^{-1}\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}). \tag{119b}$$

To do so, we also introduce $\mathbf{d}^{\mathbf{y}}_{t,\boldsymbol{\rho}}$, $\mathbf{d}^{\mathbf{v}}_{t,\boldsymbol{\rho}}$ and $\mathbf{d}^{\mathbf{x}}_{t,\boldsymbol{\rho}}$ as follows:

$$\mathbf{d}^{\mathbf{y}}_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) = \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}), \tag{120a}$$
$$\mathbf{d}^{\mathbf{v}}_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y},\mathbf{v}) = \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) + \nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})\,\mathbf{v}, \tag{120b}$$
$$\mathbf{d}^{\mathbf{x}}_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y},\mathbf{v}) = \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) + \nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})\,\mathbf{v}. \tag{120c}$$

To approximate these directions, we use (19)-(21). It can be shown that $\hat{\nabla}_{\mathbf{y}} f_t(\mathbf{x},\mathbf{y};\xi)$ and $\hat{\nabla}_{\mathbf{x}} f_t(\mathbf{x},\mathbf{y};\xi)$ are unbiased estimators of the true gradients $\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})$ and $\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})$ with respect to $\mathbf{y}$ and $\mathbf{x}$ (Flaxman et al., 2004), respectively, i.e.,

$$\mathbb{E}_{\mathbf{r}}\left[\hat{\nabla}_{\mathbf{y}} f_t(\mathbf{x},\mathbf{y};\xi)\right] = \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}), \quad \mathbb{E}_{\mathbf{z}}\left[\hat{\nabla}_{\mathbf{x}} f_t(\mathbf{x},\mathbf{y};\xi)\right] = \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}),$$

and,

$$\mathbb{E}_{\mathbf{r}}\left[\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y};\zeta)\right] = \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}), \quad \mathbb{E}_{\mathbf{z}}\left[\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x},\mathbf{y};\zeta)\right] = \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}). \tag{121}$$

Similarly,

$$\mathbb{E}_{\mathbf{r}}\left[\hat{\nabla}_{\mathbf{y}} f_t(\mathbf{x},\mathbf{y};\mathcal{B})\right] = \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}), \quad \mathbb{E}_{\mathbf{z}}\left[\hat{\nabla}_{\mathbf{x}} f_t(\mathbf{x},\mathbf{y};\mathcal{B})\right] = \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}),$$

and,

$$\mathbb{E}_{\mathbf{r}}\left[\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y};\bar{\mathcal{B}})\right] = \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}), \quad \mathbb{E}_{\mathbf{z}}\left[\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x},\mathbf{y};\bar{\mathcal{B}})\right] = \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}). \tag{122}$$

Lemma D.1. *Allen-Zhu & Li (2018, Lemma A.1.) Suppose Assumption B4. holds. Then, for any $\mathbf{x},\mathbf{v}\in\mathcal{X}$, we have:*

$$\left\|\nabla g_t(\mathbf{x}+\mathbf{v},\mathbf{y}+\mathbf{v}) - \nabla g_t(\mathbf{x},\mathbf{y}) - \nabla^2 g_t(\mathbf{x},\mathbf{y})\mathbf{v}\right\| \le \ell_{g,2}\left\|\mathbf{v}\right\|^2.$$

Lemma D.2. *Suppose that Assumptions 3.2 and 3.3 hold for all $\mathbf{x},\mathbf{x}'\in\mathcal{X}$, and $t\in[T]$, and that $\mathbf{d}^{\mathbf{x}}_{t,\boldsymbol{\rho}}$ and $\mathbf{d}^{\mathbf{v}}_{t,\boldsymbol{\rho}}$ are defined in (120). Then, we have*

$$\|\mathbf{d}^{\mathbf{x}}_{t,\boldsymbol{\rho}} - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x},\hat{\mathbf{y}}_t^*(\mathbf{x}))\|^2 \le M_f^2\left(\|\mathbf{y}-\hat{\mathbf{y}}_t^*(\mathbf{x})\|^2 + \|\mathbf{v}-\hat{\mathbf{v}}_t^*(\mathbf{x})\|^2\right), \tag{123a}$$
$$\left\|\mathbf{d}^{\mathbf{v}}_{t,\boldsymbol{\rho}}\right\|^2 \le M_{\mathbf{v}}^2\left(\|\mathbf{y}-\hat{\mathbf{y}}_t^*(\mathbf{x})\|^2 + \|\mathbf{v}-\hat{\mathbf{v}}_t^*(\mathbf{x})\|^2\right), \tag{123b}$$
$$\|\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x},\hat{\mathbf{y}}_t^*(\mathbf{x})) - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}',\hat{\mathbf{y}}_t^*(\mathbf{x}'))\| \le L_f\|\mathbf{x}-\mathbf{x}'\|, \tag{123c}$$
$$\|\hat{\mathbf{y}}_t^*(\mathbf{x}) - \hat{\mathbf{y}}_t^*(\mathbf{x}')\| \le L_{\mathbf{y}}\|\mathbf{x}-\mathbf{x}'\|, \tag{123d}$$
$$\|\hat{\mathbf{v}}_t^*(\mathbf{x}) - \hat{\mathbf{v}}_t^*(\mathbf{x}')\| \le L_{\mathbf{v}}\|\mathbf{x}-\mathbf{x}'\|. \tag{123e}$$

Here, $\hat{\mathbf{v}}_t^*(\mathbf{x})$ and $f_{t,\boldsymbol{\rho}}$, $\hat{\mathbf{y}}_t^*(\mathbf{x})$ are defined in (119b) and (17), respectively. Moreover, the constants M_f, $M_{\mathbf{v}}$, and $(L_{\mathbf{y}},L_{\mathbf{v}},L_f)$ are defined as in (40), (41), and (42), respectively.

Proof. We first show Eq. (123a).

Using Assumptions 3.2 and B1., we have $\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x})\right) \succeq \mu_g$, and

$$\|\hat{\mathbf{v}}^*_t(\mathbf{x})\| = \|\left(\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x})\right)\right)^{-1} \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x})\right)\| \leq \frac{\ell_{f,0}}{\mu_g}. \tag{124}$$

Observe that we have

$$\begin{aligned}
\|\mathbf{d}^{\mathbf{x}}_{t,\boldsymbol{\rho}} - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x}))\| &\leq \|\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x}))\| \\
&\quad + \|\mathbf{v}\nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) - \hat{\mathbf{v}}^*_t(\mathbf{x})\nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x})\right)\| \\
&\leq \|\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x}))\| \\
&\quad + \|\nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})\|\|\mathbf{v} - \hat{\mathbf{v}}^*_t(\mathbf{x})\| \\
&\quad + \|\hat{\mathbf{v}}^*_t(\mathbf{x})\|\|\nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) - \nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x}))\| \\
&\leq \left(\ell_{f,1} + \frac{\ell_{g,2}\ell_{f,0}}{\mu_g}\right)\|\mathbf{y} - \hat{\mathbf{y}}^*_t(\mathbf{x})\| + \ell_{g,1}\|\mathbf{v} - \hat{\mathbf{v}}^*_t(\mathbf{x})\| \\
&\leq M_f^2\left(\|\mathbf{y} - \hat{\mathbf{y}}^*_t(\mathbf{x})\| + \|\mathbf{v} - \hat{\mathbf{v}}^*_t(\mathbf{x})\|\right),
\end{aligned} \tag{125}$$

where M_f is defined as in (40); the third inequality is by Assumption 3.3 and the last inequality is by Eq. (124).

We now show Eq. (123b).
Since $\mathbf{d}^{\mathbf{v}*}_{t,\boldsymbol{\rho}} := \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x})) + \nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x})\right)\hat{\mathbf{v}}^*_t(\mathbf{x}) = 0$, we have

$$\begin{aligned}
\|\mathbf{d}^{\mathbf{v}}_{t,\boldsymbol{\rho}}\| &= \|\mathbf{d}^{\mathbf{v}}_{t,\boldsymbol{\rho}} - \mathbf{d}^{\mathbf{v}*}_{t,\boldsymbol{\rho}}\| \\
&= \|\mathbf{v}_t\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) + \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) \\
&\quad - \left(\hat{\mathbf{v}}^*_t(\mathbf{x})\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x})\right) + \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x}))\right)\| \\
&\leq \|\left(\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) - \nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x}))\right)\hat{\mathbf{v}}^*_t(\mathbf{x})\| \\
&\quad + \|\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})\left(\mathbf{v} - \hat{\mathbf{v}}^*_t(\mathbf{x})\right)\| \\
&\quad + \|\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}^*_t(\mathbf{x}))\|.
\end{aligned}$$

Then, from Assumption 3.3 and Eq. (124), we have

$$\begin{aligned}
\|\mathbf{d}^{\mathbf{v}}_{t,\boldsymbol{\rho}}\| &\leq \ell_{g,2}\|\mathbf{y} - \hat{\mathbf{y}}^*_t(\mathbf{x})\|\|\hat{\mathbf{v}}^*_t(\mathbf{x})\| + \ell_{g,1}\|\mathbf{v} - \hat{\mathbf{v}}^*_t(\mathbf{x})\| + \ell_{f,1}\|\mathbf{y} - \hat{\mathbf{y}}^*_t(\mathbf{x})\| \\
&\leq \left(\frac{\ell_{g,2}\ell_{f,0}}{\mu_g} + \ell_{f,1}\right)\|\mathbf{y} - \hat{\mathbf{y}}^*_t(\mathbf{x})\| + \ell_{g,1}\|\mathbf{v} - \hat{\mathbf{v}}^*_t(\mathbf{x})\| \\
&\leq M_{\mathbf{v}}\left(\|\mathbf{y} - \hat{\mathbf{y}}^*_t(\mathbf{x})\| + \|\mathbf{v} - \hat{\mathbf{v}}^*_t(\mathbf{x})\|\right),
\end{aligned}$$

where $M_{\mathbf{v}}$ is defined as in (41).

The proofs of Eqs. (123c)-(123e) follow from Tarzanagh et al. (2024, Lemma 17) by setting $(L_{\mathbf{y}}, L_{\mathbf{v}}, L_f)$ as in (42). □

D.2. Perturbation Bounds for OBO Objectives and Their Smoothing Variants

Lemma D.3. *Given $\boldsymbol{\rho} = (\rho_{\mathbf{s}}, \rho_{\mathbf{r}})$ as positive smoothing parameters, let $g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})$ and $f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})$ be the functions defined by* (17)*.*

(a) Suppose Assumption B3. holds. Then, we have

$$|g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) - g_t(\mathbf{x}, \mathbf{y})| \leq \frac{\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{2}. \tag{126}$$

(b) Suppose Assumption B2. holds. Then, we have

$$|f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) - f_t(\mathbf{x}, \mathbf{y})| \leq \frac{\ell_{f,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{2}. \tag{127}$$

Proof. Let B_1 and B_2 be the unit ball in $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, respectively. Let $\mathcal{V}(d_1)$ and $\mathcal{V}(d_2)$ be volume of the unit ball in $\mathbb{R}^{d_1}$

and $\mathbb{R}^{d_2}$, respectively. Then, we have

$$\begin{aligned}
&|g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - g_t(\mathbf{x},\mathbf{y})| \\
&= \left| \frac{1}{\mathcal{V}(d_1)\mathcal{V}(d_2)} \int_{B_1}\int_{B_2} \left(g_t(\mathbf{x}+\rho_{\mathbf{s}}\mathbf{s}, \mathbf{y}+\rho_{\mathbf{r}}\mathbf{r}) - g_t(\mathbf{x},\mathbf{y})\right) d\mathbf{s} d\mathbf{r} \right| \\
&= \left| \frac{1}{\mathcal{V}(d_1)\mathcal{V}(d_2)} \int_{B_1}\int_{B_2} \left(g_t(\mathbf{x}+\rho_{\mathbf{s}}\mathbf{s}, \mathbf{y}+\rho_{\mathbf{r}}\mathbf{r}) - g_t(\mathbf{x},\mathbf{y}) - \langle \nabla g_t(\mathbf{x},\mathbf{y}), (\rho_{\mathbf{s}}\mathbf{s}, \rho_{\mathbf{r}}\mathbf{r})\rangle\right) d\mathbf{s} d\mathbf{r} \right|.
\end{aligned}$$

Thus, we get

$$\begin{aligned}
&|g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - g_t(\mathbf{x},\mathbf{y})| \\
&\le \int_{B_1}\int_{B_2} |g_t(\mathbf{x}+\rho_{\mathbf{s}}\mathbf{s}, \mathbf{y}+\rho_{\mathbf{r}}\mathbf{r}) - g_t(\mathbf{x},\mathbf{y}) - \langle \nabla g_t(\mathbf{x},\mathbf{y}), (\rho_{\mathbf{s}}\mathbf{s}, \rho_{\mathbf{r}}\mathbf{r})\rangle| \, d\mathbf{s} d\mathbf{r} \\
&\le \int_{B_1}\int_{B_2} \frac{\ell_{g,1}}{2}\left(\rho_{\mathbf{s}}^2\|\mathbf{s}\|^2 + \rho_{\mathbf{r}}^2\|\mathbf{r}\|^2\right) d\mathbf{s} d\mathbf{r} \\
&= \frac{\ell_{g,1}\rho_{\mathbf{s}}^2}{2}\int_{B_1}\|\mathbf{s}\|^2 d\mathbf{s} + \frac{\ell_{g,1}\rho_{\mathbf{r}}^2}{2}\int_{B_2}\|\mathbf{r}\|^2 d\mathbf{r} \\
&= \frac{\ell_{g,1}\rho_{\mathbf{s}}^2}{2}\frac{d_1}{d_1+2} + \frac{\ell_{g,1}\rho_{\mathbf{r}}^2}{2}\frac{d_2}{d_2+2} \\
&\le \frac{\ell_{g,1}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)}{2},
\end{aligned}$$

where the last equality follows since $\frac{1}{\mathcal{V}(d)}\int_{s\in B}\|s\|^p ds = \frac{d}{d+p}$.

The proof of part (b) follows using similar arguments. $\square$

Lemma D.4. *Given $\boldsymbol{\rho} = (\rho_{\mathbf{s}}, \rho_{\mathbf{r}})$ as positive smoothing parameters, let $g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})$ and $f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})$ be the functions defined by (17).*

(a) Suppose Assumption B3. holds. Then, we have

$$\|\nabla g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - \nabla g_t(\mathbf{x},\mathbf{y})\| \le \frac{\ell_{g,1}(\rho_{\mathbf{s}}d_1 + \rho_{\mathbf{r}}d_2)}{2}. \tag{128}$$

(b) Suppose Assumption B2. holds. Then, we have

$$\|\nabla f_t(\mathbf{x},\mathbf{y}) - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})\| \le \frac{\ell_{f,1}(\rho_{\mathbf{s}}d_1 + \rho_{\mathbf{r}}d_2)}{2}. \tag{129}$$

Proof. Let $S(d_1)$ be the surface area of the unit sphere in $\mathbb{R}^{d_1}$. Moreover, let U_{B_1} be the unit sphere.

$$
\begin{aligned}
&\|\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} g_t(\mathbf{x},\mathbf{y})\| \\
&= \left\| \frac{1}{S(d_1)} \left(\frac{d_1}{\rho_{\mathbf{s}}} \int_{U_{B_1}} g_t(\mathbf{x}+\rho_{\mathbf{s}}\mathbf{s},\mathbf{y})\mathbf{s}d\mathbf{s} \right) - \nabla_{\mathbf{x}} g_t(\mathbf{x},\mathbf{y}) \right\| \\
&= \left\| \frac{1}{S(d_1)} \left(\frac{d_1}{\rho_{\mathbf{s}}} \int_{U_{B_1}} g_t(\mathbf{x}+\rho_{\mathbf{s}}\mathbf{s},\mathbf{y})\mathbf{s}d\mathbf{s} - \int_{U_{B_1}} \frac{d_1}{\rho_{\mathbf{s}}} g_t(\mathbf{x},\mathbf{y})\mathbf{s}d\mathbf{s} \right.\right. \\
&\left.\left. - \int_{U_{B_1}} \frac{d_1}{\rho_{\mathbf{s}}} \langle \nabla_{\mathbf{x}} g_t(\mathbf{x},\mathbf{y}), \rho_{\mathbf{s}}\mathbf{s}\rangle \mathbf{s}d\mathbf{s} \right) \right\| \\
&\leq \frac{d_1}{S(d_1)\rho_{\mathbf{s}}} \int_{U_{B_1}} \left| g_t(\mathbf{x}_t+\rho_{\mathbf{s}}\mathbf{s},\mathbf{y}) - g_t(\mathbf{x},\mathbf{y}) - \langle \nabla_{\mathbf{x}} g_t(\mathbf{x},\mathbf{y}), \rho_{\mathbf{s}}\mathbf{s}\rangle \right| \|\mathbf{s}\| \, d\mathbf{s} \\
&\leq \frac{d_1}{S(d_1)\rho_{\mathbf{s}}} \cdot \frac{\ell_{g,1}\rho_{\mathbf{s}}^2}{2} \int_{U_{B_1}} \|\mathbf{s}\|^3 \, d\mathbf{s} \\
&= \frac{\rho_{\mathbf{s}} d_1 \ell_{g,1}}{2},
\end{aligned} \tag{130}
$$

where the second equality follows from $\int_{U_{B_1}} \mathbf{s}\mathbf{s}^\top d\mathbf{s} = \frac{S(d_1)}{d_1}\mathbf{I}$.

Similarly, let $S(d_2)$ be the surface area of the unit sphere in $\mathbb{R}^{d_2}$. Moreover, let U_{B_2} be the unit sphere.

$$
\begin{aligned}
&\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y})\| \\
&= \left\| \frac{1}{S(d_2)} \left(\frac{d_2}{\rho_{\mathbf{r}}} \int_{U_{B_2}} g_t(\mathbf{x},\mathbf{y}+\rho_{\mathbf{r}}\mathbf{r})\mathbf{r}d\mathbf{r} \right) - \nabla_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y}) \right\| \\
&= \left\| \frac{1}{S(d_2)} \left(\frac{d_2}{\rho_{\mathbf{r}}} \int_{U_{B_2}} g_t(\mathbf{x},\mathbf{y}+\rho_{\mathbf{r}}\mathbf{r})\mathbf{r}d\mathbf{r} - \int_{U_{B_2}} \frac{d_2}{\rho_{\mathbf{r}}} g_t(\mathbf{x},\mathbf{y})\mathbf{r}d\mathbf{r} \right.\right. \\
&\left.\left. - \int_{U_{B_2}} \frac{d_2}{\rho_{\mathbf{r}}} \langle \nabla_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y}), \rho_{\mathbf{r}}\mathbf{r}\rangle \mathbf{r}d\mathbf{r} \right) \right\| \\
&\leq \frac{d_2}{S(d_2)\rho_{\mathbf{r}}} \int_{U_{B_2}} \left| g_t(\mathbf{x}_t,\mathbf{y}+\rho_{\mathbf{r}}\mathbf{r}) - g_t(\mathbf{x},\mathbf{y}) - \langle \nabla_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y}), \rho_{\mathbf{r}}\mathbf{r}\rangle \right| \|\mathbf{r}\| \, d\mathbf{r} \\
&\leq \frac{d_2}{S(d_2)\rho_{\mathbf{r}}} \cdot \frac{\ell_{g,1}\rho_{\mathbf{r}}^2}{2} \int_{U_{B_2}} \|\mathbf{r}\|^3 \, d\mathbf{r} \\
&= \frac{\rho_{\mathbf{r}} d_2 \ell_{g,1}}{2},
\end{aligned} \tag{131}
$$

where the second equality follows from $\int_{U_{B_2}} \mathbf{r}\mathbf{r}^\top d\mathbf{r} = \frac{S(d_2)}{d_2}\mathbf{I}$.

