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

Parvin Nazari

Amirkabir University of Technology
p.nazari17@gmail.com

Bojian Hou

University of Pennsylvania
bojianh@upenn.edu

Davoud Ataee Tarzanagh *

Samsung SDS Research America
d.tarzanagh@samsung.com

Li Shen

University of Pennsylvania
li.shen@pennmedicine.upenn.edu

George Michailidis

University of California, Los Angeles
gmichail@ucla.edu

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 [64] and formalized in mathematical optimization [9], BO finds applications in operations research, engineering, economics [16], and image processing [14]. Recently, BO has gained traction in machine learning, including hyperparameter optimization [22], meta-learning [18], reinforcement learning [65], and neural architecture search [51].

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

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

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

OBO [67] 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) [72], OBO involves iterative decision-making without prior knowledge of outcomes [67, 50, 8]. Let T be the total number of rounds. Define $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^{d_1}$ as the

*Corresponding author

OBO Method	Window Size in Regret (w)	System Iters.	Stochastic Regret	Const. Regret Min.	Only Func. Feedback	Local Regret Bound
OAGD [67]	$o(T)$	N.A. (Exact)	✗	✗	✗	$\frac{T}{w} + H_{1,T} + H_{2,T}$
SOBOW [50]	$o(T)$	$\mathcal{O}(\kappa_g \log \kappa_g)$	✗	✗	✗	$\frac{T}{w} + V_T + H_{2,T}$
SOBBO [8]	$o(T)$	$\mathcal{O}(\kappa_g \log \kappa_g)$	✓	✓	✗	$\frac{T}{w} \sigma^2 + V_T + H_{2,T}$
SOGD	1	1	✓	✓	✗	$T^{\frac{1}{3}}(\sigma^2 + \Delta_T) + T^{\frac{2}{3}}\Psi_T$
ZO-SOGD	1	1	✓	✓	✓	$(d_1 + d_2)^{\frac{3}{4}}T^{\frac{1}{4}}(\hat{\sigma}^2 + \hat{\Delta}_T) + (d_1 + d_2)^{\frac{3}{4}}T^{\frac{3}{4}}\hat{\Psi}_T$

Table 1: Comparison of OBO algorithms based on regret window w , solver iterations, stochastic/constrained regrets, feedback type, and local bounds. κ_g denotes the condition number of the inner objective g_t . V_T , $H_{p,T}$, Δ_T , Ψ_T , $\hat{\Delta}_T$, $\hat{\Psi}_T$, σ , and $\hat{\sigma}$ are defined in (11), (14), (30), (10), and (28), respectively.

decision variable and $f_t : \mathcal{X} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}$ as the outer function. Similarly, define $\mathbf{y}_t \in \mathbb{R}^{d_2}$ and $g_t : \mathcal{X} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}$ for the inner problem, where $\mathbf{y}_t^*(\mathbf{x}) = \operatorname{argmin}_{\mathbf{y} \in \mathbb{R}^{d_2}} g_t(\mathbf{x}, \mathbf{y})$. 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 [64], 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 . This framework includes online and adversarial variants of (BO), such as online actor-critic algorithms [71], online meta-learning [19], and online hyperparameter optimization [50]. The inner and outer functions may be time-varying, adversarial, unavailable *a priori*, and require *nonstationary* optimization.

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

- **Stochastic regret minimization without window-smoothing.** Existing OBO methods [67, 50, 40, 8] 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 2) and proving that both first-order and ZO methods achieve sublinear *stochastic bilevel regret without window-smoothing* ($w = 1$); see Theorems 2.6 and 3.2 and Table 1.

- **OBO with function value oracle feedback.** In large-scale and black-box settings [11, 58], 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 [21, 27]. We propose Algorithm 2, which estimates Hessians, Jacobians, and gradients using function value oracles, achieving sublinear local regret (Theorem 3.2).

- **OBO with one subproblem solver iteration.** A major challenge in BO is solving implicit systems to approximate the hypergradient [43, 12]. While efficient offline BO methods exist [43, 15], extending them to OBO is difficult due to time-varying objectives. SOBOW [50] 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 Stochastic OBO with Access to First- and Inner Second-Order Oracles

Notation. $\mathbb{R}^d$ is the d -dimensional real space; $\mathbb{R}_+^d$ and $\mathbb{R}_{++}^d$ denote its nonnegative and positive orthants. Bold lowercase letters (e.g., $\mathbf{x}, \mathbf{y}$) represent vectors, $\langle \mathbf{x}, \mathbf{y} \rangle$ is the inner product, and $\|\cdot\|$ is the Euclidean norm. $\nabla_{\mathbf{x}}$ denotes the gradient, and $\nabla_{\mathbf{x}\mathbf{y}}^2 = \nabla_{\mathbf{x}} \nabla_{\mathbf{y}}$. A function is L -smooth if its gradient is L -Lipschitz. The projection onto a convex set $\mathcal{X}$ is $\Pi_{\mathcal{X}}(\mathbf{z}) = \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} \frac{1}{2} \|\mathbf{x} - \mathbf{z}\|^2$. We use $[T]$ for $\{1, \dots, T\}$, $\mathbb{E}[\cdot]$ for expectation, and $\mathcal{O}(\cdot)$ to hide problem-independent constants.

Stochastic OBO Setting. Let T be the total rounds [67]. Define $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^{d_1}$ as the decision variable and $f_t : \mathcal{X} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}$ as the outer objective. The inner decision variable and objective are $\mathbf{y}_t \in \mathbb{R}^{d_2}$ and $g_t : \mathcal{X} \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}$, where the optimal inner decision is:

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

Further, we have

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

Here, $(\mathcal{D}_{f,t}, \mathcal{D}_{g,t})$ denote data distributions at time t . Our setting is stochastic, with only noisy evaluations of functions, gradients, and Hessians. Unlike OSO [72], where true losses are revealed, in OBO the outer function $f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))$ is inaccessible for updating $\mathbf{x}_t$ and is generally non-convex in $\mathbf{x}$, making standard regret notions from online convex optimization [33] unsuitable.

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

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

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

The local regret (2) compares the leader's decision $\mathbf{x}_t$ to the stationary points $\mathbf{x}_t^*$ satisfying $\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t^*; \nabla f_t(\mathbf{x}_t^*, \mathbf{y}_t^*(\mathbf{x}_t^*))) = 0$. 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). [67, 50] 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.

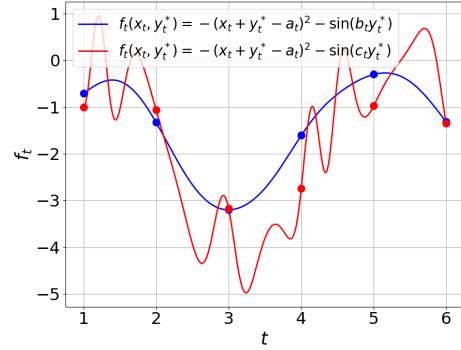


Figure 1: Smoothly and rapidly changing f_t in OBO with $g_t(x_t, y_t) = (y_t - \cos(x_t))^2$, $a_t = 1 + 0.5 \sin(t)$, $b_t = 1 + \sin(0.5t)$, and $c_t = 10b_t$.

Online Gradient Descent (OGD). One of the most widely used algorithms for online (single-level) optimization is OGD [72]. The procedure for OGD is as follows: For each $t \in [T]$, the algorithm selects $\mathbf{x}_t \in \mathcal{X}$, observes the function $f_t : \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathbb{R}$, and updates according to

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

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.

To adapt OGD to OBO, [67, 50, 8] 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

$$\nabla f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) = \nabla_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) + \nabla_{\mathbf{x}\mathbf{y}}^2 g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) \mathbf{v}_t^*(\mathbf{x}), \quad (3)$$

where $\mathbf{v}_t^*(\mathbf{x}) \in \mathbb{R}^{d_2}$ is the solution to the following linear system:

$$\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) \mathbf{v}_t^*(\mathbf{x}) + \nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) = 0. \quad (4)$$

As the exact $\mathbf{y}_t^*(\mathbf{x})$ is not available, we estimate the hypergradient of f_t at $(\mathbf{x}, \mathbf{y})$ and introduce an auxiliary variable $\mathbf{v} := \mathbf{v}(\mathbf{x}, \mathbf{y})$ to effectively decouple the nonlinear structure in $\nabla f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))$, i.e.

$$\tilde{\nabla} f_t(\mathbf{x}, \mathbf{y}) := \nabla_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}) + \nabla_{\mathbf{x}\mathbf{y}}^2 g_t(\mathbf{x}, \mathbf{y}) \mathbf{v}_t, \quad (5a)$$

where $\mathbf{v}_t$ serves as an inexact solution to the linear system

$$\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}, \mathbf{y}) \mathbf{v}_t + \nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}) = 0. \quad (5b)$$

Algorithm 1 SOGD

Require: $(\mathbf{x}_1, \mathbf{y}_1, \mathbf{v}_1) \in \mathcal{X} \times \mathbb{R}^{d_2} \times \mathcal{Z}_p$; $p \in \mathbb{R}_{++}$; $T \in \mathbb{N}$; 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 $\mathbf{d}_t^y$, $\mathbf{d}_t^v$, and $\mathbf{d}_t^x$:

$$\mathbf{d}_t^{yy}(\mathbf{z}_t; \bar{\mathcal{B}}_t) = \nabla_{\mathbf{y}} g_t(\mathbf{z}_t; \bar{\mathcal{B}}_t), \quad (9a)$$