Thus, we get

$$
\begin{aligned}
&\|\nabla g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - \nabla g_t(\mathbf{x},\mathbf{y})\| \\
&\leq \|\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} g_t(\mathbf{x},\mathbf{y})\| + \|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y})\| \\
&\leq \frac{\rho_{\mathbf{s}} d_1 \ell_{g,1}}{2} + \frac{\rho_{\mathbf{r}} d_2 \ell_{g,1}}{2}.
\end{aligned}
$$

Finally, by a similar argument as in Part (a), we obtain

$$
\|\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{x}} f_t(\mathbf{x},\mathbf{y})\| \leq \frac{\rho_{\mathbf{s}} d_1 \ell_{f,1}}{2}, \tag{132}
$$

and

$$\left\|\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}\left(\mathbf{x},\mathbf{y}\right)-\nabla_{\mathbf{y}} f_{t}\left(\mathbf{x},\mathbf{y}\right)\right\| \leq \frac{\rho_{\mathbf{r}} d_2 \ell_{f,1}}{2}, \tag{133}$$

which implies

$$\left\|\nabla f_{t,\boldsymbol{\rho}}\left(\mathbf{x},\mathbf{y}\right)-\nabla f_{t}\left(\mathbf{x},\mathbf{y}\right)\right\| \leq \frac{(\rho_{\mathbf{s}} d_1+\rho_{\mathbf{r}} d_2)\ell_{f,1}}{2}.$$

□

Lemma D.5. *Suppose Assumption B4. holds. Given $\boldsymbol{\rho}=(\rho_{\mathbf{s}},\rho_{\mathbf{r}})$ as positive smoothing parameters, let $g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})$ be the function defined in (17). Then, we have*

$$\left\|\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x},\mathbf{y}\right)-\nabla^2_{\mathbf{y}} g_{t}\left(\mathbf{x},\mathbf{y}\right)\right\|^2 \leq \frac{d_2^2\ell_{g,2}^2}{4}\rho_{\mathbf{r}}^2.$$

Proof. Similary, let $S(d_2)$ be the surface area of the unit sphere in $\mathbb{R}^{d_2}$. Moreover, let U_{B_2} be the unit sphere.

$$\begin{aligned}
&\left\|\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x},\mathbf{y}\right)-\nabla^2_{\mathbf{y}} g_{t}\left(\mathbf{x},\mathbf{y}\right)\right\| \\
&=\left\|\frac{1}{S(d_2)}\left(\frac{d_2}{\rho_{\mathbf{r}}}\int_{U_{B_2}}\nabla_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y}+\rho_{\mathbf{r}}\mathbf{r})\mathbf{r}d\mathbf{r}\right)-\nabla^2_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y})\right\| \\
&=\left\|\frac{1}{S(d_2)}\left(\frac{d_2}{\rho_{\mathbf{r}}}\int_{U_{B_2}}\nabla_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y}+\rho_{\mathbf{r}}\mathbf{r})\mathbf{r}d\mathbf{r}-\int_{U_{B_2}}\frac{d_2}{\rho_{\mathbf{r}}}\nabla_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y})\mathbf{r}d\mathbf{r}\right.\right. \\
&\left.\left.-\int_{U_{B_2}}\frac{d_2}{\rho_{\mathbf{r}}}\langle\nabla^2_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y}),\rho_{\mathbf{r}}\mathbf{r}\rangle\mathbf{r}d\mathbf{r}\right)\right\| \\
&\leq \frac{d_2}{S(d_2)\rho_{\mathbf{r}}}\int_{U_{B_2}}\left|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}+\rho_{\mathbf{r}}\mathbf{r})-\nabla_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y})-\langle\nabla^2_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y}),\rho_{\mathbf{r}}\mathbf{r}\rangle\right|\|\mathbf{r}\|\,d\mathbf{r} \\
&\leq \frac{d_2}{S(d_2)\rho_{\mathbf{r}}}\cdot\frac{\ell_{g,2}\rho_{\mathbf{r}}^2}{2}\int_{U_{B_2}}\|\mathbf{r}\|^3\,d\mathbf{r} \\
&=\frac{\rho_{\mathbf{r}} d_2\ell_{g,2}}{2},
\end{aligned}$$

where the second equality follows from $\int_{U_{B_2}}\mathbf{r}\mathbf{r}^\top d\mathbf{r}=\frac{S(d_2)}{d_2}\mathbf{I}$. □

Lemma D.6. *Suppose Assumption B3. holds. Let $\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y};\bar{\mathcal{B}})$ and $\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x},\mathbf{y};\bar{\mathcal{B}})$ be defined as in (19a) and (19b), respectively. Then, for any $(\mathbf{x},\mathbf{y})\in\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}$ and $\rho_{\mathbf{r}},\rho_{\mathbf{s}}\geq 0$, we have*

$$\mathbb{E}_{\mathbf{r}}\left[\|\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y};\bar{\mathcal{B}})-\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x},\acute{\mathbf{y}};\bar{\mathcal{B}})\|^2\right]\leq 3d_2\ell_{g,1}^2\|\mathbf{y}-\acute{\mathbf{y}}\|^2+\frac{3\ell_{g,1}^2 d_2^2\rho_{\mathbf{r}}^2}{2}\quad\forall\acute{\mathbf{y}}\in\mathbb{R}^{d_2}, \tag{134a}$$

$$\mathbb{E}_{\mathbf{z}}\left[\|\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x},\mathbf{y};\bar{\mathcal{B}})-\hat{\nabla}_{\mathbf{x}} g_t(\acute{\mathbf{x}},\mathbf{y};\bar{\mathcal{B}})\|^2\right]\leq 3d_1\ell_{g,1}^2\|\mathbf{x}-\acute{\mathbf{x}}\|^2+\frac{3\ell_{g,1}^2 d_1^2\rho_{\mathbf{s}}^2}{2}\quad\forall\acute{\mathbf{x}}\in\mathbb{R}^{d_1}. \tag{134b}$$

Proof. The proof is similar to that of Lemma 5 in (Ji et al., 2019). □

Lemma D.7. *Suppose Assumptions 3.2 and B3. hold. Let $\boldsymbol{\rho}=(\rho_{\mathbf{s}},\rho_{\mathbf{r}})$ be positive smoothing parameters. Let $\mathbf{y}_t^*(\mathbf{x})$ and $\hat{\mathbf{y}}_t^*(\mathbf{x})$ be defined in (1) and (18), respectively. Then, we have*

$$\mathbb{E}\left[\|\hat{\mathbf{y}}_t^*(\mathbf{x})-\mathbf{y}_t^*(\mathbf{x})\|^2\right]\leq\frac{\ell_{g,1}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)}{\mu_g}. \tag{135}$$

Proof. From (1), we have $\mathbf{y}_t^*(\mathbf{x})\in\arg\min_{\mathbf{y}\in\mathbb{R}^{d_2}} g_t(\mathbf{x},\mathbf{y})$. Since by Assumption 3.2, $g_t\left(\mathbf{x},\mathbf{y}\right)$ is μ_g-strongly convex in terms of $\mathbf{y}$. Then, by Lemma B.2, we get

$$\|\mathbf{y} - \mathbf{y}_t^*(\mathbf{x})\|^2 \leq \frac{2}{\mu_g}\left(g_t\left(\mathbf{x}, \mathbf{y}\right) - g_t\left(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})\right)\right).$$

By setting $\mathbf{y} = \hat{\mathbf{y}}_t^*(\mathbf{x})$, we have

$$\|\hat{\mathbf{y}}_t^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2 \leq \frac{2}{\mu_g}\left(g_t\left(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})\right) - g_t\left(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})\right)\right). \tag{136}$$

Similarly, from (18), we have

$$\hat{\mathbf{y}}_t^*(\mathbf{x}) \in \arg\min_{\mathbf{y}\in\mathbb{R}^{d_2}} \left\{g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{(\mathbf{z},\mathbf{r})}\left[g_t(\mathbf{x} + \rho_{\mathbf{s}}\mathbf{s}, \mathbf{y} + \rho_{\mathbf{r}}\mathbf{r}; \zeta)\right]\right\}.$$

Since by Assumption 3.2, $g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \mathbf{y}\right)$ is μ_g-strongly convex in terms of $\mathbf{y}$. Then, from Lemma B.2, we have

$$\|\mathbf{y} - \hat{\mathbf{y}}_t^*(\mathbf{x})\|^2 \leq \frac{2}{\mu_g}\left(g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \mathbf{y}\right) - g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})\right)\right).$$

By setting $\mathbf{y} = \mathbf{y}_t^*(\mathbf{x})$, we have

$$\|\mathbf{y}_t^*(\mathbf{x}) - \hat{\mathbf{y}}_t^*(\mathbf{x})\|^2 \leq \frac{2}{\mu_g}\left(g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) - g_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))\right). \tag{137}$$

Summing up (136) and (137), we get

$$\begin{aligned}\|\mathbf{y}_t^*(\mathbf{x}) - \hat{\mathbf{y}}_t^*(\mathbf{x})\|^2 \leq & \frac{1}{\mu_g}\left(g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) - g_t\left(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})\right)\right) \\ & + \frac{1}{\mu_g}\left(g_t\left(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})\right) - g_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))\right),\end{aligned}$$

which implies

$$\begin{aligned}\|\mathbf{y}_t^*(\mathbf{x}) - \hat{\mathbf{y}}_t^*(\mathbf{x})\|^2 \leq & \frac{1}{\mu_g}\left|g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) - g_t\left(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})\right)\right| \\ & + \frac{1}{\mu_g}\left|g_t\left(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})\right) - g_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))\right| \\ \leq & \frac{\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g},\end{aligned}$$

where the last inequality is by Eq. (126). □

Lemma D.8. *Suppose Assumptions 3.2 and 3.3 hold. Let $\mathbf{v}_t^*(\mathbf{x})$ and $\hat{\mathbf{v}}_t^*(\mathbf{x})$ be defined in (4b) and (119b), respectively. Then, we have*

$$\mathbb{E}\left[\|\hat{\mathbf{v}}_t^*(\mathbf{x}) - \mathbf{v}_t^*(\mathbf{x})\|^2\right] \leq \frac{d_2^2}{2\mu_g^4}\left(\ell_{f,1}^2\mu_g^2 + \ell_{g,2}^2\ell_{f,0}^2\right)\rho_{\mathbf{r}}^2. \tag{138}$$

Proof. From (4b) and (119b), we have

$$\begin{aligned}&\|\hat{\mathbf{v}}_t^*(\mathbf{x}) - \mathbf{v}_t^*(\mathbf{x})\|^2 \\ &= \left\|[\nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \mathbf{y}\right)]^{-1}\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) - [\nabla_{\mathbf{y}}^2 g_t\left(\mathbf{x}, \mathbf{y}\right)]^{-1}\nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y})\right\|^2 \\ &\leq 2\left\|[\nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \mathbf{y}\right)]^{-1}\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) - [\nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \mathbf{y}\right)]^{-1}\nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y})\right\|^2 \qquad (139a) \\ &+ 2\left\|[\nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}\left(\mathbf{x}, \mathbf{y}\right)]^{-1}\nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}) - [\nabla_{\mathbf{y}}^2 g_t\left(\mathbf{x}, \mathbf{y}\right)]^{-1}\nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y})\right\|^2. \qquad (139b)\end{aligned}$$

Next, we separately bound each of the above terms, (139a) and (139b).

$$
\begin{aligned}
(139a) &\leq 2 \left\| [\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})]^{-1} \right\|^2 \|\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} f_t(\mathbf{x},\mathbf{y})\|^2 \\
&\leq \frac{2}{\mu_g^2} \|\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - \nabla_{\mathbf{y}} f_t(\mathbf{x},\mathbf{y})\|^2 \\
&\leq \frac{2}{\mu_g^2} \frac{\rho_{\mathbf{r}}^2 d_2^2 \ell_{f,1}^2}{4},
\end{aligned}
\tag{140}
$$

where the second inequality holds due to the Assumption 3.2, the third inequality is by (133).

To bound (139b), note that for any invertible matrices $\mathbf{A}_1$ and $\mathbf{A}_2$, we have

$$
\|\mathbf{A}_2^{-1} - \mathbf{A}_1^{-1}\| = \|\mathbf{A}_1^{-1}(\mathbf{A}_1 - \mathbf{A}_2)\mathbf{A}_2^{-1}\| \leq \|\mathbf{A}_1^{-1}\|\|\mathbf{A}_2^{-1}\|\|\mathbf{A}_1 - \mathbf{A}_2\|,
$$

which implies

$$
\begin{aligned}
(139b) &\leq 2 \left\| [\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})]^{-1} - [\nabla^2_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y})]^{-1} \right\|^2 \|\nabla_{\mathbf{y}} f_t(\mathbf{x},\mathbf{y})\|^2 \\
&\leq 2 \left\| [\nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y})]^{-1} \right\|^2 \left\| [\nabla^2_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y})]^{-1} \right\|^2 \\
&\left\| \nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x},\mathbf{y}) - \nabla^2_{\mathbf{y}} g_t(\mathbf{x},\mathbf{y}) \right\|^2 \|\nabla_{\mathbf{y}} f_t(\mathbf{x},\mathbf{y})\|^2 \\
&\leq \frac{2}{\mu_g^4} \frac{\rho_{\mathbf{r}}^2 d_2^2 \ell_{g,2}^2}{4} \ell_{f,0}^2,
\end{aligned}
\tag{141}
$$

where the last inequality follows from Lemma D.5.

Using (139)– (140), we obtain the desired result. $\square$

D.3. Bounds on the Zeroth-Order Inner Solution

Lemma D.9. *Suppose that Assumptions B3. and D1. hold. Consider the sequence $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 2, and define*

$$
e_t^{g_\rho} := \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \hat{\mathbf{d}}_t^{\mathbf{y}}. \tag{142}
$$

Then, we have

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^{g_\rho}\|^2 \leq{}& (1-\gamma_{t+1})^2 \mathbb{E}\|e_t^{g_\rho}\|^2 + 12(1-\gamma_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_{t-1}(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2 \\
&+ 9 d_2^2 \ell_{g,1}^2 (1-\gamma_{t+1})^2 \rho_{\mathbf{r}}^2 + 24 d_2 \ell_{g,1}^2 (1-\gamma_{t+1})^2 \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
&+ 24 d_2 \ell_{g,1}^2 (1-\gamma_{t+1})^2 \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 + 2\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}} \gamma_{t+1}^2.
\end{aligned}
\tag{143}
$$

Proof. From the definition of $\hat{\mathbf{d}}_{t+1}^{\mathbf{y}}$ in Algorithm 2, we have

$$
\begin{aligned}
\hat{\mathbf{d}}_{t+1}^{\mathbf{y}} - \hat{\mathbf{d}}_t^{\mathbf{y}} ={}& -\gamma_{t+1} \hat{\mathbf{d}}_t^{\mathbf{y}} + \gamma_{t+1} \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}; \bar{\mathcal{B}}_{t+1}) \\
&+ (1-\gamma_{t+1}) \left(\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t; \bar{\mathcal{B}}_{t+1}) \right).
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{y}}_{t+1}\|^2 \\
&= \mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{y}}_{t} - (\hat{\mathbf{d}}^{\mathbf{y}}_{t+1} - \hat{\mathbf{d}}^{\mathbf{y}}_{t})\|^2 \\
&= \mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{y}}_{t} + \gamma_{t+1}\hat{\mathbf{d}}^{\mathbf{y}}_{t} - \gamma_{t+1}\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}) \\
&- (1-\gamma_{t+1})\left(\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t},\mathbf{y}_{t};\bar{\mathcal{B}}_{t+1})\right)\|^2 \\
&= \mathbb{E}\|(1-\gamma_{t+1})(\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) - \hat{\mathbf{d}}^{\mathbf{y}}_{t}) \\
&+ \gamma_{t+1}(\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1})) \\
&+ (1-\gamma_{t+1})\,(\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) \\
&+\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) - \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) \\
&-\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}) + \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t},\mathbf{y}_{t};\bar{\mathcal{B}}_{t+1})\Big)\|^2.
\end{aligned}
$$

From (122), we have

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1})\right] = \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}), \\
&\mathbb{E}\left[\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t},\mathbf{y}_{t};\bar{\mathcal{B}}_{t+1})\right] \\
&= \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}),
\end{aligned}
$$

then, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{y}}_{t+1}\|^2 \\
&= (1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) - \hat{\mathbf{d}}^{\mathbf{y}}_{t}\|^2 \\
&+ \mathbb{E}\|\gamma_{t+1}(\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1})) \\
&+ (1-\gamma_{t+1})\,(\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) + \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) - \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) \\
&-\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}) + \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t},\mathbf{y}_{t};\bar{\mathcal{B}}_{t+1})\Big)\|^2 \\
&\le (1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) - \hat{\mathbf{d}}^{\mathbf{y}}_{t}\|^2 \\
&+ 2(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) + \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) \\
&- \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}) + \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t},\mathbf{y}_{t};\bar{\mathcal{B}}_{t+1})\|^2 \\
&+ 2\gamma_{t+1}^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1})\|^2,
\end{aligned}
$$

where the second inequality holds by Cauchy-Schwarz inequality.
Then, from $\mathbb{E}\|a-\mathbb{E}[a]\|^2 = \mathbb{E}\|a\|^2 - \|\mathbb{E}[a]\|^2$ and Assumption 4.1, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{y}}_{t+1}\|^2 \\
&\le (1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) - \hat{\mathbf{d}}^{\mathbf{y}}_{t}\|^2 \\
&+ 4(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t})\|^2 \\
&+ 4(1-\gamma_{t+1})^2\mathbb{E}\|\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1};\bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t},\mathbf{y}_{t};\bar{\mathcal{B}}_{t+1})\|^2 + 2\gamma_{t+1}^2\frac{\hat{\sigma}^2_{g_{\mathbf{y}}}}{\bar{b}} \\
&\le (1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) - \hat{\mathbf{d}}^{\mathbf{y}}_{t}\|^2 \\
&+ 4(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t}) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_{t},\mathbf{y}_{t})\|^2 \\
&+ 12(1-\gamma_{t+1})^2 d_2\ell_{g,1}^2\mathbb{E}\|(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - (\mathbf{x}_{t},\mathbf{y}_{t})\|^2 \\
&+ 3(1-\gamma_{t+1})^2\ell_{g,1}^2 d_2^2\rho_{\mathbf{r}}^2 + 2\gamma_{t+1}^2\frac{\hat{\sigma}^2_{g_{\mathbf{y}}}}{\bar{b}},
\end{aligned}
$$

where the second inequality follows from Young's inequality and Lemma D.6.

From Eq. (131), we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\|^2 \\
&\leq 3\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t)\|^2 \\
&+ 3\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{y}} g_{t}(\mathbf{x}_t,\mathbf{y}_t)\|^2 \\
&+ 3\mathbb{E}\|\nabla_{\mathbf{y}} g_{t}(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\|^2 \\
&\leq 3\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{y}} g_{t}(\mathbf{x}_t,\mathbf{y}_t)\|^2 + \frac{3\rho_{\mathbf{r}}^2 d_2^2 \ell_{g,1}^2}{2}.
\end{aligned}
$$

Finally, we get

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \hat{\mathbf{d}}_{t+1}^{\mathbf{y}}\|^2 \leq (1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t) - \hat{\mathbf{d}}_t^{\mathbf{y}}\|^2 \\
&+ 12(1-\gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t)\|^2 + 6(1-\gamma_{t+1})^2\rho_{\mathbf{r}}^2 d_2^2 \ell_{g,1}^2 \\
&+ 12(1-\gamma_{t+1})^2 d_2 \ell_{g,1}^2 \mathbb{E}\|(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - (\mathbf{x}_t,\mathbf{y}_t)\|^2 + 3(1-\gamma_{t+1})^2\ell_{g,1}^2 d_2^2 \rho_{\mathbf{r}}^2 + 2\gamma_{t+1}^2\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b}.
\end{aligned}
$$

$\square$

Lemma D.10. *Suppose Assumptions 3.2 and B3. hold. Then, for the sequence $\{(\mathbf{x}_t,\mathbf{y}_t)\}_{t=1}^T$ generated by Algorithm 2, we have*

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] &\leq (1+a)\left(1 - 2\beta_t\frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right)\mathbb{E}\left[\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] \\
&+ \left(-(1+a)\left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right)\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\|^2\right] \\
&+ (1+\frac{1}{a})\beta_t^2\mathbb{E}\left[\|e_t^{g_{\boldsymbol{\rho}}}\|^2\right],
\end{aligned}
$$

where $a>0$ is a constant, $e_t^{g_{\boldsymbol{\rho}}}$ is defined in (142), and $\hat{\mathbf{y}}_t^*(\mathbf{x}_t)$ is defined in (18).