$$\mathbf{d}_t^y = \mathbf{d}_t^{yy}(\mathbf{z}_t; \bar{\mathcal{B}}_t) + (1 - \gamma_t)(\mathbf{d}_{t-1}^y - \mathbf{d}_t^{yy}(\mathbf{z}_{t-1}; \bar{\mathcal{B}}_t)),$$

$$\mathbf{d}_t^{vv}(\mathbf{z}_t; \mathcal{B}_t) = \nabla_{\mathbf{y}} f_t(\mathbf{z}_t; \mathcal{B}_t) + \nabla_{\mathbf{y}}^2 g_t(\mathbf{z}_t; \bar{\mathcal{B}}_t) \mathbf{v}_t, \quad (9b)$$

$$\mathbf{d}_t^v = \mathbf{d}_t^{vv}(\mathbf{z}_t; \mathcal{B}_t) + (1 - \lambda_t)(\mathbf{d}_{t-1}^v - \mathbf{d}_t^{vv}(\mathbf{z}_{t-1}; \mathcal{B}_t)),$$

$$\mathbf{d}_t^{xx}(\mathbf{z}_t; \mathcal{B}_t) = \nabla_{\mathbf{x}} f_t(\mathbf{z}_t; \mathcal{B}_t) + \nabla_{\mathbf{xy}}^2 g_t(\mathbf{z}_t; \bar{\mathcal{B}}_t) \mathbf{v}_t, \quad (9c)$$

$$\mathbf{d}_t^x = \mathbf{d}_t^{xx}(\mathbf{z}_t; \mathcal{B}_t) + (1 - \eta_t)(\mathbf{d}_{t-1}^x - \mathbf{d}_t^{xx}(\mathbf{z}_{t-1}; \mathcal{B}_t)).$$

S2. Update inner, system, and outer solutions:

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

An accurate solution of (5b) is crucial for tight regret bounds. [67] assumes an exact solution, which is restrictive in large-scale settings. To address this, [50] proposed an efficient OBO algorithm with window averaging, using CG methods to solve (5b), which is equivalent to:

$$\min_{\mathbf{v}_t \in \mathbb{R}^{d_2}} \frac{1}{2} \left\| \nabla_{\mathbf{y}}^2 g_t(\mathbf{x}, \mathbf{y}) \mathbf{v}_t + \nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}) \right\|^2. \quad (6)$$

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 2.1. *Let $w = t$, $W = 1/\eta$ and $\nu = 1 - \eta$ for $\eta \in (0, 1)$ in the window-smoothed gradient $\nabla F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) = \frac{1}{W} \sum_{i=0}^{w-1} \nu^i \nabla f_{t-i}(\mathbf{x}_{t-i}, \mathbf{y}_{t-i}; \mathcal{B}_{t-i})$, where $\mathcal{B}_t := \{\xi_{t,1}, \dots, \xi_{t,b}\}$ is drawn i.i.d. from $\mathcal{D}_{f,t}$. Then, $\nabla F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) = \sum_{j=1}^t \eta(1 - \eta)^{t-j} \nabla f_j(\mathbf{x}_j, \mathbf{y}_j; \mathcal{B}_j)$, and we have $\nabla F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) = \mathbf{d}_t^x$ with $\mathbf{d}_t^x = \eta \nabla f_t(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) + (1 - \eta) \mathbf{d}_{t-1}^x$, and $\mathbf{d}_1^x = \frac{1}{W} \nabla f_1(\mathbf{x}_1, \mathbf{y}_1; \mathcal{B}_1)$ for all $t \geq 2$.*

Proof is given in Appendix C.1. As shown in Lemma 2.1, 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 ($w = o(T)$) [67, 50, 8]. To address this, we propose the following search direction:

$$\mathbf{d}_t^x = \eta \nabla f_t(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) + (1 - \eta) \mathbf{d}_{t-1}^x + (1 - \eta)(\nabla f_t(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) - \nabla f_t(\mathbf{x}_{t-1}, \mathbf{y}_{t-1}; \mathcal{B}_t)). \quad (7)$$

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 formulation of (5b) in (6) motivates single-loop methods such as [15]. Building on this, we propose Simultaneous Online Gradient Descent (SOGD) for constrained OBO, presented in Algorithm 1. At each step, SOGD jointly updates the follower variable $\mathbf{y}_t$, auxiliary variable $\mathbf{v}_t$, and leader variable $\mathbf{x}_t$ using batches $\mathcal{B}_t = \{\xi_{t,1}, \dots, \xi_{t,b}\}$ and $\bar{\mathcal{B}}_t := \{\zeta_{t,1}, \dots, \zeta_{t,\bar{b}}\}$ sampled i.i.d. from $\mathcal{D}_{f,t}$ and $\mathcal{D}_{g,t}$. Step S1. only requires computing Hessian-vector products, avoiding explicit computation of $\nabla_{\mathbf{y}}^2 g_t$ or $\nabla_{\mathbf{xy}}^2 g_t$. Step S2. uses the projection:

$$\Pi_{\mathcal{Z}_p}(\mathbf{v}) = \operatorname{argmin}_{\mathbf{z} \in \mathcal{Z}_p} \frac{1}{2} \|\mathbf{v} - \mathbf{z}\|^2 = \min \left\{ 1, \frac{p}{\|\mathbf{v}\|} \right\} \mathbf{v}, \quad \text{where} \\ \mathcal{Z}_p := \{\mathbf{v} \in \mathbb{R}^{d_2} \mid \|\mathbf{v}\| \leq p\}. \quad (8)$$

Unlike OAGD [67] with alternating loops, and SOBOW [50] using CG, SOGD performs a single OGD step for all variables.