Proof. From Lemma B.5, we have

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] &= \mathbb{E}\left[\|\mathbf{y}_t - \beta_t\hat{\mathbf{d}}_t^{\mathbf{y}} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] \\
&\leq (1+a)\mathbb{E}\left[\|\mathbf{y}_t - \beta_t\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] \\
&+ (1+\frac{1}{a})\beta_t^2\mathbb{E}\left[\|\hat{\mathbf{d}}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\|^2\right].
\end{aligned} \tag{144}
$$

Next, we will separately bound the first term on the RHS of the above inequality.
We have

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_t - \beta_t\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] &= \mathbb{E}\left[\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] + \beta_t^2\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\|^2\right] \\
&- 2\beta_t\mathbb{E}\left[\langle\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t), \mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\rangle\right] \\
&\leq \left(1 - 2\beta_t\frac{\mu_g\ell_{g,1}}{\mu_g + \ell_{g,1}}\right)\mathbb{E}\left[\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] \\
&- \left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\|^2\right],
\end{aligned} \tag{145}
$$

where the inequality results from the strong convexity of $g_{t,\boldsymbol{\rho}}$ by Assumption 3.2, which implies

$$
\langle\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t), \mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\rangle \geq \frac{\mu_g\ell_{g,1}}{\mu_g + \ell_{g,1}}\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2 + \frac{1}{\mu_g + \ell_{g,1}}\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\|^2.
$$

Substituting (145) into (144), gives the desired result.

$\square$

For notational brevity in the analysis, we define

$$\hat{\theta}_t^{\mathbf{y}} := \|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2, \quad \hat{\theta}_t^{\mathbf{v}} := \|\mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2, \tag{146}$$

where $\hat{\mathbf{y}}_t^*(\mathbf{x})$ and $\hat{\mathbf{v}}_t^*(\mathbf{x})$ are defined in (18) and (119b), respectively.

Lemma D.11. *Suppose Assumptions 3.2, B2. and B3. hold. Let $\hat{\theta}_t^{\mathbf{y}}$ be defined in (146). Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$ generated by Algorithm 2 guarantees the following bound:*

$$\begin{aligned}
&\sum_{t=1}^T \left(\mathbb{E}[\hat{\theta}_{t+1}^{\mathbf{y}}] - \mathbb{E}[\hat{\theta}_t^{\mathbf{y}}]\right) \\
&\leq \left(-\frac{L_{\mu_g}}{2}\sum_{t=1}^T \mathbb{E}[\hat{\theta}_t^{\mathbf{y}}] + \frac{2}{L_{\mu_g}}\sum_{t=1}^T \mathbb{E}\left[\|e_t^{g_{\boldsymbol{\rho}}}\|^2\right]\right)\beta_t + \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}}\sum_{t=1}^T \mathbb{E}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \frac{1}{\beta_t} \\
&+ \sum_{t=1}^T \left(\frac{24\ell_{g,1}}{L_{\mu_g}\mu_g}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2) + \frac{12}{L_{\mu_g}}\sup_{\mathbf{x}\in\mathcal{X}}\|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2\right)\frac{1}{\beta_t} \\
&+ \sum_{t=1}^T \left(-\frac{2\beta_t}{\mu_g + \ell_{g,1}} + \beta_t^2\right)\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2\right],
\end{aligned} \tag{147}$$

where $L_{\mu_g} = \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}$, and $L_{\mathbf{y}} = \frac{\ell_{g,1}}{\mu_g}$ is defined as in (42).

Proof. From Lemma B.5, we have for any $c > 0$

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1})\|^2\right] &= \mathbb{E}\left[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t) + \hat{\mathbf{y}}_t^*(\mathbf{x}_t) - \hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1})\|^2\right] \\
&\leq (1+c)\,\mathbb{E}\left[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] \\
&+ \left(1 + \frac{1}{c}\right)\mathbb{E}\left[\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right].
\end{aligned} \tag{148}$$

From Lemma D.10, we have for any $a > 0$

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] &\leq (1+a)\left(1 - 2\beta_t\frac{\mu_g\ell_{g,1}}{\mu_g + \ell_{g,1}}\right)\mathbb{E}\left[\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] \\
&+ \left(-(1+a)\left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right)\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2\right] \\
&+ \left(1 + \frac{1}{a}\right)\beta_t^2\mathbb{E}\left[\|e_t^{g_{\boldsymbol{\rho}}}\|^2\right].
\end{aligned} \tag{149}$$

Substituting (149) into (148), we get

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1})\|^2\right] \\
&\leq (1+c)(1+a)\left(1 - 2\beta_t\frac{\mu_g\ell_{g,1}}{\mu_g + \ell_{g,1}}\right)\mathbb{E}\left[\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right] \\
&+ \left(-(1+c)(1+a)\left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right)\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2\right] \\
&+ (1+c)(1+\frac{1}{a})\beta_t^2\mathbb{E}\left[\|e_t^{g_{\boldsymbol{\rho}}}\|^2\right] + \left(1 + \frac{1}{c}\right)\mathbb{E}\left[\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\right].
\end{aligned} \tag{150}$$

Choose $c = \frac{\beta_t L_{\mu_g}/2}{1-\beta_t L_{\mu_g}}$ and $a = \frac{\beta_t L_{\mu_g}}{1-2\beta_t L_{\mu_g}}$. Then, the following equations and inequalities are satisfied.

$$
\begin{aligned}
&(1+c)(1+a)\left(1-2\beta_t L_{\mu_g}\right) = 1 - \frac{\beta_t L_{\mu_g}}{2},\\
&(1+a)\left(1-2\beta_t L_{\mu_g}\right) = 1 - \beta_t L_{\mu_g},\\
&(1+c)\left(1-\beta_t L_{\mu_g}\right) = 1 - \frac{\beta_t L_{\mu_g}}{2},\\
&1+\frac{1}{a} \le \frac{1}{\beta_t L_{\mu_g}}, \quad 1+\frac{1}{c} \le \frac{2}{\beta_t L_{\mu_g}},
\end{aligned} \tag{151}
$$

where $L_{\mu_g} = \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}$. Based on (150) and (151), we get

$$
\begin{aligned}
&\mathbb{E}\left[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_{t+1})\|^2\right] - \mathbb{E}\left[\|\mathbf{y}_t - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2\right]\\
&\le -\frac{\beta_t L_{\mu_g}}{2}\mathbb{E}\left[\|\mathbf{y}_t - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2\right] + \left(-\left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right)\mathbb{E}\left[\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2\right]\\
&+ \frac{2}{\beta_t L_{\mu_g}}\beta_t^2 \mathbb{E}\left[\|e_t^{g_{\boldsymbol{\rho}}}\|^2\right] + \frac{2}{\beta_t L_{\mu_g}}\mathbb{E}\left[\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2\right].
\end{aligned} \tag{152}
$$

Next, we upper-bound the last term of the above inequality.

$$
\begin{aligned}
&\mathbb{E}\left[\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2\right]\\
&\le 2\left(\mathbb{E}\left[\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_t)\|^2\right] + \mathbb{E}\left[\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_t) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2\right]\right)\\
&\le 2\left(L_{\mathbf{y}}^2 \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_t) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2\right]\right),
\end{aligned} \tag{153}
$$

where the second inequality is by Lemma D.2.
Moreover, from Lemma D.7, we get

$$
\begin{aligned}
\mathbb{E}\left[\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_t) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2\right] &\le 3\mathbb{E}\left[\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_t) - \mathbf{y}^*_{t+1}(\mathbf{x}_t)\|^2\right]\\
&+ 3\mathbb{E}\left[\|\mathbf{y}^*_{t+1}(\mathbf{x}_t) - \mathbf{y}^*_t(\mathbf{x}_t)\|^2\right] + 3\mathbb{E}\left[\|\mathbf{y}^*_t(\mathbf{x}_t) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2\right]\\
&\le 3\mathbb{E}\left[\|\mathbf{y}^*_{t+1}(\mathbf{x}_t) - \mathbf{y}^*_t(\mathbf{x}_t)\|^2\right] + \frac{6\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g}.
\end{aligned} \tag{154}
$$

Combining (153) and (154) yields

$$
\begin{aligned}
&\mathbb{E}\left[\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2\right]\\
&\le 2\left(L_{\mathbf{y}}^2 \mathbb{E}\left[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2\right] + 3\mathbb{E}\left[\|\mathbf{y}^*_{t+1}(\mathbf{x}_t) - \mathbf{y}^*_t(\mathbf{x}_t)\|^2\right] + \frac{6\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g}\right).
\end{aligned} \tag{155}
$$

Substituting (155) into (152) and summing over $t \in [T]$, give the desired result.

□

D.4. Bounds on the Zeroth-Order System Solution

Lemma D.12. *Suppose Assumptions B2. and B3. hold. Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 2, we have*

$$
\begin{aligned}
&\mathbb{E}\|\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t; \mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\|^2\\
&\le (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2\mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2\\
&+ \frac{9}{2}d_2\ell_{g,1}^2\mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + (3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_2^2\rho_{\mathbf{r}}^2,
\end{aligned}
$$

where $\hat{\nabla}_{\mathbf{y}} f_t$ and $\hat{\nabla}^2_{\mathbf{y}} g_t$ are defined in (20a) and (21a), respectively.

Proof. From Lemma D.6, we have

$$\begin{aligned}
&\|\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1})\|^2 \\
&\leq 3d_2\ell_{f,1}^2\|\mathbf{z}_{t+1}-\mathbf{z}_t\|^2 + \frac{3}{2}\ell_{f,1}^2 d_2^2\rho_{\mathbf{r}}^2 \\
&\leq 6d_2\ell_{f,1}^2\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 6d_2\ell_{f,1}^2\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2 + \frac{3}{2}\ell_{f,1}^2 d_2^2\rho_{\mathbf{r}}^2.
\end{aligned} \tag{156}$$

Moreover, from (21a), we have

$$\begin{aligned}
&\|\hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\|^2 \\
&= \frac{1}{4\rho_{\mathbf{v}}^2}\|\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}+\rho_{\mathbf{v}}\mathbf{v}_{t+1};\bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t;\bar{\mathcal{B}}_{t+1})\|^2 \\
&\leq \frac{3}{4\rho_{\mathbf{v}}^2} d_2\ell_{g,1}^2\|(\mathbf{x}_{t+1},\mathbf{y}_{t+1}+\rho_{\mathbf{v}}\mathbf{v}_{t+1}) - (\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t)\|^2 + \frac{3}{8\rho_{\mathbf{v}}^2}\ell_{g,1}^2 d_2^2\rho_{\mathbf{r}}^2 \\
&\leq \frac{9}{4\rho_{\mathbf{v}}^2} d_2\ell_{g,1}^2\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + \frac{9}{4\rho_{\mathbf{v}}^2} d_2\ell_{g,1}^2\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2 \\
&+ \frac{9}{4} d_2\ell_{g,1}^2\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 + \frac{3}{8\rho_{\mathbf{v}}^2}\ell_{g,1}^2 d_2^2\rho_{\mathbf{r}}^2,
\end{aligned} \tag{157}$$

where the first inequality follows from Lemma D.6.

From $\|a+b\|^2 \leq 2\left(\|a\|^2+\|b\|^2\right)$, we get

$$\begin{aligned}
&\|\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\|^2 \\
&\leq 2\|\hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\|^2 \\
&+ 2\|\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1})\|^2 \\
&\leq (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2 \\
&+ \frac{9}{2}d_2\ell_{g,1}^2\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 + (3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_2^2\rho_{\mathbf{r}}^2,
\end{aligned}$$

where the second inequality follows from (156) and (157).

□

Lemma D.13. *Suppose Assumptions B2., B3., D1., and D3. hold. Consider the sequence* $\{(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\}_{t=1}^T$ *generated by Algorithm 2, and define*

$$e_{t+1}^M := \nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) - \hat{\mathbf{d}}_{t+1}^{\mathbf{v}}, \quad \text{where} \tag{158}$$

$$\begin{aligned}
\tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}) = \frac{1}{2\rho_{\mathbf{v}}}(&\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}+\rho_{\mathbf{v}}\mathbf{v}_{t+1}) \\
&- \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}-\rho_{\mathbf{v}}\mathbf{v}_{t+1})).
\end{aligned} \tag{159}$$

Then, we have

$$\begin{aligned}\mathbb{E}\|e^M_{t+1}\|^2 \leq\ & (1-\lambda_{t+1})^2\mathbb{E}\|e^M_t\|^2 + 36\mathbb{E}\left\|\nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t,\mathbf{y}_t)\right\|^2 \\ & + \left(18d_2^2\ell_{f,1}^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_2^2\right)\rho_{\mathbf{r}}^2 + 18d_2^2\ell_{g,1}^2\frac{\rho_{\mathbf{r}}^2}{\rho_{\mathbf{v}}^2} \\ & + \frac{18}{\rho_{\mathbf{v}}^2}\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\ & + \frac{18}{\rho_{\mathbf{v}}^2}\mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\ & + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2\mathbb{E}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2 \\ & + 27d_2\ell_{g,1}^2\mathbb{E}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 + 3(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b})\lambda_{t+1}^2. \end{aligned} \tag{160}$$

Proof. According to the definition of $\hat{\mathbf{d}}_t^{\mathbf{v}}$ in Algorithm 2, we have

$$\begin{aligned}\hat{\mathbf{d}}_{t+1}^{\mathbf{v}} - \hat{\mathbf{d}}_t^{\mathbf{v}} =\ & -\lambda_{t+1}\hat{\mathbf{d}}_t^{\mathbf{v}} + \lambda_{t+1}(\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})) \\ & + (1-\lambda_{t+1})\left(\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\right. \\ & \left. -\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\right). \end{aligned}$$

Then we have

$$\begin{aligned} & \mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_{t+1}^{\mathbf{v}}\|^2 \\ = & \mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_t^{\mathbf{v}} - (\hat{\mathbf{d}}_{t+1}^{\mathbf{v}} - \hat{\mathbf{d}}_t^{\mathbf{v}})\|^2 \\ = & \mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_t^{\mathbf{v}} + \lambda_{t+1}\hat{\mathbf{d}}_t^{\mathbf{v}} \\ & - \lambda_{t+1}\left(\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\right) \\ & - (1-\lambda_{t+1})\left(\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\right. \\ & \left.-\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\right)\|^2 \\ = & \mathbb{E}\|(1-\lambda_{t+1})(\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}_t^{\mathbf{v}}) \\ & + \lambda_{t+1}(\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})) \\ & + (1-\lambda_{t+1})\left(\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t)\right. \\ & +\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t) \\ & \left.-\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) + \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\right)\|^2. \end{aligned}$$

Since

$$\begin{aligned} & \mathbb{E}\left[\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\right] = \nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}), \\ & \mathbb{E}\left[\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\right] \\ & = \nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t), \end{aligned}$$

then, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{v}}_{t+1}\|^2 \\
&= (1-\lambda_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}^{\mathbf{v}}_t\|^2 \\
&+ \|\lambda_{t+1}(\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})) \\
&+ (1-\lambda_{t+1})\Big(\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t) \\
&+\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t) \\
&-\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) + \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\Big)\|^2 \\
&\le (1-\lambda_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}^{\mathbf{v}}_t\|^2 \\
&+ 3(1-\lambda_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t) \\
&+ \nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t) \\
&- \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) + \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\|^2 \\
&+ 3\lambda^2_{t+1}\mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1})\|^2 \\
&+ 3\lambda^2_{t+1}\mathbb{E}\|\tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\|^2,
\end{aligned}
\tag{161}
$$

where the second inequality holds by Cauchy-Schwarz inequality.
Note that, for the last term on the right-hand side of (161), from (21a) and (159), we have

$$
\begin{aligned}
&\|\tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\|^2 \\
&\le 2\|\frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}+\rho_{\mathbf{v}}\mathbf{v}_{t+1}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}+\rho_{\mathbf{v}}\mathbf{v}_{t+1};\bar{\mathcal{B}}_{t+1}))\|^2 \\
&+ 2\|\frac{1}{2\rho_{\mathbf{v}}}(\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}-\rho_{\mathbf{v}}\mathbf{v}_{t+1};\bar{\mathcal{B}}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}-\rho_{\mathbf{v}}\mathbf{v}_{t+1}))\|^2 \\
&\le \frac{\hat{\sigma}^2_{g_{\mathbf{y}}}}{\bar{b}\rho^2_{\mathbf{v}}},
\end{aligned}
$$

where the last inequality follows from Assumption 4.1.
Then, from $\mathbb{E}\|a-\mathbb{E}[a]\|^2 = \mathbb{E}\|a\|^2 - \|\mathbb{E}[a]\|^2$ and Assumption 4.1, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{v}}_{t+1}\|^2 \\
&\le (1-\lambda_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}^{\mathbf{v}}_t\|^2 \\
&+ 6(1-\lambda_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t)\|^2 \\
&+ 6(1-\lambda_{t+1})^2\mathbb{E}\|\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) \\
&- \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\|^2 + 3\lambda^2_{t+1}\left(\frac{\hat{\sigma}^2_{g_{\mathbf{y}}}}{\bar{b}\rho^2_{\mathbf{v}}} + \frac{\hat{\sigma}^2_{f_{\mathbf{y}}}}{b}\right).
\end{aligned}
$$

Then, from Young's inequality and Lemma D.12, we obtain

$$\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{v}}_{t+1}\|^2 \\
&\leq (1-\lambda_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}^{\mathbf{v}}_t\|^2 \\
&+ 12(1-\lambda_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t)\|^2 \\
&+ 12(1-\lambda_{t+1})^2 \mathbb{E}\|\tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t)\|^2 \\
&+ 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_2 \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\
&+ 27 d_2 \ell_{g,1}^2 \|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) d_2^2 \rho_{\mathbf{r}}^2 + 3\lambda_{t+1}^2 (\frac{\hat{\sigma}^2_{g_{\mathbf{y}}}}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}^2_{f_{\mathbf{y}}}}{b}),
\end{aligned} \tag{162}$$

For the third term on the right-hand side of (162), based on (159), we have

$$\begin{aligned}
&\|\tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\leq \frac{1}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 && \text{(163a)} \\
&+ \frac{1}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)\|^2. && \text{(163b)}
\end{aligned}$$

For (163a), we get

$$\begin{aligned}
&\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&\leq 3\|\nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&+ 3\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&+ 3\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&\leq 3\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 + \frac{3\rho_{\mathbf{r}}^2 d_2^2 \ell_{g,1}^2}{2},
\end{aligned}$$

where the last inequality follows from Eq. (131).
Similary, for (163b), we have

$$\begin{aligned}
&\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&\leq 3\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 + \frac{3\rho_{\mathbf{r}}^2 d_2^2 \ell_{g,1}^2}{2}.
\end{aligned}$$

Substituting the above inequalities in (163), we have

$$\begin{aligned}
\|\tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 &\leq \frac{3}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
+ \frac{3}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t) &- \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 + \frac{3\rho_{\mathbf{r}}^2 d_2^2 \ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}.
\end{aligned} \tag{164}$$

For the second term on the right-hand side of (162), we have

$$\begin{aligned}
&\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\leq 3\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&+ 3\|\nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&+ 3\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\leq 3\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \frac{3\rho_{\mathbf{r}}^2 d_2^2 \ell_{f,1}^2}{2},
\end{aligned} \tag{165}$$

where the last inequality follows from Eq. (133).

From (164), (165) and (162), we get

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{v}}_{t+1}\|^2 \\
&\leq (1-\lambda_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}^{\mathbf{v}}_t\|^2 \\
&+ 36\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_t,\mathbf{y}_t)\|^2 + 18\rho_{\mathbf{r}}^2 d_2^2 \ell_{f,1}^2 \\
&+ \frac{18}{\rho_{\mathbf{v}}^2}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&+ \frac{18}{\rho_{\mathbf{v}}^2}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t)\|^2 + \frac{18\rho_{\mathbf{r}}^2 d_2^2 \ell_{g,1}^2}{\rho_{\mathbf{v}}^2} \\
&+ 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_2 \|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_2 \|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2 \\
&+ 27 d_2 \ell_{g,1}^2 \|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) d_2^2 \rho_{\mathbf{r}}^2 + 3\lambda_{t+1}^2 (\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b}).
\end{aligned}
$$

□

Lemma D.14. *Suppose Assumptions B3. and B4. hold. Let*

$$e_t^H := \tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t) - \nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\mathbf{v}_t, \tag{166a}$$

$$e_t^J := \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{x}_t,\mathbf{y}_t) - \nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\mathbf{v}_t, \tag{166b}$$

where

$$
\begin{aligned}
\tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t) &= \frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t)), \\
\tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{x}_t,\mathbf{y}_t) &= \frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t)).
\end{aligned}
$$

Then, for $(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)$ presented to Algorithm 2, we have

(a)

$$\mathbb{E}\left[\left\|e_t^H\right\|^2\right] \leq \ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4. \tag{167a}$$

(b)

$$\mathbb{E}\left[\left\|e_t^J\right\|^2\right] \leq \ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4. \tag{167b}$$

Proof. **For part (a)**: From Lemma D.1, We have

$$
\begin{aligned}
\mathbb{E}\left[\left\|e_t^H\right\|\right] &= \mathbb{E}\left[\left\|\tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t) - \nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\mathbf{v}_t\right\|\right] \\
&\leq \frac{1}{2\rho_{\mathbf{v}}}\mathbb{E}\left[\left\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t) - \nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\rho_{\mathbf{v}}\mathbf{v}_t\right\|\right] \\
&+ \frac{1}{2\rho_{\mathbf{v}}}\mathbb{E}\left[\left\|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\rho_{\mathbf{v}}\mathbf{v}_t\right\|\right] \\
&\leq \ell_{g,2}\rho_{\mathbf{v}}\mathbb{E}\left[\left\|\mathbf{v}_t\right\|^2\right] \\
&\leq \ell_{g,2}\rho_{\mathbf{v}} p^2,
\end{aligned} \tag{168}
$$

where the last inequality follows from (8).