Assumption 2.2. $g_t(\mathbf{x}, \mathbf{y})$ is twice continuously differentiable and μ_g -strongly convex in $\mathbf{y}$ for all $\mathbf{x} \in \mathcal{X}, t \in [T]$.

Assumption 2.3. 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. $\|f_t(\mathbf{z}; \xi_t) - f_t(\mathbf{z}'; \xi_t)\| \leq \ell_{f,0} \|\mathbf{z} - \mathbf{z}'\|$;
- B2. $\exists \ell_{f,1} \in \mathbb{R}_+$ s.t. $\|\nabla f_t(\mathbf{z}; \xi_t) - \nabla f_t(\mathbf{z}'; \xi_t)\| \leq \ell_{f,1} \|\mathbf{z} - \mathbf{z}'\|$;
- B3. $\exists \ell_{g,1} \in \mathbb{R}_+$ s.t. $\|\nabla g_t(\mathbf{z}; \zeta_t) - \nabla g_t(\mathbf{z}'; \zeta_t)\| \leq \ell_{g,1} \|\mathbf{z} - \mathbf{z}'\|$;
- B4. $\exists \ell_{g,2} \in \mathbb{R}_+$ s.t. $\|\nabla^2 g_t(\mathbf{z}; \zeta_t) - \nabla^2 g_t(\mathbf{z}'; \zeta_t)\| \leq \ell_{g,2} \|\mathbf{z} - \mathbf{z}'\|$.

Assumption 2.4. For any $t \in [T]$, $|f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))| \leq M$ for some $M \in \mathbb{R}_{++}$ and any $\mathbf{x} \in \mathcal{X}$.

Assumption 2.5. There exist constants $\sigma_{g_y}, \sigma_{g_{yy}}, \sigma_{g_{xy}}, \sigma_{f_y}, \sigma_{f_x}$ such that, for all $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$:

- C1. $\mathbb{E} \|\nabla_y g_t(\mathbf{z}; \zeta_t) - \nabla_y g_t(\mathbf{z})\|^2 \leq \sigma_{g_y}^2$;
- C2. $\mathbb{E} \|\nabla_y^2 g_t(\mathbf{z}; \zeta_t) - \nabla_y^2 g_t(\mathbf{z})\|^2 \leq \sigma_{g_{yy}}^2$;
- C3. $\mathbb{E} \|\nabla_{xy}^2 g_t(\mathbf{z}; \zeta_t) - \nabla_{xy}^2 g_t(\mathbf{z})\|^2 \leq \sigma_{g_{xy}}^2$;
- C4. $\mathbb{E} \|\nabla_y f_t(\mathbf{z}; \xi_t) - \nabla_y f_t(\mathbf{z})\|^2 \leq \sigma_{f_y}^2$;
- C5. $\mathbb{E} \|\nabla_x f_t(\mathbf{z}; \xi_t) - \nabla_x f_t(\mathbf{z})\|^2 \leq \sigma_{f_x}^2$.

Throughout this paper, we define

$$\sigma^2 := \sigma_{g_y}^2 + \sigma_{g_{yy}}^2 + \sigma_{f_y}^2 + \sigma_{g_{xy}}^2 + \sigma_{f_x}^2. \quad (10)$$

Assumptions 2.2 and 2.3 are standard in BO [12, 43] and OBO [67], and hold for many bilevel ML problems [22]. Assumption 2.4 is typical in non-convex OSO [36, 50], while Assumption 2.5 assumes unbiased stochastic gradient, Hessian, and Jacobian estimators with bounded variance [12].