For part (b): From Lemma D.1, We have

$$
\begin{aligned}
\mathbb{E}\left[\|e_t^J\|\right] &= \mathbb{E}\left[\left\|\tilde{\nabla}^2_{\mathbf{xy}} g_t\left(\mathbf{x}_t, \mathbf{y}_t\right) - \nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x}_t, \mathbf{y}_t\right)\mathbf{v}_t\right\|\right] \\
&\leq \frac{1}{2\rho_{\mathbf{v}}}\mathbb{E}\left[\|\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \nabla^2_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x}_t, \mathbf{y}_t\right)\rho_{\mathbf{v}}\mathbf{v}_t\|\right] \\
&+ \frac{1}{2\rho_{\mathbf{v}}}\mathbb{E}\left[\|\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{xy}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x}_t, \mathbf{y}_t\right)\rho_{\mathbf{v}}\mathbf{v}_t\|\right] \\
&\leq \ell_{g,2}\rho_{\mathbf{v}}\mathbb{E}\left[\|\mathbf{v}_t\|^2\right] \\
&\leq \ell_{g,2}\rho_{\mathbf{v}}p^2,
\end{aligned} \tag{169}
$$

where the last inequality follows from (8). □

Lemma D.15. *Suppose Assumptions B2., B3. and B4. hold. Then, for directions $\hat{\mathbf{d}}_t^{\mathbf{v}}$ and $\hat{\mathbf{d}}_t^{\mathbf{x}}$ presented to Algorithm 2, and*

(a) $\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}$ defined in (120b), we have

$$
\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{v}} - \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}\right\|^2\right] \leq 2\mathbb{E}\left[\|e_t^M\|^2\right] + 2\ell_{g,2}^2\rho_{\mathbf{v}}^2p^4 := B_t, \tag{170a}
$$

where $e_t^M := \hat{\mathbf{d}}_t^{\mathbf{v}} - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_t\left(\mathbf{x}_t, \mathbf{y}_t\right)$, with $\tilde{\nabla}^2_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)$ is defined in (171).

(b) $\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{x}}$ defined in (120c), we have

$$
\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{x}}\right\|^2\right] \leq 2\mathbb{E}\left[\|e_t^L\|^2\right] + 2\ell_{g,2}^2\rho_{\mathbf{v}}^2p^4, \tag{170b}
$$

where $e_t^L := \hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}^2_{\mathbf{xy}} g_t\left(\mathbf{x}_t, \mathbf{y}_t\right)$ with $\tilde{\nabla}^2_{\mathbf{xy}} g_t\left(\mathbf{x}_t, \mathbf{y}_t\right)$ is defined in (176).

Proof. **For part (a)**: Let

$$
\tilde{\nabla}^2_{\mathbf{y}} g_t\left(\mathbf{x}_t, \mathbf{y}_t\right) = \frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)). \tag{171}
$$

According to the definition of $\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}$ in (120b), we have

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{v}} - \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}\right\|^2\right] &= \mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{v}} - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x}_t, \mathbf{y}_t\right)\mathbf{v}\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{v}} - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_t\left(\mathbf{x}_t, \mathbf{y}_t\right)\right\|^2\right] && \text{(172a)} \\
&+ 2\mathbb{E}\left[\left\|\tilde{\nabla}^2_{\mathbf{y}} g_t\left(\mathbf{x}_t, \mathbf{y}_t\right) - \nabla^2_{\mathbf{y}} g_{t,\boldsymbol{\rho}}\left(\mathbf{x}_t, \mathbf{y}_t\right)\mathbf{v}\right\|^2\right]. && \text{(172b)}
\end{aligned}
$$

Next, we separately bound (172a)–(172b) on the RHS of the above inequality.
Bounding (172a) . We have

$$
2\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{v}} - \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}^2_{\mathbf{y}} g_t\left(\mathbf{x}_t, \mathbf{y}_t\right)\right\|^2\right] := 2\mathbb{E}\left[\|e_t^M\|^2\right]. \tag{173}
$$

Bounding (172b) . From Lemmas D.1 and D.14, we have

$$
(172\text{b}) = \mathbb{E}\left[\|e_t^H\|^2\right] \leq 3\ell_{g,2}^2\rho_{\mathbf{v}}^2p^4. \tag{174}
$$

Combining (173)-(174) yields

$$\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{v}} - \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}\right\|^2\right] \leq 2\mathbb{E}\left[\left\|e_t^M\right\|^2\right] + 2\ell_{g,2}^2\rho_{\mathbf{v}}^2 p^4. \tag{175}$$

For part (b): Let

$$\tilde{\nabla}_{\mathbf{xy}}^2 g_t\left(\mathbf{x}_t, \mathbf{y}_t\right) = \frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)). \tag{176}$$

According to the definition of $\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{x}}$ in (120c), we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{x}}\right\|^2\right] &= \mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{xy}}^2 g_{t,\boldsymbol{\rho}}\left(\mathbf{x}_t, \mathbf{y}_t\right)\mathbf{v}\right\|^2\right] \\
&\leq 2\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}_{\mathbf{xy}}^2 g_t\left(\mathbf{x}_t, \mathbf{y}_t\right)\right\|^2\right] \qquad (177a) \\
&\quad + 2\mathbb{E}\left[\left\|\tilde{\nabla}_{\mathbf{xy}}^2 g_t\left(\mathbf{x}_t, \mathbf{y}_t\right) - \nabla_{\mathbf{xy}}^2 g_{t,\boldsymbol{\rho}}\left(\mathbf{x}_t, \mathbf{y}_t\right)\mathbf{v}_t\right\|^2\right]. \qquad (177b)
\end{aligned}$$

Next, we separately bound (177a)–(177b) on the RHS of the above inequality.
Bounding (177a) . We have

$$2\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}_{\mathbf{xy}}^2 g_t\left(\mathbf{x}_t, \mathbf{y}_t\right)\right\|^2\right] := 2\mathbb{E}\left[\left\|e_t^L\right\|^2\right]. \tag{178}$$

Bounding (177b) . From Lemmas D.1 and D.14, we have

$$(177b) = \mathbb{E}\left[\left\|e_t^J\right\|^2\right] \leq 2\ell_{g,2}^2\rho_{\mathbf{v}}^2 p^4. \tag{179}$$

Combining (178)–(179) yields

$$\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{x}}\right\|^2\right] \leq 2\mathbb{E}\left[\left\|e_t^L\right\|^2\right] + 2\ell_{g,2}^2\rho_{\mathbf{v}}^2 p^4.$$

□

Lemma D.16. *Suppose Assumptions 3.2, B1. and B3. hold. Set the step size δ_t and the parameter p in (8), as*

$$\delta_t \leq \left(2 + \frac{1}{\ell_{g,1}^2}\right)\frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}, \ \forall t \in [T], \quad \text{ and } \quad p = \frac{\ell_{f,0}}{\mu_g}. \tag{180}$$

Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 2, we have

$$\mathbb{E}\left[\left\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\right\|^2\right] \leq (1+\acute{a})\left(1 - \delta_t\frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right)\mathbb{E}[\hat{\theta}_t^{\mathbf{v}}] + \left(1 + \frac{1}{\acute{a}}\right)\delta_t^2 B_t,$$

for some $\acute{a} > 0$, where $\hat{\theta}_t^{\mathbf{v}}$ and B_t are defined in Eq. (146) and Lemma D.15, respectively.

Proof. By setting the radius $p := \frac{\ell_{f,0}}{\mu_g}$ in (8), we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2\right] &= \mathbb{E}\left[\left\|\Pi_{\mathcal{Z}_p}\left[\mathbf{v}_t - \delta_t \hat{\mathbf{d}}_t^{\mathbf{v}}\right] - \Pi_{\mathcal{Z}_p}\left[\hat{\mathbf{v}}_t^*(\mathbf{x}_t)\right]\right\|^2\right] \\
&\le \mathbb{E}\left[\|\mathbf{v}_t - \delta_t \hat{\mathbf{d}}_t^{\mathbf{v}} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2\right] \\
&\le (1+\acute{a})\underbrace{\mathbb{E}\left[\|\mathbf{v}_t - \delta_t \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t) - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2\right]}_{I_t} \\
&+ \left(1 + \frac{1}{\acute{a}}\right)\delta_t^2 \underbrace{\mathbb{E}\left[\|\hat{\mathbf{d}}_t^{\mathbf{v}} - \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2\right]}_{K_t},
\end{aligned} \tag{181}$$

where $\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)$ is defined in (120b); the first inequality follows from non-expansiveness property of a projection operator.

We next bound the I_t, and K_t terms in (181), respectively.

Bounding I_t . We have

$$\begin{aligned}
I_t &= \mathbb{E}\left[\|\mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2\right] - 2\delta_t \mathbb{E}\left[\langle \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t), \mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\rangle\right] + \delta_t^2 \mathbb{E}\left[\|\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2\right] \\
&\le \left(1 - 2\delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right)\mathbb{E}\left[\|\mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2\right] \\
&- \left(2\delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}} - \delta_t^2\right)\mathbb{E}\left[\|\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2\right],
\end{aligned}$$

where the inequality holds since $\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}$ in (120) is the gradient of the strongly convex quadratic program $\frac{1}{2}\mathbf{v}^\top \nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})\mathbf{v} + \mathbf{v}^\top \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})$.

Thus, we have

$$\begin{aligned}
&\mathbb{E}\left[\langle \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t), \mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\rangle\right] \\
&\ge \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\mathbb{E}\left[\|\mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2\right] + \frac{1}{\mu_g + \ell_{g,1}}\mathbb{E}\left[\|\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2\right].
\end{aligned}$$

Since $\delta_t \le \left(2 + \frac{1}{\ell_{g,1}^2}\right)\frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}$, then we have

$$\begin{aligned}
I_t &\le \left(1 - 2\delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right)\mathbb{E}\left[\|\mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2\right] + \frac{1}{\ell_{g,1}^2}\left(\frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\delta_t\right)\mathbb{E}\left[\|\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2\right] \\
&\le \left(1 - \delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right)\mathbb{E}\left[\|\mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2\right],
\end{aligned} \tag{182}$$

where the second inequality holds since from (119b), we have

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2\right] &= \mathbb{E}\left[\|\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) + \nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})\mathbf{v}\|^2\right] \\
&= \mathbb{E}\left[\|\nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})(\mathbf{v} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t))\|^2\right] \\
&\le \ell_{g,1}^2 \mathbb{E}\left[\|\mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2\right].
\end{aligned}$$

Bounding K_t . Let

$$\tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) = \frac{1}{2\rho_{\mathbf{v}}}\left(\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)\right).$$

From Lemma D.15, we have

$$K_t = \mathbb{E}\left[\|\hat{\mathbf{d}}_t^{\mathbf{v}} - \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2\right] \le B_t. \tag{183}$$

Putting (182), and (183) together with Eq. (181) yields the desired result.

$$\mathbb{E}\left[\left\|\mathbf{v}_{t+1}-\hat{\mathbf{v}}_t^*(\mathbf{x}_t)\right\|^2\right] \leq (1+\acute{a})\left(1-\delta_t\frac{\mu_g\ell_{g,1}}{\mu_g+\ell_{g,1}}\right)\mathbb{E}\left[\left\|\mathbf{v}_t-\hat{\mathbf{v}}_t^*(\mathbf{x}_t)\right\|^2\right]+\left(1+\frac{1}{\acute{a}}\right)\delta_t^2 B_t.$$

□

Lemma D.17. *Suppose Assumptions 3.2 and 3.3 hold. Let $\hat{\theta}_t^{\mathbf{v}}$ be defined in (146). Set the parameter p in (8) as $p=\frac{\ell_{f,0}}{\mu_g}$. Then, for any positive choice of step sizes satisfying*

$$\delta_t \leq \left(2+\frac{1}{\ell_{g,1}^2}\right)\frac{\mu_g\ell_{g,1}}{\mu_g+\ell_{g,1}},$$

the sequence $\{(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 2 guarantees the following bound:

$$\begin{aligned}\sum_{t=1}^T\left(\mathbb{E}[\hat{\theta}_{t+1}^{\mathbf{v}}]-\mathbb{E}[\hat{\theta}_t^{\mathbf{v}}]\right) \leq{}& \sum_{t=1}^T\left(-\frac{L_{\mu_g}}{4}\mathbb{E}[\hat{\theta}_t^{\mathbf{v}}]+\frac{4}{L_{\mu_g}}B_t\right)\delta_t\\ &+\frac{16\nu^2}{L_{\mu_g}\mu_g^2}(2L_{\mathbf{y}}^2+1)\sum_{t=1}^T\mathbb{E}\left\|\mathbf{x}_{t+1}-\mathbf{x}_t\right\|^2\frac{1}{\delta_t}\\ &+\sum_{t=1}^T\left(\frac{96\ell_{g,1}\nu^2}{L_{\mu_g}\mu_g^3}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)+\frac{48\nu^2}{L_{\mu_g}\mu_g^2}\sup_{\mathbf{x}\in\mathcal{X}}\left\|\mathbf{y}_{t-1}^*(\mathbf{x})-\mathbf{y}_t^*(\mathbf{x})\right\|^2\right)\frac{1}{\delta_t},\end{aligned}\tag{184}$$

where B_t and ν are defined in Lemmas D.15 and C.7, respectively.

Proof. From Lemma B.5, we have, for any $\acute{c}>0$

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{v}_{t+1}-\hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2\right] &= \mathbb{E}\left[\left\|\mathbf{v}_{t+1}-\hat{\mathbf{v}}_t^*(\mathbf{x}_t)+\hat{\mathbf{v}}_t^*(\mathbf{x}_t)-\hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2\right]\\ &\leq (1+\acute{c})\,\mathbb{E}\left[\left\|\mathbf{v}_{t+1}-\hat{\mathbf{v}}_t^*(\mathbf{x}_t)\right\|^2\right]\\ &\quad+\left(1+\frac{1}{\acute{c}}\right)\mathbb{E}\left[\left\|\hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1})-\hat{\mathbf{v}}_t^*(\mathbf{x}_t)\right\|^2\right].\end{aligned}\tag{185}$$

From Lemma D.16, we have, for any $\acute{a}>0$

$$\mathbb{E}\left[\left\|\mathbf{v}_{t+1}-\hat{\mathbf{v}}_t^*(\mathbf{x}_t)\right\|^2\right] \leq (1+\acute{a})\left(1-\delta_t\frac{\mu_g\ell_{g,1}}{\mu_g+\ell_{g,1}}\right)\hat{\theta}_t^{\mathbf{v}}+\left(1+\frac{1}{\acute{a}}\right)\delta_t^2 B_t.\tag{186}$$

Substituting (186) into (185), we get

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{v}_{t+1}-\hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1})\right\|^2\right] \leq{}& (1+\acute{c})\,(1+\acute{a})\left(1-\delta_t\frac{\mu_g\ell_{g,1}}{\mu_g+\ell_{g,1}}\right)\hat{\theta}_t^{\mathbf{v}}\\ &+(1+\acute{c})\left(1+\frac{1}{\acute{a}}\right)\delta_t^2 B_t\\ &+\left(1+\frac{1}{\acute{c}}\right)\mathbb{E}\left[\left\|\hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1})-\hat{\mathbf{v}}_t^*(\mathbf{x}_t)\right\|^2\right].\end{aligned}\tag{187}$$

Choose $\acute{c} = \frac{\delta_t L_{\mu_g}/4}{1-\frac{\delta_t L_{\mu_g}}{2}}$ and $\acute{a} = \frac{\delta_t L_{\mu_g}/2}{1-\delta_t L_{\mu_g}}$. Then, the following equations and inequalities are satisfied.

$$\begin{aligned}
&(1+\acute{c})(1+\acute{a})\left(1-\delta_t L_{\mu_g}\right) = 1 - \frac{\delta_t L_{\mu_g}}{4},\\
&(1+\acute{c})\left(1+\frac{1}{\acute{a}}\right) \leq \frac{4}{\delta_t L_{\mu_g}},\\
&1+\frac{1}{\acute{a}} \leq \frac{2}{\delta_t L_{\mu_g}}, \quad 1+\frac{1}{\acute{c}} \leq \frac{4}{\delta_t L_{\mu_g}},
\end{aligned} \tag{188}$$

where $L_{\mu_g} = \frac{\mu_g \ell_{g,1}}{\mu_g+\ell_{g,1}}$.

Thus, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}^*_{t+1}(\mathbf{x}_{t+1})\right\|^2\right] \leq & \left(1-\frac{\delta_t L_{\mu_g}}{4}\right)\hat{\theta}^{\mathbf{v}}_t + \frac{4}{L_{\mu_g}}\delta_t B_t\\
& + \frac{4}{L_{\mu_g}}\frac{1}{\delta_t}\mathbb{E}\left[\left\|\hat{\mathbf{v}}^*_{t+1}(\mathbf{x}_{t+1}) - \hat{\mathbf{v}}^*_t(\mathbf{x}_t)\right\|^2\right].
\end{aligned} \tag{189}$$

We now bound the last term on the right-hand side of (189). By Lemma C.7, we have:

$$\begin{aligned}
&\left\|\hat{\mathbf{v}}^*_{t+1}(\mathbf{x}_{t+1}) - \hat{\mathbf{v}}^*_t(\mathbf{x}_t)\right\|^2\\
&\leq 2\frac{\nu^2}{\mu_g^2}\left(\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2 + \|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right)\\
&\leq 2\frac{\nu^2}{\mu_g^2}\left(2\left\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_t)\right\|^2\right.\\
&\left.+2\left\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_t) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\right\|^2 + \|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right)\\
&\leq 2\frac{\nu^2}{\mu_g^2}\left(2L_{\mathbf{y}}^2\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 2\left\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_t) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\right\|^2 + \|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\right),
\end{aligned} \tag{190}$$

where the last inequality follows from Lemma D.2.
From (154), we have

$$\begin{aligned}
\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_t) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2 \leq\ & 3\|\hat{\mathbf{y}}^*_{t+1}(\mathbf{x}_t) - \mathbf{y}^*_{t+1}(\mathbf{x}_t)\|^2\\
& + 3\|\mathbf{y}^*_{t+1}(\mathbf{x}_t) - \mathbf{y}^*_t(\mathbf{x}_t)\|^2 + 3\|\mathbf{y}^*_t(\mathbf{x}_t) - \hat{\mathbf{y}}^*_t(\mathbf{x}_t)\|^2\\
\leq\ & 3\|\mathbf{y}^*_{t+1}(\mathbf{x}_t) - \mathbf{y}^*_t(\mathbf{x}_t)\|^2 + \frac{6\ell_{g,1}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)}{\mu_g}.
\end{aligned} \tag{191}$$

Plugging (191) into (190), we get

$$\begin{aligned}
&\left\|\hat{\mathbf{v}}^*_{t+1}(\mathbf{x}_{t+1}) - \hat{\mathbf{v}}^*_t(\mathbf{x}_t)\right\|^2\\
&\leq 4\frac{\nu^2}{\mu_g^2}(2L_{\mathbf{y}}^2+1)\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2\\
&+4\frac{\nu^2}{\mu_g^2}\left(3\|\mathbf{y}^*_{t+1}(\mathbf{x}_t) - \mathbf{y}^*_t(\mathbf{x}_t)\|^2 + \frac{6\ell_{g,1}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)}{\mu_g}\right).
\end{aligned} \tag{192}$$

Then, substituting (192) into (189), rearranging the resulting inequality and summing over $t \in [T]$, we obtain the desired result. □

D.5. Bounds on the Zeroth-Order Objective Function and its Projected Gradients

Lemma D.18. *Suppose Assumptions B2., B3. hold. Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 2, we have*

$$
\begin{aligned}
&\|\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t; \mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\|^2 \\
&\leq (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\
&+ \frac{9}{2} d_1 \ell_{g,1}^2 \|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + (3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) d_1^2 \rho_{\mathbf{s}}^2,
\end{aligned}
$$

where $\hat{\nabla}_{\mathbf{x}} f_{t+1}$ and $\hat{\nabla}^2_{\mathbf{xy}} g_{t+1}$ are defined in (20b) and (21b), respectively.