Achieving sublinear dynamic regret is generally infeasible under arbitrary time variations [7]. Prior analyses [67, 50] bound regret by enforcing regularity on the comparator sequence. To attain sublinear regret, [67] introduces the following regularity metrics for bilevel sequences:

$$H_{p,T} := \sum_{t=2}^T \sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^p, \quad V_T := \sum_{t=2}^T \sup_{\mathbf{x} \in \mathcal{X}} |f_{t-1}(\mathbf{x}, \mathbf{y}_{t-1}^*(\mathbf{x})) - f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))|. \quad (11)$$

Path-length $H_{p,T}$ measures changes in the follower's costs, while V_T captures the leader's objective smoothness. We use path-length for the follower and function variation for the leader due to the follower's strong convexity (Assumption 2.2) versus the leader's nonconvexity. Another regularity is the sequential gradient difference of the outer objective:

$$D_{\mathbf{x},T} := \sum_{t=2}^T \sup_{\mathbf{x}, \mathbf{y}} \|\nabla_{\mathbf{x}} f_{t-1}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y})\|^2, \quad (12a)$$

$$D_{\mathbf{y},T} := \sum_{t=2}^T \sup_{\mathbf{x}, \mathbf{y}} \|\nabla_{\mathbf{y}} f_{t-1}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y})\|^2. \quad (12b)$$

As in [41, 31], $D_{\mathbf{x},T}$ 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 define deviations in the gradient, Hessian, and Jacobian of the inner objective as:

$$G_{\mathbf{y},T} := \sum_{t=2}^T \|\nabla_{\mathbf{y}} g_{t-1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2, \quad G_{\mathbf{yy},T} := \sum_{t=2}^T \|\nabla_{\mathbf{y}}^2 g_{t-1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2, \\ G_{\mathbf{xy},T} := \sum_{t=2}^T \|\nabla_{\mathbf{xy}}^2 g_{t-1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2. \quad (13)$$

We introduce the following notations for simplicity:

$$\Delta_T := E_1 + V_T, \quad \Psi_T := H_{2,T} + G_T + D_T, \quad (14)$$

where $(V_T, H_{p,T})$ are defined in (11), and

$$E_1 := \|\mathbf{y}_1 - \mathbf{y}_1^*(\mathbf{x}_1)\|^2 + \|\mathbf{v}_1 - \mathbf{v}_1^*(\mathbf{x}_1)\|^2, \quad G_T := G_{\mathbf{y},T} + G_{\mathbf{yy},T} + G_{\mathbf{xy},T}, \\ D_T := D_{\mathbf{x},T} + D_{\mathbf{y},T}. \quad (15)$$

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

Theorem 2.6. Let $\{(f_t, g_t)\}_{t=1}^T$ be the sequence of functions presented to Algorithm 1, satisfying Assumptions 2.2-2.5. For all $t \in [T]$, let

$$\alpha_t = \frac{1}{(c+t)^{1/3}}, \quad \beta_t = c_\beta \alpha_t, \quad \delta_t = c_\delta \alpha_t, \quad b = \bar{b} = 1, \\ \gamma_{t+1} = c_\gamma \alpha_t^2, \quad \eta_{t+1} = c_\eta \alpha_t^2, \quad \lambda_{t+1} = c_\lambda \alpha_t^2. \quad (16)$$

Here, c , c_β , c_δ , c_γ , c_η , and c_λ are specified in (107). Algorithm 1 guarantees:

$$\text{BL-Reg}_T \leq \mathcal{O}\left(T^{\frac{1}{3}}(\sigma^2 + \Delta_T) + T^{\frac{2}{3}}\Psi_T\right), \quad (17)$$

where σ and (Δ_T, Ψ_T) are defined in (10) and (14).

Remark 2.7 (Stochastic Regret Guarantee for OBO and OSO with $w = 1$). Theorem 2.6 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 , Ψ_T , and σ . Specifically, if $\Delta_T = o(T^{2/3})$, $\Psi_T = o(T^{1/3})$, and $\sigma = o(T^{1/3})$, then the dynamic regret grows sublinearly, i.e., $\text{BL-Reg}_T = o(T)$; see Appendix B.2 for further examples and discussion. This result also yields a sharper $T^{-2/3}\sigma^2$ regret—improving over the $T^{-1/2}\sigma^2$ bound for stochastic OBO [8]—and removes the need for window-smoothing [8, 67, 50, 40]. For OSO, this result surpasses the $T^{-1/2}\sigma^2$ rate in [31].

3 Stochastic OBO with Zeroth-Order Oracles

Black-box optimization arises when gradients are unavailable [11]. We study ZO-OBO methods with limited access to leader and follower objectives. Let $\mathbf{s} \in \mathbb{R}^{d_1}$ and $\mathbf{r} \in \mathbb{R}^{d_2}$ be vectors uniformly sampled from unit balls B_1 and B_2 . Given smoothing parameters $\boldsymbol{\rho} = (\rho_{\mathbf{s}}, \rho_{\mathbf{r}})$, we define Gaussian-smoothed objectives using [59]:

$$f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) = \mathbb{E}_{(\mathbf{s}, \mathbf{r}, \xi_t)} [f_t(\mathbf{x} + \rho_{\mathbf{s}}\mathbf{s}, \hat{\mathbf{y}}_t^*(\mathbf{x}) + \rho_{\mathbf{r}}\mathbf{r}; \xi_t)], \quad \text{where} \quad (18)$$

$$\hat{\mathbf{y}}_t^*(\mathbf{x}) \in \arg \min_{\mathbf{y} \in \mathbb{R}^{d_2}} \{g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) := \mathbb{E}_{(\mathbf{s}, \mathbf{r}, \zeta_t)} [g_t(\mathbf{x} + \rho_{\mathbf{s}}\mathbf{s}, \mathbf{y} + \rho_{\mathbf{r}}\mathbf{r}; \zeta_t)]\}. \quad (19)$$