Proof. From Lemma D.6, we have

$$
\begin{aligned}
&\|\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t; \mathcal{B}_{t+1})\|^2 \\
&\leq 3 d_1 \ell_{g,1}^2 \|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 + \frac{3}{2} \ell_{f,1}^2 d_1^2 \rho_{\mathbf{s}}^2 \\
&\leq 6 d_1 \ell_{f,1}^2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6 d_1 \ell_{f,1}^2 \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 + \frac{3}{2} \ell_{f,1}^2 d_1^2 \rho_{\mathbf{s}}^2.
\end{aligned} \tag{193}
$$

Moreover, from (21a), we have

$$
\begin{aligned}
&\|\hat{\nabla}^2_{\mathbf{yy}} g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}^2_{\mathbf{yy}} g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\|^2 \\
&= \frac{1}{4\rho_{\mathbf{v}}^2} \|\hat{\nabla}_{\mathbf{x}} g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1} + \rho_{\mathbf{v}} \mathbf{v}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t; \bar{\mathcal{B}}_{t+1})\|^2 \\
&\leq \frac{3}{4\rho_{\mathbf{v}}^2} d_1 \ell_{g,1}^2 \|(\mathbf{x}_{t+1}, \mathbf{y}_{t+1} + \rho_{\mathbf{v}} \mathbf{v}_{t+1}) - (\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 + \frac{3}{8\rho_{\mathbf{v}}^2} \ell_{g,1}^2 d_1^2 \rho_{\mathbf{s}}^2 \\
&\leq \frac{9}{4\rho_{\mathbf{v}}^2} d_1 \ell_{g,1}^2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \frac{9}{4\rho_{\mathbf{v}}^2} d_1 \ell_{g,1}^2 \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\
&+ \frac{9}{4} d_1 \ell_{g,1}^2 \|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + \frac{3}{8\rho_{\mathbf{v}}^2} \ell_{g,1}^2 d_1^2 \rho_{\mathbf{s}}^2,
\end{aligned} \tag{194}
$$

where the first inequality follows from Lemma D.6.

From $\|a+b\|^2 \leq 2\left(\|a\|^2 + \|b\|^2\right)$, we get

$$
\begin{aligned}
&\|\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t; \mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\|^2 \\
&\leq 2\|\hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\|^2 \\
&+ 2\|\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t; \mathcal{B}_{t+1})\|^2 \\
&\leq (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\
&+ \frac{9}{2} d_1 \ell_{g,1}^2 \|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + (3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) d_1^2 \rho_{\mathbf{s}}^2,
\end{aligned}
$$

where the second inequality follows from (193) and (194).

□

Lemma D.19. *Suppose Assumptions B2., B3., D2., and D4. hold. Consider the sequence $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 2, and define*

$$
e_t^L := \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) + \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{x}_t, \mathbf{y}_t) - \hat{\mathbf{d}}_t^{\mathbf{x}}, \quad \text{where} \tag{195}
$$

$$
\tilde{\nabla}^2_{\mathbf{xy}} g_t\left(\mathbf{x}_t, \mathbf{y}_t\right) = \frac{1}{2\rho_{\mathbf{v}}} (\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)). \tag{196}
$$

Then, we have

$$
\begin{aligned}
\mathbb{E}\|e_{t+1}^{L}\|^2 \leq\ & (1-\eta_{t+1})^2\mathbb{E}\|e_t^L\|^2 + 36\mathbb{E}\left\|\nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t,\mathbf{y}_t)\right\|^2 \\
&+ \left(18d_1^2\ell_{f,1}^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_1^2\right)\rho_{\mathbf{s}}^2 + 18d_1^2\ell_{g,1}^2\frac{\rho_{\mathbf{s}}^2}{\rho_{\mathbf{v}}^2} \\
&+ \frac{18}{\rho_{\mathbf{v}}^2}\mathbb{E}\|\nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&+ \frac{18}{\rho_{\mathbf{v}}^2}\mathbb{E}\|\nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&+ 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1\mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1\mathbb{E}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2 \\
&+ 27d_1\ell_{g,1}^2\mathbb{E}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 + 3(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b})\eta_{t+1}^2.
\end{aligned} \tag{197}
$$

Proof. According to the definition of $\hat{\mathbf{d}}_t^{\mathbf{x}}$ in Algorithm 2, we have

$$
\begin{aligned}
\hat{\mathbf{d}}_{t+1}^{\mathbf{x}} - \hat{\mathbf{d}}_t^{\mathbf{x}} =\ & -\eta_{t+1}\hat{\mathbf{d}}_t^{\mathbf{x}} + \eta_{t+1}(\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})) \\
&+ (1-\eta_{t+1})\Big(\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) \\
&-\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\Big).
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_{t+1}^{\mathbf{x}}\|^2 \\
&= \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_t^{\mathbf{x}} - (\hat{\mathbf{d}}_{t+1}^{\mathbf{x}} - \hat{\mathbf{d}}_t^{\mathbf{x}})\|^2 \\
&= \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_t^{\mathbf{x}} + \eta_{t+1}\hat{\mathbf{d}}_t^{\mathbf{x}} \\
&- \eta_{t+1}(\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})) \\
&- (1-\eta_{t+1})\Big(\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) \\
&-\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\Big)\|^2 \\
&= \mathbb{E}\|(1-\eta_{t+1})(\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{z}_t) - \hat{\mathbf{d}}_t^{\mathbf{x}}) \\
&+ \eta_{t+1}(\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})) \\
&+ (1-\eta_{t+1})\Big(\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{z}_t) \\
&+\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_t) \\
&-\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) + \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) + \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\Big)\|^2.
\end{aligned}
$$

Since

$$
\begin{aligned}
&\mathbb{E}\left[\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\right] = \nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}), \\
&\mathbb{E}\left[\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\right] \\
&= \nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_t),
\end{aligned}
$$

then, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{x}}_{t+1}\|^2 \\
&= (1-\eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}^{\mathbf{x}}_t\|^2 \\
&+ \|\eta_{t+1}(\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})) \\
&+ (1-\eta_{t+1})\Big(\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{z}_t) \\
&+\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t) \\
&-\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) + \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\Big)\|^2 \\
&\le (1-\eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}^{\mathbf{x}}_t\|^2 \\
&+ 3(1-\eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{z}_t) \\
&+ \nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t) \\
&- \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) - \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) + \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\|^2 \\
&+ 3\eta_{t+1}^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1})\|^2 \\
&+ 3\eta_{t+1}^2 \mathbb{E}\|\tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\|^2,
\end{aligned} \tag{198}
$$

where the second inequality holds by Cauchy-Schwarz inequality.
Note that for the last term on the right-hand side of (198), using (196) and (21b), we have

$$
\begin{aligned}
&\|\tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1})\|^2 \\
&\le 2\|\frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{x}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}+\rho_{\mathbf{v}}\mathbf{v}_{t+1}) - \hat{\nabla}_{\mathbf{x}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}+\rho_{\mathbf{v}}\mathbf{v}_{t+1};\bar{\mathcal{B}}_{t+1}))\|^2 \\
&+ 2\|\frac{1}{2\rho_{\mathbf{v}}}(\hat{\nabla}_{\mathbf{x}} g_{t+1}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}-\rho_{\mathbf{v}}\mathbf{v}_{t+1};\bar{\mathcal{B}}_{t+1}) - \nabla_{\mathbf{x}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\mathbf{y}_{t+1}-\rho_{\mathbf{v}}\mathbf{v}_{t+1}))\|^2 \\
&\le \frac{\hat{\sigma}^2_{g_{\mathbf{x}}}}{\bar{b}\rho_{\mathbf{v}}^2},
\end{aligned}
$$

where the last inequality follows from Assumption 4.1.
Then, from $\mathbb{E}\|a-\mathbb{E}[a]\|^2 = \mathbb{E}\|a\|^2 - \|\mathbb{E}[a]\|^2$ and Assumption 4.1, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{x}}_{t+1}\|^2 \\
&\le (1-\eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}^{\mathbf{x}}_t\|^2 \\
&+ 6(1-\eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{z}_t)\|^2 \\
&+ 6(1-\eta_{t+1})^2 \mathbb{E}\|\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1};\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1};\bar{\mathcal{B}}_{t+1}) \\
&+ \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t;\mathcal{B}_{t+1}) + \hat{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t;\bar{\mathcal{B}}_{t+1})\|^2 + 3\eta_{t+1}^2(\frac{\hat{\sigma}^2_{g_{\mathbf{x}}}}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}^2_{f_{\mathbf{x}}}}{b}),
\end{aligned} \tag{199}
$$

Then, from Young's inequality and Lemma D.18, we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{x}}_{t+1}\|^2 \\
&\le (1-\eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}^{\mathbf{x}}_t\|^2 \\
&+ 12(1-\eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_t) - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t)\|^2 + 12(1-\eta_{t+1})^2 \mathbb{E}\|\tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_t) - \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{z}_t)\|^2 \\
&+ 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2 \\
&+ 27 d_1 \ell_{g,1}^2 \|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) d_1^2 \rho_{\mathbf{s}}^2 + 3\eta_{t+1}^2(\frac{\hat{\sigma}^2_{g_{\mathbf{x}}}}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}^2_{f_{\mathbf{x}}}}{b}).
\end{aligned} \tag{200}
$$

For the third term on the right-hand side of (199), we have

$$
\begin{aligned}
&\|\tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\leq \frac{1}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{x}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \qquad (201a) \\
&+ \frac{1}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2. \qquad (201b)
\end{aligned}
$$

For (201a), we get

$$
\begin{aligned}
&\|\nabla_{\mathbf{x}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
&\leq 3\|\nabla_{\mathbf{x}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
&+ 3\|\nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
&+ 3\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
&\leq 3\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 + \frac{3\rho_{\mathbf{s}}^2 d_1^2 \ell_{g,1}^2}{2},
\end{aligned}
$$

where the last inequality follows from Lemma 131.
Similary, for (163b), we have

$$
\begin{aligned}
&\|\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
&\leq 3\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\| + \frac{3\rho_{\mathbf{s}}^2 d_1^2 \ell_{g,1}^2}{2}.
\end{aligned}
$$

Substituting these inequalities in (201), we have

$$
\begin{aligned}
&\|\tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\leq \frac{3}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
&+ \frac{3}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 + \frac{3\rho_{\mathbf{s}}^2 d_1^2 \ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}. \qquad (202)
\end{aligned}
$$

For the second term on the right-hand side of (199), we have

$$
\begin{aligned}
&\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\leq 3\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&+ 3\|\nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&+ 3\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\leq 3\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \frac{3\rho_{\mathbf{s}}^2 d_1^2 \ell_{f,1}^2}{2}, \qquad (203)
\end{aligned}
$$

where the last inequality follows from Eq. (133).

From (202), (203) and (200), we get

$$\begin{aligned}
&\mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1,\boldsymbol{\rho}}(\mathbf{z}_{t+1}) + \tilde{\nabla}^2_{\mathbf{xy}} g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}^{\mathbf{x}}_{t+1}\|^2 \\
&\leq (1-\eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{z}_t) + \tilde{\nabla}^2_{\mathbf{xy}} g_t(\mathbf{z}_t) - \hat{\mathbf{d}}^{\mathbf{x}}_t\|^2 \\
&+ 36\mathbb{E}\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t,\mathbf{y}_t) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t,\mathbf{y}_t)\|^2 + 18\rho_{\mathbf{s}}^2 d_1^2 \ell_{f,1}^2 \\
&+ \frac{18}{\rho_{\mathbf{v}}^2}\mathbb{E}\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&+ \frac{18}{\rho_{\mathbf{v}}^2}\mathbb{E}\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t)\|^2 + \frac{18\rho_{\mathbf{s}}^2 d_1^2 \ell_{g,1}^2}{\rho_{\mathbf{v}}^2} \\
&+ 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \mathbb{E}\|\mathbf{x}_{t+1}-\mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \mathbb{E}\|\mathbf{y}_{t+1}-\mathbf{y}_t\|^2 \\
&+ 27 d_1 \ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) d_1^2\rho_{\mathbf{s}}^2 + 3\eta_{t+1}^2(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b}).
\end{aligned}$$

□

Lemma D.20. *Suppose Assumptions 3.2, B2., B3., and 3.4 hold. Then, for the sequence of functions $\{f_{t,\boldsymbol{\rho}}\}_{t=1}^T$ defined in Eq. (17), we have*

$$\begin{aligned}
&\sum_{t=1}^T \left(f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_{t,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))\right) \\
&\leq 2M + V_T + \ell_{f,1}\left(1+2\frac{\ell_{g,1}}{\mu_g}\right) T\left(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2\right).
\end{aligned}$$

Here, V_T is defined in (10); and M is defined in Assumption 3.4.

Proof. Note that, we have

$$\begin{aligned}
&\sum_{t=1}^T \left(f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_{t,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))\right) \\
&= \sum_{t=1}^T \left(f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right) && (204)\\
&+ \sum_{t=1}^T \left(f_t(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1},\hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))\right) && (205)\\
&+ \sum_{t=1}^T \left(f_t(\mathbf{x}_{t+1},\hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1})) - f_{t,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))\right). && (206)
\end{aligned}$$

From (127), we have

$$(204) \leq T\frac{\ell_{f,1}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)}{2}, \tag{207}$$

and

$$(206) \leq T\frac{\ell_{f,1}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)}{2}. \tag{208}$$

Moreover, from Lemma D.7, we have

$$
\begin{aligned}
(205) = & \sum_{t=1}^{T} \left(f_t(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \right) \\
& + \sum_{t=1}^{T} \left(f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1})) \right) \\
& + \sum_{t=1}^{T} \left(f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1})) - f_t(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1})) \right) \\
\leq & \ \ell_{f,1} \sum_{t=1}^{T} \|\hat{\mathbf{y}}_t^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\| + \ell_{f,1} \sum_{t=1}^{T} \|\hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_{t+1})\| \\
& + \sum_{t=1}^{T} \left(f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1})) \right) \\
\leq & \ 2T\ell_{f,1} \frac{\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g} + \sum_{t=1}^{T} \left(f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1})) \right). \qquad (209)
\end{aligned}
$$

For the last term of the above inequality, we have

$$
\begin{aligned}
\sum_{t=1}^{T} \left(f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1})) \right) = & \ f_1(\mathbf{x}_1, \mathbf{y}_1^*(\mathbf{x}_1)) - f_T(\mathbf{x}_{T+1}, \mathbf{y}_T^*(\mathbf{x}_{T+1})) \\
& + \sum_{t=2}^{T} \big(f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - f_{t-1}(\mathbf{x}_t, \mathbf{y}_{t-1}^*(\mathbf{x}_t)) \big) \\
\leq & \ 2M + V_T,
\end{aligned}
$$

which implies that

$$
(205) \leq 2T\ell_{f,1} \frac{\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g} + 2M + V_T. \qquad (210)
$$

From (207), (208), and (210), we get the desired result. □

Lemma D.21. *Suppose that Assumptions 3.2 and 3.3 hold. Let $f_{t,\boldsymbol{\rho}}$ be defined as in (17). Then, for $\hat{\mathbf{d}}_t^{\mathbf{x}}$ generated by Algorithm 2, for all $t \in [T]$, we have*

$$
\begin{aligned}
\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right\|^2\right] \leq & \ 4\mathbb{E}\left[\left\|e_t^L\right\|^2\right] + 4\ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4 \\
& + 2M_f^2 \left(\mathbb{E}[\hat{\theta}_t^{\mathbf{y}}] + \mathbb{E}[\hat{\theta}_t^{\mathbf{v}}]\right) := A_t, \qquad (211)
\end{aligned}
$$

where e_t^L is defined in Lemma D.15, and $\hat{\theta}_t^{\mathbf{y}}$, $\hat{\theta}_t^{\mathbf{v}}$ are as defined in (146). Additionally, M_f is given in Lemma D.2.

Proof. From $\|a+b\|^2 \leq 2\left(\|a\|^2 + \|b\|^2\right)$, we get

$$
\begin{aligned}
& \mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right\|^2\right] \\
\leq & \ 2\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{x}}\right\|^2\right] \qquad (212a) \\
& + 2\mathbb{E}\left[\left\|\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{x}} - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right\|^2\right], \qquad (212b)
\end{aligned}
$$

where $\mathbf{d}^{\mathbf{x}}_{t,\boldsymbol{\rho}}$ is defined in (120c). From Lemma D.15, we have

$$(212a) \leq 4\mathbb{E}\left[\left\|e_t^L\right\|^2\right] + 4\ell_{g,2}^2\rho_{\mathbf{v}}^2 p^4. \tag{213}$$

Moreover, from Eq. (123a), we get

$$(212b) \leq 2M_f^2\left(\mathbb{E}[\hat{\theta}_t^{\mathbf{y}}] + \mathbb{E}[\hat{\theta}_t^{\mathbf{v}}]\right). \tag{214}$$

Substituting (213) and (214) into (212), we conclude the desired result. □

Lemma D.22. *Suppose Assumptions 3.2, 3.3, and 3.4 hold. Let the sequence of functions $\{f_{t,\boldsymbol{\rho}}\}_{t=1}^T$ be defined in(17), and $\mathcal{P}_{\mathcal{X},\alpha_t}$ be given in Definition B.1. Then, for any positive choice of step sizes as $\alpha_t \leq 1/4L_f$, for all $t \in [T]$, Algorithm 2 guarantees the following bound:*

$$\begin{aligned}
&\sum_{t=1}^T \left(\alpha_t - L_f\alpha_t^2\right)\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2\right] \\
&\leq 12M + 6V_T + \sum_{t=1}^T \left(6\alpha_t - 3L_f\alpha_t^2\right) A_t \\
&+ \sum_{t=1}^T \left(6\ell_{f,1}(1 + 2\frac{\ell_{g,1}}{\mu_g}) + \frac{3\ell_{f,1}\ell_{g,1}}{\mu_g}(\alpha_t - L_f\alpha_t^2)\right)\left(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2\right),
\end{aligned} \tag{215}$$

where V_T and A_t are respectively defined in Eq. (10) and Lemma D.21.

Proof. Due to the L_f-smoothness of f_t function by Lemma C.1, $f_{t,\boldsymbol{\rho}}$ is L_f-smooth as well. Hence,

$$\begin{aligned}
&f_{t,\boldsymbol{\rho}}(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1})) - f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) \\
&\leq \langle \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)), \mathbf{x}_{t+1} - \mathbf{x}_t\rangle + \frac{L_f}{2}\left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|^2 \\
&= -\alpha_t\left\langle \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)), \mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\rangle + \frac{L_f\alpha_t^2}{2}\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\|^2.
\end{aligned} \tag{216}$$

For the first term on the R.H.S of Eq. (216), we have that

$$\begin{aligned}
&-\mathbb{E}\left\langle \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)), \mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\rangle \\
&= -\mathbb{E}\left\langle \hat{\mathbf{d}}_t^{\mathbf{x}}, \mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\rangle \\
&- \mathbb{E}\left\langle \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - \hat{\mathbf{d}}_t^{\mathbf{x}}, \mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\rangle \\
&\leq -\frac{1}{2}\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right\|^2\right] \\
&\leq -\frac{1}{2}\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\|^2\right] + \frac{A_t}{2},
\end{aligned} \tag{217}$$

where the first inequality follows from Lemma B.8; the last inequality follows from Lemma D.21.

Plugging the bound (217) into (216), we have that

$$\begin{aligned}
&\mathbb{E}\left[f_{t,\boldsymbol{\rho}}(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1})) - f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right] \\
&\leq \frac{(L_f\alpha_t^2 - \alpha_t)}{2}\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\|^2\right] + \frac{\alpha_t A_t}{2},
\end{aligned}$$

which can be rearranged into

$$
\begin{aligned}
&(\alpha_t - L_f\alpha_t^2)\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\|^2\right] \\
&\leq 2\mathbb{E}\left[f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_{t,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))\right] + \alpha_t A_t.
\end{aligned} \tag{218}
$$

In addition, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2\right] \\
&\leq 3\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\hat{\mathbf{d}}_t^{\mathbf{x}}\right) - \mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right)\right\|^2\right] \\
&+ 3\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right) - \mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2\right] \\
&+ 3\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\|^2\right] \\
&\leq 3\mathbb{E}\left[\left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right\|^2\right] \\
&+ 3\mathbb{E}\left[\left\|\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right\|^2\right] \\
&+ 3\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\|^2\right],
\end{aligned}
$$

where the second inequaliy follows from non-expansiveness of the projection operator.

Then, from Lemma D.21 and Assumption 3.3, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2\right] \\
&\leq 3A_t + 3\ell_{f,1}\mathbb{E}\left[\left\|\hat{\mathbf{y}}_t^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right] + 3\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\|^2\right] \\
&\leq 3A_t + 3\ell_{f,1}\frac{\ell_{g,1}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)}{\mu_g} + 3\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\|^2\right],
\end{aligned} \tag{219}
$$

where the last inequality is by Lemma D.7.

Combining (218) and (219) and summing over $t = 1$ to T, we have

$$
\begin{aligned}
&\sum_{t=1}^{T}\left(\alpha_t - L_f\alpha_t^2\right)\mathbb{E}\left[\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2\right] \\
&\leq 6\sum_{t=1}^{T}\left(f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_{t,\boldsymbol{\rho}}(\mathbf{x}_{t+1},\hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))\right) \\
&+ \frac{3\ell_{f,1}\ell_{g,1}}{\mu_g}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)\sum_{t=1}^{T}\left(\alpha_t - L_f\alpha_t^2\right) + 3\sum_{t=1}^{T}\left(2\alpha_t - L_f\alpha_t^2\right)A_t \\
&\leq 12M + 6V_T + 6\ell_{f,1}\left(1 + 2\frac{\ell_{g,1}}{\mu_g}\right)T\left(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2\right) \\
&+ \frac{3\ell_{f,1}\ell_{g,1}}{\mu_g}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)\sum_{t=1}^{T}\left(\alpha_t - L_f\alpha_t^2\right) + 3\sum_{t=1}^{T}\left(2\alpha_t - L_f\alpha_t^2\right)A_t,
\end{aligned}
$$

where the second inequality is due to Lemma D.20. □

Lemma D.23. *Let the sequence $\{(\mathbf{x}_t,\mathbf{y}_t,\mathbf{v}_t)\}_{t=1}^{T}$ be generated by Algorithm 2.*

(a) Then, we have

$$\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \le 2\beta_t^2 \|e_t^{g_{\boldsymbol{\rho}}}\|^2 + 2\beta_t^2 \|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2,$$

where $e_t^{g_{\boldsymbol{\rho}}}$ is defined in (142).

(b) Suppose Assumptions 3.2, B2. and B3. hold. Then, we have

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \le\ & 4\alpha_t^2 \left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2 \\ &+ \frac{4\ell_{f,1}\ell_{g,1}\alpha_t^2(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g} + 2A_t\alpha_t^2, \end{aligned} \tag{220}$$

where A_t is defined in (211).

(c) Suppose Assumptions B1., B2. and B3. hold. Then, we have

$$\begin{aligned}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 \le\ & 2\delta_t^2 \|e_t^M\|^2 + 3d_2^2\ell_{f,1}^2\delta_t^2\rho_{\mathbf{r}}^2 \\ &+ (12\ell_{f,0}^2 + 6\ell_{g,1}^2 p^2)\delta_t^2 + 6\ell_{g,1}^2 \frac{\delta_t^2}{\rho_{\mathbf{v}^2}} \hat{\theta}_t^{\mathbf{y}}, \end{aligned}$$

where e_t^M and $\hat{\theta}_t^{\mathbf{y}}$ are defined in (158) and (146), respectively.

Proof. **For part (a):** From Algorithm 2, we have

$$\begin{aligned}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 &= \beta_t^2 \|\hat{\mathbf{d}}_t^{\mathbf{y}}\|^2 \\ &\le 2\beta_t^2 \|\hat{\mathbf{d}}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2 + 2\beta_t^2 \|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &= 2\beta_t^2 \|e_t^{g_{\boldsymbol{\rho}}}\|^2 + 2\beta_t^2 \|\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t)\|^2, \end{aligned} \tag{221}$$

and from (220), we get
For part (b):

From the update rule in Algorithm 2, we obtain

$$\begin{aligned}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 &= \alpha_t^2 \left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}\right)\right\|^2 \\ &\le 2\alpha_t^2 \Big(\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right)\right\|^2 \\ &+ \left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}\right) - \mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right)\right\|^2\Big) \\ &\le 2\alpha_t^2 \Big(\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right)\right\|^2 \\ &+ \left\|\hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right\|^2\Big) \\ &\le 2\alpha_t^2 \left(\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t; \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right)\right\|^2 + A_t\right), \end{aligned} \tag{222}$$

where the first inequality is by $(a+b)^2 \le 2a^2 + 2b^2$; the second inequality follows from non-expansiveness of the projection operator; and the last inequality follows from Lemma D.21.

The first term in the above inequality can be bounded as

$$\begin{aligned}
&\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right)\right\|^2\\
&\le 2\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t))\right)-\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2\\
&+2\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2\\
&\le 2\left\|\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\hat{\mathbf{y}}_t^*(\mathbf{x}_t))-\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right\|^2\\
&+2\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2\\
&\le 2\ell_{f,1}\|\hat{\mathbf{y}}_t^*(\mathbf{x}_t)-\mathbf{y}_t^*(\mathbf{x}_t)\|^2+2\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2\\
&\le 2\ell_{f,1}\frac{\ell_{g,1}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)}{\mu_g}+2\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2,
\end{aligned}\tag{223}$$

where the last inequality follows from Lemma D.7.

Based on (223) and (222), we get

$$\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2\le 2\alpha_t^2\left(2\left\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\right\|^2+\frac{2\ell_{f,1}\ell_{g,1}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)}{\mu_g}+A_t\right).$$

For part (c): Note that, we have

$$\begin{aligned}
\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2&=\delta_t^2\|\hat{\mathbf{d}}_t^{\mathbf{v}}\|^2\\
&\le 2\delta_t^2\|\hat{\mathbf{d}}_t^{\mathbf{v}}-\nabla_{\mathbf{y}}f_{t,\boldsymbol{\rho}}(\mathbf{z}_t)-\tilde{\nabla}_{\mathbf{y}}^2g_t(\mathbf{z}_t)\|^2+2\delta_t^2\|\nabla_{\mathbf{y}}f_{t,\boldsymbol{\rho}}(\mathbf{z}_t)+\tilde{\nabla}_{\mathbf{y}}^2g_t(\mathbf{z}_t)\|^2\\
&=2\delta_t^2\|e_t^M\|^2\\
&+2\delta_t^2\|\nabla_{\mathbf{y}}f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)+\frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{y}}g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t)-\nabla_{\mathbf{y}}g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t))\|^2\\
&\le 2\delta_t^2\|e_t^M\|^2+6\delta_t^2\|\nabla_{\mathbf{y}}f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\|^2\\
&+\frac{3\delta_t^2}{2\rho_{\mathbf{v}^2}}\|\nabla_{\mathbf{y}}g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t)\|^2+\frac{3\delta_t^2}{2\rho_{\mathbf{v}^2}}\|\nabla_{\mathbf{y}}g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t)\|^2.
\end{aligned}\tag{224}$$

From Assumption B3., Lemma B.3 and (8), we have

$$\begin{aligned}
\|\nabla_{\mathbf{y}}g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t)\|^2&\le\ell_{g,1}^2\|\mathbf{y}_t+\rho_{\mathbf{v}}\mathbf{v}_t-\hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\\
&\le 2\ell_{g,1}^2\|\rho_{\mathbf{v}}\mathbf{v}_t\|^2+2\ell_{g,1}^2\|\mathbf{y}_t-\hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2\\
&\le 2\ell_{g,1}^2\rho_{\mathbf{v}}^2p^2+2\ell_{g,1}^2\|\mathbf{y}_t-\hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2.
\end{aligned}\tag{225}$$

Similarly, we get

$$\|\nabla_{\mathbf{y}}g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t-\rho_{\mathbf{v}}\mathbf{v}_t)\|^2\le 2\ell_{g,1}^2\rho_{\mathbf{v}}^2p^2+2\ell_{g,1}^2\|\mathbf{y}_t-\hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2.\tag{226}$$

Moreover, from Eq. (133) and Assumption B1., we have

$$\begin{aligned}
\|\nabla_{\mathbf{y}}f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\|^2&\le 2\|\nabla_{\mathbf{y}}f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)-\nabla_{\mathbf{y}}f_t(\mathbf{x}_t,\mathbf{y}_t)\|^2+2\|\nabla_{\mathbf{y}}f_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\\
&\le\frac{d_2^2\ell_{f,1}^2\rho_{\mathbf{r}}^2}{2}+2\|\nabla_{\mathbf{y}}f_t(\mathbf{x}_t,\mathbf{y}_t)\|^2\\
&\le\frac{d_2^2\ell_{f,1}^2\rho_{\mathbf{r}}^2}{2}+2\ell_{f,0}^2.
\end{aligned}\tag{227}$$

Substituting (225), (226) and (227), into (224), we get

$$\begin{aligned}
\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2\le{}&2\delta_t^2\|e_t^M\|^2+3d_2^2\ell_{f,1}^2\delta_t^2\rho_{\mathbf{r}}^2\\
&+(12\ell_{f,0}^2+6\ell_{g,1}^2p^2)\delta_t^2+\frac{6\ell_{g,1}^2}{\rho_{\mathbf{v}^2}}\delta_t^2\|\mathbf{y}_t-\hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2.
\end{aligned}$$

$\square$

D.6. Proof of Theorem 4.2

Proof. Since $(1-\gamma_{t+1})^2 \leq 1-\gamma_{t+1}$ and $\gamma_{t+1} = c_\gamma \alpha_t$, from (143), we have

$$\begin{aligned}\mathbb{E}\|e_{t+1}^{g_\rho}\|^2 - \mathbb{E}\|e_t^{g_\rho}\|^2 \leq & -c_\gamma \alpha_t \mathbb{E}\|e_t^{g_\rho}\|^2 \\ & + 12(1-\gamma_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_{t-1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ & + 9d_2^2 \ell_{g,1}^2 (1-\gamma_{t+1})^2 \rho_{\mathbf{r}}^2 + 24 d_2 \ell_{g,1}^2 (1-\gamma_{t+1})^2 \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ & + 24 d_2 \ell_{g,1}^2 (1-\gamma_{t+1})^2 \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 + 2\frac{\hat{\sigma}_{g_\mathbf{y}}^2}{\bar{b}} \gamma_{t+1}^2. \end{aligned} \tag{228}$$

Since $(1-\eta_{t+1})^2 \leq 1-\eta_{t+1}$ and $\eta_{t+1} = c_\eta \alpha_t$, from (197), we have

$$\begin{aligned}\mathbb{E}\|e_{t+1}^{L}\|^2 - \mathbb{E}\|e_t^{L}\|^2 \leq & -c_\eta \alpha_t \mathbb{E}\|e_t^{L}\|^2 + 36\mathbb{E}\left\|\nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t)\right\|^2 \\ & + \left(18 d_1^2 \ell_{f,1}^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) d_1^2\right) \rho_{\mathbf{s}}^2 + 18 d_1^2 \ell_{g,1}^2 \frac{\rho_{\mathbf{s}}^2}{\rho_{\mathbf{v}}^2} \\ & + \frac{36}{\rho_{\mathbf{v}}^2} \mathbb{E}\|\nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\ & + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\ & + 27\ell_{g,1}^2 d_1 \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + 3(\frac{\hat{\sigma}_{g_\mathbf{x}}^2}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_\mathbf{x}}^2}{b}) \eta_{t+1}^2. \end{aligned} \tag{229}$$

Since $(1-\lambda_{t+1})^2 \leq 1-\lambda_{t+1}$ and $\lambda_{t+1} = c_\lambda \alpha_t$, from (160), we have

$$\begin{aligned}\mathbb{E}\|e_{t+1}^{M}\|^2 - \mathbb{E}\|e_t^{M}\|^2 \leq & -c_\lambda \alpha_t \mathbb{E}\|e_t^{M}\|^2 + 36\mathbb{E}\left\|\nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t)\right\|^2 \\ & + \left(18 d_2^2 \ell_{f,1}^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) d_2^2\right) \rho_{\mathbf{r}}^2 + 18 d_2^2 \ell_{g,1}^2 \frac{\rho_{\mathbf{r}}^2}{\rho_{\mathbf{v}}^2} \\ & + \frac{36}{\rho_{\mathbf{v}}^2} \mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\ & + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_2 \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_2 \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\ & + 27 d_2 \ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + 3(\frac{\hat{\sigma}_{g_\mathbf{y}}^2}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_\mathbf{y}}^2}{b}) \lambda_{t+1}^2. \end{aligned} \tag{230}$$

Combining the outcomes .
Let

$$\begin{aligned}\Lambda := & \Gamma \sum_{t=1}^{T} \left(\mathbb{E}[\hat{\theta}_{t+1}^{\mathbf{y}}] - \mathbb{E}[\hat{\theta}_t^{\mathbf{y}}]\right) + \Upsilon \sum_{t=1}^{T} \left(\mathbb{E}[\hat{\theta}_{t+1}^{\mathbf{v}}] - \mathbb{E}[\hat{\theta}_t^{\mathbf{v}}]\right) + \frac{1}{\Phi} \sum_{t=1}^{T} \left(\mathbb{E}\|e_{t+1}^{g_\rho}\|^2 - \mathbb{E}\|e_t^{g_\rho}\|^2\right) \\ & + \frac{1}{\Psi} \sum_{t=1}^{T} \left(\mathbb{E}\|e_{t+1}^{M}\|^2 - \mathbb{E}\|e_t^{M}\|^2\right) + \frac{1}{\Omega} \sum_{t=1}^{T} \left(\mathbb{E}\|e_{t+1}^{L}\|^2 - \mathbb{E}\|e_t^{L}\|^2\right). \end{aligned}$$

Here, we have

$$\Gamma = \frac{11M_f^2}{L_{\mu_g}c_\beta}, \quad \Upsilon = \frac{52M_f^2}{L_{\mu_g}c_\delta}, \quad \Phi = \max\left\{240\frac{d_2\ell_{g,1}^2}{L_f}, \frac{12d_2\ell_{g,1}^2L_{\mu_g}^2c_\beta^2}{L_fM_f^2}\right\},$$
$$\Psi = \max\left\{720\frac{d_2\ell_{f,1}^2}{L_f}, 27\frac{L_{\mu_g}}{\Upsilon L_f}\ell_{g,1}^2d_2c_\delta, \frac{144d_2\ell_{f,1}^2(\mu_g+\ell_{g,1})c_\beta}{L_f\Gamma}, \frac{36\ell_{f,1}^2d_2L_{\mu_g}^2c_\beta^2}{L_fM_f^2}\right\}, \tag{231}$$
$$\Omega = \max\left\{720\frac{d_1\ell_{f,1}^2}{L_f}, 27\frac{L_{\mu_g}}{\Upsilon L_f}\ell_{g,1}^2d_1c_\delta, \frac{144d_1\ell_{f,1}^2(\mu_g+\ell_{g,1})c_\beta}{L_f\Gamma}, \frac{36\ell_{f,1}^2d_1L_{\mu_g}^2c_\beta^2}{L_fM_f^2}\right\},$$

with

$$c_\beta \geq \sqrt{1760\frac{L_{\mathbf{y}}^2M_f^2}{L_{\mu_g}^2}}, \qquad c_\delta \geq \sqrt{33280\frac{\nu^2M_f^2}{L_{\mu_g}^2\mu_g^2}(1+2L_{\mathbf{y}}^2)},$$
$$c \geq \left(\max\left\{4L_f, c_\beta(\mu_g+\ell_{g,1}), \frac{48L_{\mu_g}^2d_2\ell_{g,1}^2c_\beta^2}{M_f^2\Phi}\right\}\right)^3 + 1, \tag{232}$$
$$c_{\mathbf{v}} = \max\left\{1080\ell_{g,1}^2, \frac{324}{M_f^2}\ell_{g,1}^4c_\delta^2, \frac{54L_{\mu_g}^2}{M_f^2}\ell_{g,1}^2c_\beta^2, \frac{216}{\Gamma}\ell_{g,1}^2c_\beta(\mu_g+\ell_{g,1})\right\}(\frac{d_2}{\Psi}+\frac{d_1}{\Omega}),$$
$$c_\gamma = \frac{26M_f^2\Phi}{L_{\mu_g}^2}, \qquad c_\eta = 26\Omega, \qquad c_\lambda = \frac{10\Upsilon}{L_{\mu_g}}c_\delta\Psi.$$

By adding (229), (228), (230), (147), and (184), along with (215) and considering the fact that α_t decreases with respect to t, and by applying Lemma D.23, we obtain:

$$
\begin{aligned}
&\sum_{t=1}^{T} A(\alpha_t,\beta_t,\delta_t)\mathbb{E}\left[\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\|^2\right]+\Lambda \\
&\le 12M+6V_T+\sum_{t=1}^{T}B(\alpha_t,\beta_t,\delta_t)\hat{\theta}_t^{\mathbf{v}}+\sum_{t=1}^{T}C(\alpha_t,\beta_t,\delta_t)\hat{\theta}_t^{\mathbf{y}} && \text{(233a)}\\
&+\frac{4\ell_{f,1}\ell_{g,1}}{\mu_g}\sum_{t=1}^{T}E(\beta_t,\delta_t,\rho_{\mathbf{v}})\alpha_t^2(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)+\sum_{t=1}^{T}L(\alpha_t,\beta_t,\delta_t)\mathbb{E}\|e_t^L\|^2 && \text{(233b)}\\
&+\frac{8\ell_{g,2}^2p^4\Upsilon}{L_{\mu_g}}\sum_{t=1}^{T}\delta_t\rho_{\mathbf{v}}^2+4\ell_{g,2}^2p^4\sum_{t=1}^{T}\left(6\alpha_t-3L_f\alpha_t^2+2\alpha_t^2E(\beta_t,\delta_t,\rho_{\mathbf{v}})\right)\rho_{\mathbf{v}}^2 && \text{(233c)}\\
&+\left(\frac{12}{L_{\mu_g}}\frac{\Gamma}{\beta_T}+\frac{48\nu^2}{L_{\mu_g}\mu_g^2}\frac{\Upsilon}{\delta_T}\right)H_{2,T}+\sum_{t=1}^{T}M(\alpha_t,\beta_t,\delta_t)\mathbb{E}\|e_t^M\|^2 && \text{(233d)}\\
&+\sum_{t=1}^{T}Q(\alpha_t,\beta_t,\delta_t)\mathbb{E}\|e_t^{g_{\rho}}\|^2+\sum_{t=1}^{T}S(\alpha_t,\beta_t,\delta_t)\mathbb{E}\left[\|\nabla_{\mathbf{y}}g_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t)\|^2\right] && \text{(233e)}\\
&+\sum_{t=1}^{T}Z(\alpha_t)\left(3d_2^2\ell_{f,1}^2\delta_t^2\rho_{\mathbf{r}}^2+(12\ell_{f,0}^2+6\ell_{g,1}^2p^2)\delta_t^2\right) && \text{(233f)}\\
&+\frac{36}{\Psi}D_{\mathbf{y},T}+\frac{36}{\Omega}D_{\mathbf{x},T}+\frac{12}{\Phi}G_{\mathbf{y},T}+\frac{18}{\Psi\rho_{\mathbf{v}}^2}G_{\mathbf{v},T}+\frac{18}{\Omega\rho_{\mathbf{v}}^2}G_{\mathbf{x},T} && \text{(233g)}\\
&+2\sum_{t=1}^{T}\frac{\gamma_{t+1}^2}{\Phi}\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}}+3\left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}\rho_{\mathbf{v}}^2}+\frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b}\right)\sum_{t=1}^{T+1}\frac{\lambda_{t+1}^2}{\Psi}+3\left(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{\bar{b}\rho_{\mathbf{v}}^2}+\frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b}\right)\sum_{t=1}^{T}\frac{\eta_{t+1}^2}{\Omega} && \text{(233h)}\\
&+R(\rho_{\mathbf{v}})T\rho_{\mathbf{r}}^2+\acute{R}(\rho_{\mathbf{v}})T\rho_{\mathbf{s}}^2+18d_1^2\ell_{g,1}^2\frac{T\rho_{\mathbf{s}}^2}{\Omega\rho_{\mathbf{v}}^2}+18d_2^2\ell_{g,1}^2\frac{T\rho_{\mathbf{r}}^2}{\Psi\rho_{\mathbf{v}}^2}+\sum_{t=1}^{T}D(\alpha_t,\beta_t,\delta_t)\left(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2\right). && \text{(233i)}
\end{aligned}
$$