To solve stochastic OBO with (18), we need to obtain the hyper-gradient of $f_{t,\boldsymbol{\rho}}$ in (18) at $(\mathbf{x}, \mathbf{y})$ as

$$\begin{aligned} \nabla f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) &:= \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) + \nabla_{\mathbf{xy}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) \hat{\mathbf{v}}_t^*(\mathbf{x}), \quad \text{where} \\ \hat{\mathbf{v}}_t^*(\mathbf{x}) &\text{ is the solution to } \nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) \hat{\mathbf{v}}_t^*(\mathbf{x}) + \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) = 0. \end{aligned} \quad (20)$$

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 $(\mathbf{x}, \mathbf{y})$, we introduce an auxiliary variable $\mathbf{v} = \mathbf{v}(\mathbf{x}, \mathbf{y})$ and define:

$$\tilde{\nabla} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) := \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) + \nabla_{\mathbf{xy}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) \mathbf{v}, \quad \text{where} \quad (21a)$$

$$\mathbf{v} \text{ is the solution to } \nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) \mathbf{v} + \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) = 0. \quad (21b)$$

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

$$\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{y}}(\mathbf{x}, \mathbf{y}) = \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}), \quad (22a)$$

$$\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{v}}(\mathbf{x}, \mathbf{y}, \mathbf{v}) = \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) + \nabla_{\mathbf{y}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) \mathbf{v}, \quad (22b)$$

$$\mathbf{d}_{t,\boldsymbol{\rho}}^{\mathbf{x}}(\mathbf{x}, \mathbf{y}, \mathbf{v}) = \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) + \nabla_{\mathbf{xy}}^2 g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}) \mathbf{v}. \quad (22c)$$

Next, we approximate these directions using stochastic zeroth-order oracles (SZO), which produce the quantities $\hat{\nabla}_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}; \xi_t)$, $\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}; \zeta_t)$, $\hat{\nabla}_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}; \xi_t)$, and $\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}; \zeta_t)$. These are unbiased estimators of the true gradients $\nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})$, $\nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})$, $\nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})$, and $\nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})$, respectively, as shown in [20], such that the following assumption holds:

$$\begin{aligned} \mathbb{E}_{(\mathbf{r}, \xi_t)} [\hat{\nabla}_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}; \xi_t)] &= \nabla_{\mathbf{y}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}), & \mathbb{E}_{(\mathbf{s}, \xi_t)} [\hat{\nabla}_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}; \xi_t)] &= \nabla_{\mathbf{x}} f_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}), \\ \mathbb{E}_{(\mathbf{r}, \zeta_t)} [\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}; \zeta_t)] &= \nabla_{\mathbf{y}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}), & \mathbb{E}_{(\mathbf{s}, \zeta_t)} [\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}; \zeta_t)] &= \nabla_{\mathbf{x}} g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y}). \end{aligned} \quad (23)$$

Specifically, following [62], we estimate the gradient of a function $h : \mathbb{R}^d \rightarrow \mathbb{R}$, querying at $\mathbf{x} - \lambda \mathbf{s}$ and $\mathbf{x} + \lambda \mathbf{s}$, yielding an estimator $(d/2\lambda)(h(\mathbf{x} + \lambda \mathbf{s}) - h(\mathbf{x} - \lambda \mathbf{s})) \mathbf{s}$. Using this strategy, the finite-difference estimation of $\nabla g_{t,\boldsymbol{\rho}}(\mathbf{x}, \mathbf{y})$, denoted by $\hat{\nabla} g_t(\mathbf{x}, \mathbf{y})$, is constructed for given smoothing

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 (24)–(26), get:

$$\mathbf{d}_t^{\mathbf{y}}(\mathbf{z}_t; \bar{\mathcal{B}}_t) = \hat{\nabla}_{\mathbf{y}} g_t(\mathbf{z}_t; \bar{\mathcal{B}}_t), \quad (27a)$$

$$\begin{aligned} \hat{\mathbf{d}}_t^{\mathbf{y}} &= \mathbf{d}_t^{\mathbf{y}}(\mathbf{z}_t; \bar{\mathcal{B}}_t) + (1 - \gamma_t)(\hat{\mathbf{d}}_{t-1}^{\mathbf{y}} - \mathbf{d}_t^{\mathbf{y}}(\mathbf{z}_{t-1}; \bar{\mathcal{B}}_t)), \\ \mathbf{d}_t^{\mathbf{vv}}(\mathbf{z}_t; \mathcal{B}_t) &= \hat{\nabla}_{\mathbf{y}} f_t(\mathbf{z}_t; \mathcal{B}_t) + \hat{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{z}_t; \bar{\mathcal{B}}_t), \end{aligned} \quad (27b)$$

$$\begin{aligned} \hat{\mathbf{d}}_t^{\mathbf{v}} &= \mathbf{d}_t^{\mathbf{vv}}(\mathbf{z}_t; \mathcal{B}_t) + (1 - \lambda_t)(\hat{\mathbf{d}}_{t-1}^{\mathbf{v}} - \mathbf{d}_t^{\mathbf{vv}}(\mathbf{z}_{t-1}; \mathcal{B}_t)), \\ \mathbf{d}_t^{\mathbf{xy}}(\mathbf{z}_t; \mathcal{B}_t) &= \hat{\nabla}_{\mathbf{x}} f_t(\mathbf{z}_t; \mathcal{B}_t) + \hat{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{z}_t; \bar{\mathcal{B}}_t), \\ \hat{\mathbf{d}}_t^{\mathbf{x}} &= \mathbf{d}_t^{\mathbf{xy}}(\mathbf{z}_t; \mathcal{B}_t) + (1 - \eta_t)(\hat{\mathbf{d}}_{t-1}^{\mathbf{x}} - \mathbf{d}_t^{\mathbf{xy}}(\mathbf{z}_{t-1}; \mathcal{B}_t)), \end{aligned} \quad (27c)$$

S2. Update inner, system, and outer solutions:

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

parameters $\rho = (\rho_{\mathbf{s}}, \rho_{\mathbf{r}})$, and a set $\bar{\mathcal{B}}_t = \{\zeta_{t,1}, \dots, \zeta_{t,\bar{b}}\}$ drawn i.i.d. from $\mathcal{D}_{g,t}$, as follows:

$$\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathcal{B}}_t) := \frac{d_2}{2\bar{b}\rho_{\mathbf{r}}} \sum_{i=1}^{\bar{b}} (g_t(\mathbf{x}, \mathbf{y} + \rho_{\mathbf{r}} \mathbf{r}_i; \zeta_{t,i}) - g_t(\mathbf{x}, \mathbf{y} - \rho_{\mathbf{r}} \mathbf{r}_i; \zeta_{t,i})) \mathbf{r}_i, \quad (24a)$$

$$\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathcal{B}}_t) := \frac{d_1}{2\bar{b}\rho_{\mathbf{s}}} \sum_{i=1}^{\bar{b}} (g_t(\mathbf{x} + \rho_{\mathbf{s}} \mathbf{s}_i, \mathbf{y}; \zeta_{t,i}) - g_t(\mathbf{x} - \rho_{\mathbf{s}} \mathbf{s}_i, \mathbf{y}; \zeta_{t,i})) \mathbf{s}_i. \quad (24b)$$

Similarly, we estimate $\nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}, \mathbf{y}; \mathcal{B}_t)$ and $\nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}, \mathbf{y}; \mathcal{B}_t)$, respectively, using a batch $\mathcal{B}_t = \{\xi_{t,1}, \dots, \xi_{t,b}\}$ drawn i.i.d. from $\mathcal{D}_{f,t}$, by

$$\hat{\nabla}_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}; \mathcal{B}_t) := \frac{d_2}{2b\rho_{\mathbf{r}}} \sum_{i=1}^b (f_t(\mathbf{x}, \mathbf{y} + \rho_{\mathbf{r}} \mathbf{r}_i; \xi_{t,i}) - f_t(\mathbf{x}, \mathbf{y} - \rho_{\mathbf{r}} \mathbf{r}_i; \xi_{t,i})) \mathbf{r}_i, \quad (25a)$$

$$\hat{\nabla}_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}; \mathcal{B}_t) := \frac{d_1}{2b\rho_{\mathbf{s}}} \sum_{i=1}^b (f_t(\mathbf{x} + \rho_{\mathbf{s}} \mathbf{s}_i, \mathbf{y}; \xi_{t,i}) - f_t(\mathbf{x} - \rho_{\mathbf{s}} \mathbf{s}_i, \mathbf{y}; \xi_{t,i})) \mathbf{s}_i. \quad (25b)$$

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

$$\hat{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}, \mathbf{y}; \bar{\mathcal{B}}_t) := \frac{1}{2\bar{b}\rho_{\mathbf{v}}} \sum_{i=1}^{\bar{b}} (\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y} + \rho_{\mathbf{v}} \mathbf{v}; \zeta_{t,i}) - \hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y} - \rho_{\mathbf{v}} \mathbf{v}; \zeta_{t,i})), \quad (26a)$$

$$\hat{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}, \mathbf{y}; \bar{\mathcal{B}}_t) := \frac{1}{2\bar{b}\rho_{\mathbf{v}}} \sum_{i=1}^{\bar{b}} (\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y} + \rho_{\mathbf{v}} \mathbf{v}; \zeta_{t,i}) - \hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y} - \rho_{\mathbf{v}} \mathbf{v}; \zeta_{t,i})). \quad (26b)$$