Here

$$
\begin{aligned}
E(\beta_t,\delta_t,\rho_{\mathbf{v}}) &:= \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}}\frac{\Gamma}{\beta_t}+\frac{16\nu^2}{L_{\mu_g}\mu_g^2}(2L_{\mathbf{y}}^2+1)\frac{\Upsilon}{\delta_t}\\
&\quad+24d_2\frac{\ell_{g,1}^2}{\Phi}+6\left(12\ell_{f,1}^2+\frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}\right)\left(\frac{d_2}{\Psi}+\frac{d_1}{\Omega}\right),\\
A(\alpha_t,\beta_t,\delta_t) &:= \alpha_t-L_f\alpha_t^2-4E(\beta_t,\delta_t,\rho_{\mathbf{v}})\alpha_t^2,\\
B(\alpha_t,\beta_t,\delta_t) &:= -\frac{L_{\mu_g}}{4}\Upsilon\delta_t+2M_f^2\left(6\alpha_t-3L_f\alpha_t^2+2\alpha_t^2E(\beta_t,\delta_t,\rho_{\mathbf{v}})\right),\\
Z(\alpha_t) &:= 27\ell_{g,1}^2\left(\frac{d_2}{\Psi}+\frac{d_1}{\Omega}\right),\\
C(\alpha_t,\beta_t,\delta_t) &:= -\frac{L_{\mu_g}}{2}\Gamma\beta_t+Z(\alpha_t)6\ell_{g,1}^2\frac{\delta_t^2}{\rho_{\mathbf{v}}^2}+2M_f^2\left(6\alpha_t-3L_f\alpha_t^2+2\alpha_t^2E(\beta_t,\delta_t,\rho_{\mathbf{v}})\right).
\end{aligned} \tag{234}
$$

Moreover,

$$\begin{aligned}
M(\alpha_t,\beta_t,\delta_t) &:= -\frac{\lambda_{t+1}}{\Psi} + Z(\alpha_t)2\delta_t^2 + \frac{8\Upsilon}{L_{\mu_g}}\delta_t,\\
D(\alpha_t,\beta_t,\delta_t) &:= 6\ell_{f,1}(1+2\frac{\ell_{g,1}}{\mu_g}) + \frac{3\ell_{f,1}\ell_{g,1}}{\mu_g}(\alpha_t - L_f\alpha_t^2)\\
&\quad + \frac{24\ell_{g,1}}{L_{\mu_g}\mu_g}\frac{\Gamma}{\beta_t} + \frac{96\ell_{g,1}\nu^2}{L_{\mu_g}\mu_g^3}\frac{\Upsilon}{\delta_t},\\
F(\alpha_t) &:= 24d_2\frac{\ell_{g,1}^2}{\Phi} + (72\ell_{f,1}^2 + \frac{27\ell_{g,1}^2}{\rho_{\mathbf{v}}^2})(\frac{d_2}{\Psi}+\frac{d_1}{\Omega}),\\
S(\alpha_t,\beta_t,\delta_t) &:= -\frac{2\beta_t\Gamma}{\mu_g+\ell_{g,1}} + \beta_t^2\Gamma + 2F(\alpha_t)\beta_t^2,\\
Q(\alpha_t,\beta_t,\delta_t) &:= \frac{2}{L_{\mu_g}}\Gamma\beta_t - \frac{\gamma_{t+1}}{\Phi} + 2F(\alpha_t)\beta_t^2,\\
R(\rho_{\mathbf{v}}) &:= 9d_2^2\frac{\ell_{g,1}^2}{\Phi} + 18d_2^2\frac{\ell_{f,1}^2}{\Psi} + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})\frac{d_2^2}{\Psi},\\
\acute{R}(\rho_{\mathbf{v}}) &:= 18d_1^2\frac{\ell_{f,1}^2}{\Omega} + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})\frac{d_1^2}{\Omega},\\
L(\alpha_t,\beta_t,\delta_t) &:= -\frac{\eta_{t+1}}{\Omega} + 4\left(6\alpha_t - 3L_f\alpha_t^2 + 2\alpha_t^2E(\beta_t,\delta_t,\rho_{\mathbf{v}})\right).
\end{aligned} \tag{235}$$

We then provide bounds for the terms in (233a)-(233i).
Note that, we have

$$\begin{aligned}
E(\beta_t,\delta_t,\rho_{\mathbf{v}}) &:= \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}}\frac{\Gamma}{\beta_t} + \frac{16\nu^2}{L_{\mu_g}\mu_g^2}(2L_{\mathbf{y}}^2+1)\frac{\Upsilon}{\delta_t}\\
&\quad + 24d_2\frac{\ell_{g,1}^2}{\Phi} + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})(\frac{d_2}{\Psi}+\frac{d_1}{\Omega}),
\end{aligned}$$

which together with $\beta_t = c_\beta\alpha_t$, $\delta_t = c_\delta\alpha_t$, we have

$$\begin{aligned}
\alpha_t^2E(\beta_t,\delta_t,\rho_{\mathbf{v}}) &= \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}}\frac{\Gamma\alpha_t^2}{\beta_t} + \frac{16\nu^2}{L_{\mu_g}\mu_g^2}(2L_{\mathbf{y}}^2+1)\frac{\Upsilon\alpha_t^2}{\delta_t}\\
&\quad + 24d_2\frac{\ell_{g,1}^2}{\Phi}\alpha_t^2 + (72\ell_{f,1}^2\alpha_t^2 + \frac{27\ell_{g,1}^2}{\rho_{\mathbf{v}}^2}\alpha_t^2)(\frac{d_2}{\Psi}+\frac{d_1}{\Omega})\\
&\le \frac{44L_{\mathbf{y}}^2}{L_{\mu_g}^2}M_f^2\frac{\alpha_t}{c_\beta^2} + \frac{832\nu^2}{L_{\mu_g}^2\mu_g^2}(1+2L_{\mathbf{y}}^2)M_f^2\frac{\alpha_t}{c_\delta^2}\\
&\quad + 6\frac{d_2\ell_{g,1}^2}{L_f\Phi}\alpha_t + (\frac{18\ell_{f,1}^2}{L_f}\alpha_t + \frac{27\ell_{g,1}^2}{c_{\mathbf{v}}}\alpha_t)(\frac{d_2}{\Psi}+\frac{d_1}{\Omega})\\
&\le \frac{\alpha_t}{8},
\end{aligned} \tag{236}$$

where the first inequality is by $\Gamma = \frac{11M_f^2}{L_{\mu_g}c_\beta}$, $\Upsilon = \frac{52M_f^2}{L_{\mu_g}c_\delta}$ in (231), $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}}\alpha_t$ and $\alpha_t \le 1/4L_f$; the second inequality follows from $c_\beta \ge \sqrt{1760\frac{L_{\mathbf{y}}^2M_f^2}{L_{\mu_g}^2}}$, $c_\delta \ge \sqrt{33280\frac{\nu^2M_f^2}{L_{\mu_g}^2\mu_g^2}(1+2L_{\mathbf{y}}^2)}$, in (232); and $\Phi = 240\frac{d_2\ell_{g,1}^2}{L_f}$, $\Psi = 720\frac{d_2\ell_{f,1}^2}{L_f}$, $\Omega = 720\frac{d_1\ell_{f,1}^2}{L_f}$ and $c_{\mathbf{v}} \ge 1080\ell_{g,1}^2(\frac{d_2}{\Psi}+\frac{d_1}{\Omega})$ in (231).

Moreover, we have

$$A(\alpha_t,\beta_t,\delta_t) = \alpha_t - L_f\alpha_t^2 - 4E(\beta_t,\delta_t,\rho_{\mathbf{v}})\alpha_t^2$$
$$\geq \alpha_t - L_f\alpha_t^2 - \frac{\alpha_t}{2}$$
$$\geq \frac{\alpha_t}{4}, \tag{237}$$

where the last inequality is by $\alpha_t \leq 1/4L_f$ in (232).

Bounding (233a) .

From $\delta_t = c_\delta\alpha_t$, we have

$$B(\alpha_t,\beta_t,\delta_t) = -\frac{L_{\mu_g}}{4}\Upsilon\delta_t + 2M_f^2\left(6\alpha_t - 3L_f\alpha_t^2 + 2\alpha_t^2E(\beta_t,\delta_t,\rho_{\mathbf{v}})\right)$$
$$\leq -\frac{L_{\mu_g}}{4}\Upsilon c_\delta\alpha_t + 12M_f^2\alpha_t - 6M_f^2L_f\alpha_t^2 + \frac{M_f^2}{2}\alpha_t$$
$$\leq \left(-\frac{L_{\mu_g}}{4}\Upsilon c_\delta + \frac{25}{2}M_f^2\right)\alpha_t$$
$$\leq -\frac{1}{2}M_f^2\alpha_t, \tag{238}$$

where the first inequality follows from (236); the last inequality is by $\Upsilon = \frac{52M_f^2}{L_{\mu_g}c_\delta}$ in (231).

From (234), we obtain

$$Z(\alpha_t) = 27\ell_{g,1}^2\left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}\right).$$

Thus, from $\beta_t = c_\beta\alpha_t$, $\delta_t = c_\delta\alpha_t$ and $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}}\alpha_t$, we have

$$C(\alpha_t,\beta_t,\delta_t) = -\frac{L_{\mu_g}}{2}\Gamma\beta_t + 162\left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}\right)\ell_{g,1}^4\frac{\delta_t^2}{\rho_{\mathbf{v}}^2}$$
$$+ 2M_f^2\left(6\alpha_t - 3L_f\alpha_t^2 + 2\alpha_t^2E(\beta_t,\delta_t,\rho_{\mathbf{v}})\right)$$
$$\leq -\frac{L_{\mu_g}}{2}\Gamma c_\beta\alpha_t + 162\left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}\right)\ell_{g,1}^4\frac{c_\delta^2}{c_{\mathbf{v}}}\alpha_t + \frac{9}{2}M_f^2\alpha_t$$
$$= -\frac{11}{2}M_f^2\alpha_t + 162\left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}\right)\ell_{g,1}^4\frac{c_\delta^2}{c_{\mathbf{v}}}\alpha_t + \frac{9}{2}M_f^2\alpha_t$$
$$\leq -\frac{1}{2}M_f^2\alpha_t, \tag{239}$$

where the first inequality follows from (236); the second equality follows from $\Gamma = \frac{11M_f^2}{L_{\mu_g}c_\beta}$ in (231); the last inequality is by $c_{\mathbf{v}} \geq \frac{324}{M_f^2}\ell_{g,1}^4(\frac{d_2}{\Psi} + \frac{d_1}{\Omega})c_\delta^2$.

Thus, from (238) and (239), we get

$$(233a) \leq \mathcal{O}\left(V_T\right). \tag{240}$$

Bounding (233b) .

From (236), we also obtain

$$
\begin{aligned}
&\frac{4\ell_{f,1}\ell_{g,1}}{\mu_g}\sum_{t=1}^{T}E(\beta_t,\delta_t,\rho_{\mathbf{v}})\alpha_t^2(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)\\
&\leq \frac{4\ell_{f,1}\ell_{g,1}}{\mu_g}\sum_{t=1}^{T}\frac{\alpha_t}{8}(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)\\
&=\mathcal{O}\left(\sum_{t=1}^{T}\alpha_t(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)\right).
\end{aligned}
\tag{241}
$$

From (235) and $\eta_{t+1}=c_\eta\alpha_t$, we have

$$
\begin{aligned}
L(\alpha_t,\beta_t,\delta_t)&=-\frac{\eta_{t+1}}{\Omega}+4\left(6\alpha_t-3L_f\alpha_t^2+2\alpha_t^2E(\beta_t,\delta_t,\rho_{\mathbf{v}})\right)\\
&\leq-\frac{c_\eta}{\Omega}\alpha_t+25\alpha_t\\
&\leq-\alpha_t,
\end{aligned}
$$

where the last inequality is by $c_\eta\geq 26\Omega$ and (236).
Thus, we get

$$
\sum_{t=1}^{T}L(\alpha_t,\beta_t,\delta_t)\mathbb{E}\|e_t^L\|^2\leq 0. \tag{242}
$$

From (242) and (241), we have

$$
(233b)\leq\mathcal{O}\left(\sum_{t=1}^{T}\alpha_t(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)\right). \tag{243}
$$

Bounding (233c) .
From $\delta_t=c_\delta\alpha_t$ and (236), we have

$$
\begin{aligned}
&\frac{8\ell_{g,2}^2p^4\Upsilon}{L_{\mu_g}}\sum_{t=1}^{T}\delta_t\rho_{\mathbf{v}}^2+4\ell_{g,2}^2p^4\sum_{t=1}^{T}\left(6\alpha_t-3L_f\alpha_t^2+2\alpha_t^2E(\beta_t,\delta_t,\rho_{\mathbf{v}})\right)\rho_{\mathbf{v}}^2\\
&\leq\frac{8\ell_{g,2}^2p^4\Upsilon}{L_{\mu_g}}\sum_{t=1}^{T}c_\delta\alpha_t\rho_{\mathbf{v}}^2+4\ell_{g,2}^2p^4\sum_{t=1}^{T}\frac{25}{4}\alpha_t\rho_{\mathbf{v}}^2.
\end{aligned}
$$

Thus, from $\rho_{\mathbf{v}}^2=c_{\mathbf{v}}\alpha_t$, we have

$$
(233c)\leq\mathcal{O}\left(\sum_{t=1}^{T}\alpha_t^2\right). \tag{244}
$$

Bounding (233d) .
From (234), we obtain

$$
Z(\alpha_t)=27\ell_{g,1}^2\left(\frac{d_2}{\Psi}+\frac{d_1}{\Omega}\right). \tag{245}
$$

From (235), $\lambda_{t+1} = c_\lambda \alpha_t$ and $\delta_t = c_\delta \alpha_t$, we have

$$\begin{aligned}
M(\alpha_t, \beta_t, \delta_t) &= -\frac{\lambda_{t+1}}{\Psi} + Z(\alpha_t) 2\delta_t^2 + \frac{8\Upsilon}{L_{\mu_g}} \delta_t \\
&= -\frac{c_\lambda \alpha_t}{\Psi} + 27\ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right) 2c_\delta^2 \alpha_t^2 + \frac{8\Upsilon}{L_{\mu_g}} c_\delta \alpha_t \\
&\leq -\frac{2\Upsilon}{L_{\mu_g}} c_\delta \alpha_t + \frac{27}{4L_f} \ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right) 2c_\delta^2 \alpha_t \\
&\leq -\frac{\Upsilon}{L_{\mu_g}} c_\delta \alpha_t,
\end{aligned}$$

where the first inequality is by $c_\lambda \geq \frac{10\Upsilon}{L_{\mu_g}} c_\delta \Psi$ and $\alpha_t \leq 1/4L_f$; the last inequality follows from $\Psi \geq 27 \frac{L_{\mu_g}}{\Upsilon L_f} \ell_{g,1}^2 d_2 c_\delta$ and $\Omega \geq 27 \frac{L_{\mu_g}}{\Upsilon L_f} \ell_{g,1}^2 d_1 c_\delta$.

Since $\beta_t = c_\beta \alpha_t$ and $\delta_t = c_\delta \alpha_t$, we get

$$\begin{aligned}
(233d) &= \left(\frac{12}{L_{\mu_g}} \frac{\Gamma}{\beta_T} + \frac{48\nu^2}{L_{\mu_g} \mu_g^2} \frac{\Upsilon}{\delta_T} \right) H_{2,T} + \sum_{t=1}^{T} M(\alpha_t, \beta_t, \delta_t) \mathbb{E} \| e_t^M \|^2 \\
&\leq \mathcal{O} \left(\frac{H_{2,T}}{\alpha_T} \right).
\end{aligned} \tag{246}$$

Bounding (233e) .
From (235), we have

$$F(\alpha_t) = 24d_2 \frac{\ell_{g,1}^2}{\Phi} + (72\ell_{f,1}^2 + \frac{27\ell_{g,1}^2}{\rho_{\mathbf{v}}^2})(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \tag{247}$$

From (235), $\gamma_{t+1} = c_\gamma \alpha_t$, $\beta_t = c_\beta \alpha_t$, we have

$$\begin{aligned}
&Q(\alpha_t, \beta_t, \delta_t) = -\frac{\gamma_{t+1}}{\Phi} + \frac{2}{L_{\mu_g}} \Gamma \beta_t + 2F(\alpha_t) \beta_t^2 \\
&= -\frac{c_\gamma \alpha_t}{\Phi} + \frac{22M_f^2}{L_{\mu_g}^2} \alpha_t + \left(24d_2 \frac{\ell_{g,1}^2}{\Phi} + (72\ell_{f,1}^2 + \frac{27\ell_{g,1}^2}{c_{\mathbf{v}} \alpha_t})(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \alpha_t^2 \\
&\leq -\frac{4M_f^2}{L_{\mu_g}^2} \alpha_t + \left(24d_2 \frac{\ell_{g,1}^2}{\Phi} \alpha_t^2 + (72\ell_{f,1}^2 \alpha_t^2 + \frac{27\ell_{g,1}^2 \alpha_t}{c_{\mathbf{v}}})(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \\
&\leq -\frac{4M_f^2}{L_{\mu_g}^2} \alpha_t + \left(\frac{6d_2}{L_f} \frac{\ell_{g,1}^2}{\Phi} \alpha_t + (\frac{18}{L_f} \ell_{f,1}^2 \alpha_t + \frac{27\ell_{g,1}^2 \alpha_t}{c_{\mathbf{v}}})(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \\
&\leq -\frac{M_f^2}{L_{\mu_g}^2} \alpha_t,
\end{aligned} \tag{248}$$

where the first equality is by $\Gamma = \frac{11M_f^2}{L_{\mu_g} c_\beta}$ and $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}} \alpha_t$; the first inequality follows from $c_\gamma \geq \frac{26M_f^2 \Phi}{L_{\mu_g}^2}$; the second inequality is by $\alpha_t \leq 1/4L_f$; the last inequality follows from $c_{\mathbf{v}} \geq \frac{54L_{\mu_g}^2}{M_f^2} \ell_{g,1}^2 (\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) c_\beta^2$, $\Phi \geq \frac{12d_2 \ell_{g,1}^2 L_{\mu_g}^2 c_\beta^2}{L_f M_f^2}$,and $\Psi \geq \frac{36\ell_{f,1}^2 d_2 L_{\mu_g}^2 c_\beta^2}{L_f M_f^2}$, and $\Omega \geq \frac{36\ell_{f,1}^2 d_1 L_{\mu_g}^2 c_\beta^2}{L_f M_f^2}$.

From (235), $\beta_t = c_\beta \alpha_t$, $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}} \alpha_t$ and (247), we have

$$
\begin{aligned}
S(\alpha_t, \beta_t, \delta_t) &= -\frac{2\beta_t \Gamma}{\mu_g + \ell_{g,1}} + \beta_t^2 \Gamma + 2F(\alpha_t)\beta_t^2 \\
&= -\frac{2c_\beta \alpha_t \Gamma}{\mu_g + \ell_{g,1}} + c_\beta^2 \alpha_t^2 \Gamma + \left(24 d_2 \frac{\ell_{g,1}^2}{\Phi} + (72\ell_{f,1}^2 + \frac{27\ell_{g,1}^2}{c_{\mathbf{v}} \alpha_t})(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \alpha_t^2 \\
&\leq -\frac{c_\beta \alpha_t \Gamma}{\mu_g + \ell_{g,1}} + \left(24 d_2 \frac{\ell_{g,1}^2}{\Phi} \alpha_t^2 + (72\ell_{f,1}^2 \alpha_t^2 + \frac{27\ell_{g,1}^2 \alpha_t}{c_{\mathbf{v}}})(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \\
&\leq -\frac{c_\beta \alpha_t \Gamma}{\mu_g + \ell_{g,1}} + \left(\frac{6 d_2}{L_f} \frac{\ell_{g,1}^2}{\Phi} \alpha_t + (\frac{18}{L_f} \ell_{f,1}^2 \alpha_t + \frac{27\ell_{g,1}^2 \alpha_t}{c_{\mathbf{v}}})(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \\
&\leq -\frac{c_\beta \alpha_t \Gamma}{4(\mu_g + \ell_{g,1})},
\end{aligned}
\tag{249}
$$

where the first inequality follows from $\alpha_t \leq 1/c_\beta(\mu_g + \ell_{g,1})$; the second inequality is by $\alpha \leq 1/4L_f$; the last inequality is by $c_{\mathbf{v}} \geq \frac{216}{\Gamma} \ell_{g,1}^2 (\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) c_\beta (\mu_g + \ell_{g,1})$ and $\Phi \geq \frac{24 d_2 \ell_{g,1}^2 (\mu_g + \ell_{g,1})}{L_f c_\beta \Gamma}$, and $\Psi \geq \frac{144 d_2 \ell_{f,1}^2 (\mu_g + \ell_{g,1}) c_\beta}{L_f \Gamma}$, and $\Omega \geq \frac{144 d_1 \ell_{f,1}^2 (\mu_g + \ell_{g,1}) c_\beta}{L_f \Gamma}$. Thus, we get

$$
(233e) = \sum_{t=1}^{T} Q(\alpha_t, \beta_t, \delta_t) \mathbb{E} \| e_t^{g_{\boldsymbol{\rho}}} \|^2 + \sum_{t=1}^{T} S(\alpha_t, \beta_t, \delta_t) \mathbb{E} \left[\| \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}_t, \mathbf{y}_t) \|^2 \right] \leq 0. \tag{250}
$$

Bounding (233f) .
From (234), we obtain

$$
Z(\alpha_t) = 27\ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right).
$$

Thus, from $\delta_t = c_\delta \alpha_t$, we have

$$
\begin{aligned}
(233f) &= \sum_{t=1}^{T} Z(\alpha_t) \left(3 d_2^2 \ell_{f,1}^2 \delta_t^2 \rho_{\mathbf{r}}^2 + (12\ell_{f,0}^2 + 6\ell_{g,1}^2 p^2) \delta_t^2 \right) \\
&= \sum_{t=1}^{T} 27\ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right) \left(3 d_2^2 \ell_{f,1}^2 \rho_{\mathbf{r}}^2 + (12\ell_{f,0}^2 + 6\ell_{g,1}^2 p^2) \right) c_\delta^2 \alpha_t^2 \\
&= \mathcal{O} \left(\sum_{t=1}^{T} (d_1 + d_2)(\alpha_t^2 \rho_{\mathbf{r}}^2 + \alpha_t^2) \right).
\end{aligned}
\tag{251}
$$

Bounding (233g) . From $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}} \alpha_t$, we have

$$
\begin{aligned}
(233g) &= \frac{36}{\Psi} D_{\mathbf{y},T} + \frac{36}{\Omega} D_{\mathbf{x},T} + \frac{12}{\Phi} G_{\mathbf{y},T} + \frac{36}{\Psi \rho_{\mathbf{v}}^2} G_{\mathbf{v},T} + \frac{36}{\Omega \rho_{\mathbf{v}}^2} G_{\mathbf{x},T} \\
&= \mathcal{O} \left(D_{\mathbf{y},T} + D_{\mathbf{x},T} + G_{\mathbf{y},T} + \frac{1}{\alpha_T} (G_{\mathbf{v},T} + G_{\mathbf{x},T}) \right).
\end{aligned}
\tag{252}
$$

Bounding (233h) . From $\gamma_{t+1} = c_\gamma \alpha_t$, $\eta_{t+1} = c_\eta \alpha_t$, $\lambda_{t+1} = c_\lambda \alpha_t$ and $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}} \alpha_t$, we have

$$
\begin{aligned}
(233h) &= 2\sum_{t=1}^{T} \frac{\gamma_{t+1}^2}{\Phi} \frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}} + 3\left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b}\right) \sum_{t=1}^{T+1} \frac{\lambda_{t+1}^2}{\Psi} + 3\left(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b}\right) \sum_{t=1}^{T} \frac{\eta_{t+1}^2}{\Omega} \\
&= 2\sum_{t=1}^{T} \frac{c_\gamma^2 \alpha_t^2}{\Phi} \frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}} + 3\left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b}\right) \sum_{t=1}^{T+1} \frac{c_\lambda^2 \alpha_t^2}{\Psi} + 3\left(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{\bar{b}\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b}\right) \sum_{t=1}^{T} \frac{c_\eta^2 \alpha_t^2}{\Omega} \\
&= \mathcal{O}\left(\left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}} + \frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\bar{b}\alpha_t} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b} + \frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{\bar{b}\alpha_t} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b} \right) \sum_{t=1}^{T} \alpha_t^2 \right).
\end{aligned} \tag{253}
$$

Bounding (233i) . From $\beta_t = c_\beta \alpha_t$, $\delta_t = c_\delta \alpha_t$, we have

$$
\begin{aligned}
D(\alpha_t, \beta_t, \delta_t) &= 6\ell_{f,1}\left(1 + 2\frac{\ell_{g,1}}{\mu_g}\right) + \frac{3\ell_{f,1}\ell_{g,1}}{\mu_g}(\alpha_t - L_f \alpha_t^2) + \frac{24\ell_{g,1}}{L_{\mu_g}\mu_g} \frac{\Gamma}{\beta_t} + \frac{96\ell_{g,1}\nu^2}{L_{\mu_g}\mu_g^3} \frac{\Upsilon}{\delta_t} \\
&= 6\ell_{f,1}\left(1 + 2\frac{\ell_{g,1}}{\mu_g}\right) + \frac{3\ell_{f,1}\ell_{g,1}}{\mu_g}(\alpha_t - L_f \alpha_t^2) + \frac{24\ell_{g,1}}{L_{\mu_g}\mu_g} \frac{\Gamma}{c_\beta \alpha_t} + \frac{96\ell_{g,1}\nu^2}{L_{\mu_g}\mu_g^3} \frac{\Upsilon}{c_\delta \alpha_t} \\
&= \mathcal{O}\left(\alpha_t + \frac{1}{\alpha_t}\right),
\end{aligned}
$$

and

$$
\sum_{t=1}^{T} D(\alpha_t, \beta_t, \delta_t)\left(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2\right) := \mathcal{O}\left(\sum_{t=1}^{T} \left(\alpha_t + \frac{1}{\alpha_t}\right)\left(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2\right)\right). \tag{254}
$$

Moreover, we have

$$
\begin{aligned}
R(\rho_{\mathbf{v}}) &= 9d_2^2 \frac{\ell_{g,1}^2}{\Phi} + 18d_2^2 \frac{\ell_{f,1}^2}{\Psi} + 6\left(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}\right)\frac{d_2^2}{\Psi} = \mathcal{O}\left(\left(1 + \frac{1}{\rho_{\mathbf{v}}^2}\right)d_2^2\right), \\
\acute{R}(\rho_{\mathbf{v}}) &= 18d_1^2 \frac{\ell_{f,1}^2}{\Omega} + 6\left(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}\right)\frac{d_1^2}{\Omega} = \mathcal{O}\left(\left(1 + \frac{1}{\rho_{\mathbf{v}}^2}\right)d_1^2\right),
\end{aligned}
$$

which, implies that

$$
\begin{aligned}
& R(\rho_{\mathbf{v}})T\rho_{\mathbf{r}}^2 + \acute{R}(\rho_{\mathbf{v}})T\rho_{\mathbf{s}}^2 + 18d_1^2\ell_{g,1}^2 \frac{T\rho_{\mathbf{s}}^2}{\Omega\rho_{\mathbf{v}}^2} + 18d_2^2\ell_{g,1}^2 \frac{T\rho_{\mathbf{r}}^2}{\Psi\rho_{\mathbf{v}}^2} \\
&= \mathcal{O}\left(\left(1 + \frac{1}{\rho_{\mathbf{v}}^2}\right)T(d_1^2\rho_{\mathbf{s}}^2 + d_2^2\rho_{\mathbf{r}}^2) + \frac{T}{\rho_{\mathbf{v}}^2}(d_2^2\rho_{\mathbf{r}}^2 + d_1^2\rho_{\mathbf{s}}^2)\right).
\end{aligned} \tag{255}
$$

From (254), (255) and $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}} \alpha_t$, we get

$$
(233i) \leq \mathcal{O}\left(\sum_{t=1}^{T} \left(\alpha_t + \frac{1}{\alpha_t}\right)\left(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2\right) + \left(1 + \frac{1}{\alpha_T}\right)T(d_2^2\rho_{\mathbf{r}}^2 + d_1^2\rho_{\mathbf{s}}^2)\right). \tag{256}
$$

Combining the outcomes (233i) . Combining inequalities (240), (243), (244), (246), (250), (251), (252), (253), and (256) leads to

$$
\begin{aligned}
&\frac{\alpha_T}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\|^2\right]+\Lambda\\
&\le\mathcal{O}\left(V_T+\sum_{t=1}^{T}\alpha_t(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)+\sum_{t=1}^{T}\alpha_t^2+\frac{H_{2,T}}{\alpha_T}+\sum_{t=1}^{T}(d_1+d_2)(\alpha_t^2\rho_{\mathbf{r}}^2+\alpha_t^2)\right)\\
&+\mathcal{O}\left(D_{\mathbf{y},T}+D_{\mathbf{x},T}+G_{\mathbf{y},T}+\frac{1}{\alpha_T}(G_{\mathbf{v},T}+G_{\mathbf{x},T})\right)\\
&+\mathcal{O}\left(\sum_{t=1}^{T}\left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2\alpha_t^2}{\bar{b}}+\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2\alpha_t}{\bar{b}}+\frac{\hat{\sigma}_{f_{\mathbf{y}}}^2\alpha_t^2}{b}+\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2\alpha_t}{\bar{b}}+\frac{\hat{\sigma}_{f_{\mathbf{x}}}^2\alpha_t^2}{b}\right)\right)\\
&+\mathcal{O}\left(\sum_{t=1}^{T}(\alpha_t+\frac{1}{\alpha_t})\left(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2\right)+(1+\frac{1}{\alpha_T})T(d_2^2\rho_{\mathbf{r}}^2+d_1^2\rho_{\mathbf{s}}^2)\right).
\end{aligned}
$$

From the definition of Λ in (102), we have

$$
\begin{aligned}
-\Lambda&=\Gamma\sum_{t=1}^{T}\left(\mathbb{E}[\theta_t^{\mathbf{y}}]-\mathbb{E}[\theta_{t+1}^{\mathbf{y}}]\right)+\Upsilon\sum_{t=1}^{T}\left(\mathbb{E}[\theta_t^{\mathbf{v}}]-\mathbb{E}[\theta_{t+1}^{\mathbf{v}}]\right)+\frac{1}{\Phi}\sum_{t=1}^{T}\left(\mathbb{E}\|e_t^g\|^2-\mathbb{E}\|e_{t+1}^g\|^2\right)\\
&+\frac{1}{\Psi}\sum_{t=1}^{T}\left(\mathbb{E}\|e_t^{\mathbf{v}}\|^2-\mathbb{E}\|e_{t+1}^{\mathbf{v}}\|^2\right)+\frac{1}{\Omega}\sum_{t=1}^{T}\left(\mathbb{E}\|e_t^f\|^2-\mathbb{E}\|e_{t+1}^f\|^2\right)\\
&\le\Gamma\theta_1^{\mathbf{y}}+\Upsilon\theta_1^{\mathbf{v}}+\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{\Phi}+\frac{\hat{\sigma}_{g_{\mathbf{yy}}}^2+\hat{\sigma}_{f_{\mathbf{y}}}^2}{\Psi}+\frac{\hat{\sigma}_{g_{\mathbf{xy}}}^2+\hat{\sigma}_{f_{\mathbf{x}}}^2}{\Omega}.
\end{aligned}
\tag{257}
$$

From (23), we have

$$
\hat{\sigma}^2:=\hat{\sigma}_{g_{\mathbf{y}}}^2+\hat{\sigma}_{g_{\mathbf{yy}}}^2+\hat{\sigma}_{f_{\mathbf{y}}}^2+\hat{\sigma}_{g_{\mathbf{xy}}}^2+\hat{\sigma}_{f_{\mathbf{x}}}^2.
$$

From (26), we have

$$
\begin{aligned}
&\alpha_t=\frac{1}{(d_1+d_2)^{3/4}(c+t)^{1/3}},\quad\beta_t=c_\beta\alpha_t,\quad\delta_t=c_\delta\alpha_t,\quad\rho_{\mathbf{v}}^2=c_{\mathbf{v}}\alpha_t,\\
&\gamma_{t+1}=c_\gamma\alpha_t,\quad\eta_{t+1}=c_\eta\alpha_t,\quad\lambda_{t+1}=c_\lambda\alpha_t,\quad\rho_{\mathbf{r}}^2=\frac{1}{d_2^2T},\quad\rho_{\mathbf{s}}^2=\frac{1}{d_1^2T},\\
&b=\frac{T^{1/3}}{(d_1+d_2)^{3/2}},\quad\bar{b}=\frac{T^{2/3}}{(d_1+d_2)^{3/4}}.
\end{aligned}
\tag{258}
$$

Thus, using (257), (258), and rearranging the terms, we get

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\|^2\right]\\
&\leq \frac{2}{\alpha_T}\mathcal{O}\left(V_T+\sum_{t=1}^{T}\alpha_t(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2)+\sum_{t=1}^{T}\alpha_t^2+\frac{H_{2,T}}{\alpha_T}+\sum_{t=1}^{T}(d_1+d_2)(\alpha_t^2\rho_{\mathbf{r}}^2+\alpha_t^2)\right)\\
&+\frac{2}{\alpha_T}\mathcal{O}\left(D_{\mathbf{y},T}+D_{\mathbf{x},T}+G_{\mathbf{y},T}+\frac{1}{\alpha_T}(G_{\mathbf{v},T}+G_{\mathbf{x},T})\right)\\
&+\frac{2}{\alpha_T}\mathcal{O}\left(\sum_{t=1}^{T}\left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2\alpha_t^2}{\bar{b}}+\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2\alpha_t}{\bar{b}}+\frac{\hat{\sigma}_{f_{\mathbf{y}}}^2\alpha_t^2}{b}+\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2\alpha_t}{\bar{b}}+\frac{\hat{\sigma}_{f_{\mathbf{x}}}^2\alpha_t^2}{b}\right)\right)\\
&+\frac{2}{\alpha_T}\mathcal{O}\left(\sum_{t=1}^{T}(\alpha_t+\frac{1}{\alpha_t})\left(\rho_{\mathbf{s}}^2+\rho_{\mathbf{r}}^2\right)+(1+\frac{1}{\alpha_T})T(d_2^2\rho_{\mathbf{r}}^2+d_1^2\rho_{\mathbf{s}}^2)\right)\\
&+\frac{2}{\alpha_T}\mathcal{O}\left(\theta_1^{\mathbf{y}}+\theta_1^{\mathbf{v}}+\hat{\sigma}^2\right)\\
&\leq \mathcal{O}\Big((d_1+d_2)^{3/4}T^{1/3}\left(V_T+D_{\mathbf{y},T}+D_{\mathbf{x},T}+G_{\mathbf{y},T}+\Delta_1+\hat{\sigma}^2\right)\\
&\qquad+(d_1+d_2)^{3/2}T^{2/3}\left(H_{2,T}+G_{\mathbf{v},T}+G_{\mathbf{x},T}\right)\Big),
\end{aligned} \tag{259}$$

where second inequality holds because we have

$$\begin{aligned}
&\sum_{t=1}^{T}\alpha_t^3=\sum_{t=1}^{T}\frac{1}{(d_1+d_2)^{9/4}(c+t)}\leq\sum_{t=1}^{T}\frac{1}{(d_1+d_2)^{9/4}(1+t)}\leq\frac{\log(T+1)}{(d_1+d_2)^{9/4}},\\
&\sum_{t=1}^{T}\alpha_t^2=\sum_{t=1}^{T}\frac{1}{(d_1+d_2)^{3/2}(c+t)^{2/3}}\leq\sum_{t=1}^{T}\frac{1}{(d_1+d_2)^{3/2}(1+t)^{2/3}}\leq\frac{T^{1/3}}{(d_1+d_2)^{3/2}},\\
&\sum_{t=1}^{T}\alpha_t=\sum_{t=0}^{T}\frac{1}{(d_1+d_2)^{3/4}(c+t)^{1/3}}\leq\sum_{t=1}^{T}\frac{1}{(d_1+d_2)^{3/4}(1+t)^{1/3}}\leq\frac{3T^{2/3}}{2(d_1+d_2)^{3/4}},\\
&\sum_{t=1}^{T}\frac{1}{\alpha_t}=\sum_{t=0}^{T}(d_1+d_2)^{3/4}(c+t)^{1/3}\leq\frac{3}{2}(d_1+d_2)^{3/4}T^{4/3}.
\end{aligned}$$

Then, note that, we have

$$\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\|^2\right]\\
&\leq\sum_{t=1}^{T}\mathbb{E}\left[\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\|^2\right]\\
&+\sum_{t=1}^{T}\mathbb{E}\left[\|\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)-\mathcal{P}_{\mathcal{X},\alpha_t}\left(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right)\|^2\right].
\end{aligned}$$

From non-expansiveness of the projection operator and Lemma D.4, we have

$$\begin{aligned}
&\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))) - \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\|^2 \\
\leq &\|\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)) - \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\|^2 \\
\leq &\frac{(\rho_{\mathbf{s}}d_1+\rho_{\mathbf{r}}d_2)^2\ell_{f,1}^2}{4} \\
\leq &\frac{(\rho_{\mathbf{s}}^2d_1^2+\rho_{\mathbf{r}}^2d_2^2)\ell_{f,1}^2}{2}.
\end{aligned}$$

This implies

$$\begin{aligned}
&\frac{1}{2}\sum_{t=1}^T \mathbb{E}\left[\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\|^2\right] \\
\leq &\sum_{t=1}^T \mathbb{E}\left[\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\|^2\right] + \frac{T(\rho_{\mathbf{s}}^2d_1^2+\rho_{\mathbf{r}}^2d_2^2)\ell_{f,1}^2}{2}.
\end{aligned}$$

Applying the upper bound in (259) yields

$$\begin{aligned}
&\frac{1}{2}\sum_{t=1}^T \mathbb{E}\left[\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\|^2\right] \\
\leq &\mathcal{O}\Big((d_1+d_2)^{3/4}T^{1/3}\left(V_T + D_{\mathbf{y},T} + D_{\mathbf{x},T} + G_{\mathbf{y},T} + \Delta_1 + \hat{\sigma}^2\right) \\
&\qquad +(d_1+d_2)^{3/2}T^{2/3}\left(H_{2,T} + G_{\mathbf{v},T} + G_{\mathbf{x},T}\right)\Big) \\
+ &\frac{T(\rho_{\mathbf{s}}^2d_1^2+\rho_{\mathbf{r}}^2d_2^2)\ell_{f,1}^2}{2}.
\end{aligned}$$

Thus, from $\rho_{\mathbf{r}}^2 = \frac{1}{d_2^2T}$ and $\rho_{\mathbf{s}}^2 = \frac{1}{d_1^2T}$, we get

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}\left[\|\mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t;\nabla f_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t)))\|^2\right] \\
\leq &\mathcal{O}\Big((d_1+d_2)^{3/4}T^{1/3}\left(V_T + D_{\mathbf{y},T} + D_{\mathbf{x},T} + G_{\mathbf{y},T} + \Delta_1 + \hat{\sigma}^2\right) \\
&\qquad +(d_1+d_2)^{3/2}T^{2/3}\left(H_{2,T} + G_{\mathbf{v},T} + G_{\mathbf{x},T}\right)\Big).
\end{aligned}$$

This completes the proof. □