Using (24)–(26), the first-order terms in (9) are approximated by $\hat{\mathbf{d}}_t^{\mathbf{y}}$, $\hat{\mathbf{d}}_t^{\mathbf{v}}$, and $\hat{\mathbf{d}}_t^{\mathbf{x}}$ in (27). The approximations in (26a) and (26b) introduce errors in the hypergradient, which must be controlled. (26) depends on the dimension of $\mathbf{y}$, as in ZO optimization [59, 62]. The projection $\Pi_{\mathcal{Z}_p}$ in (8) bounds $\mathbf{v}$, controlling variance in $\mathbf{v}$ and $\mathbf{x}$ updates for convergence.

Assumption 3.1. 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 $\mathbf{z} = [\mathbf{x}, \mathbf{y}]$:

$$\begin{aligned} \text{D1. } \mathbb{E} \|\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{z}; \zeta_t) - \nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{z})\|^2 &\leq \hat{\sigma}_{g_{\mathbf{y}}}^2, & \text{D3. } \mathbb{E} \|\hat{\nabla}_{\mathbf{y}} f_t(\mathbf{z}; \xi_t) - \nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{z})\|^2 &\leq \hat{\sigma}_{f_{\mathbf{y}}}^2, \\ \text{D2. } \mathbb{E} \|\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{z}; \zeta_t) - \nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{z})\|^2 &\leq \hat{\sigma}_{g_{\mathbf{x}}}^2, & \text{D4. } \mathbb{E} \|\hat{\nabla}_{\mathbf{x}} f_t(\mathbf{z}; \xi_t) - \nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{z})\|^2 &\leq \hat{\sigma}_{f_{\mathbf{x}}}^2. \end{aligned}$$

Assumption 3.1 is analogous to the upper bound on the variance of stochastic partial gradients discussed in [54, 68]. We simplify the notation by introducing the following shorthand.

$$\hat{\sigma}^2 := \hat{\sigma}_{g_y}^2 + \hat{\sigma}_{g_x}^2 + \hat{\sigma}_{f_y}^2 + \hat{\sigma}_{f_x}^2. \quad (28)$$

Next, we establish a regret bound for ZO-SOGD. Similar to the previous results, we introduce regularity conditions for the smoothed functions defined in (18) and (19).

Inner Gradient Variations: In ZO setting, we use a set of gradient variations at the perturbed point as follows:

$$G_{\mathbf{v},T} := \sum_{t=2}^T (\chi_{1t} + \chi_{2t}), \quad G_{\mathbf{x},T} := \sum_{t=2}^T (\chi_{3t} + \chi_{4t}), \quad (29)$$

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

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

Further, for simplicity of notation, we define

$$\hat{\Delta}_T := E_1 + V_T + D_T + G_{\mathbf{y},T}, \quad \hat{\Psi}_T := H_{2,T} + G_{\mathbf{v},T} + G_{\mathbf{x},T}, \quad (30)$$

where $(V_T, H_{p,T})$ and (E_1, D_T) are defined in (11), and (15), respectively. Moreover, $G_{\mathbf{y},T}$ and $(G_{\mathbf{v},T}, G_{\mathbf{x},T})$ are defined in (13) and (29), respectively.

Theorem 3.2. Let $\{(f_t, g_t)\}_{t=1}^T$ be the sequence of functions presented to Algorithm 2, satisfying Assumptions 2.2-2.4 and 3.1. For all $t \in [T]$, let

$$\begin{aligned} \alpha_t &= \frac{1}{(d_1 + d_2)^{3/4}(c+t)^{1/3}}, & \beta_t &= c_\beta \alpha_t, & \delta_t &= c_\delta \alpha_t, & \gamma_{t+1} &= c_\gamma \alpha_t, \\ \eta_{t+1} &= c_\eta \alpha_t, & \lambda_{t+1} &= c_\lambda \alpha_t, & \rho_{\mathbf{v}}^2 &= c_{\mathbf{v}} \alpha_t, & \rho_{\mathbf{r}}^2 &= \frac{1}{d_2^2 T}, & \rho_{\mathbf{s}}^2 &= \frac{1}{d_1^2 T}, \\ b &= \frac{T^{1/3}}{(d_1 + d_2)^{3/2}}, & \bar{b} &= \frac{T^{2/3}}{(d_1 + d_2)^{3/4}}, \end{aligned} \quad (31)$$

where $c, c_\beta, c_\delta, c_\gamma, c_\eta, c_{\mathbf{v}}$, and c_λ are specified in (226). Let $p = \ell_{f,0}/\mu_g$ for the set $\mathcal{Z}_p$ defined in (8). Then, Algorithm 2 guarantees:

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

where $\hat{\sigma}^2$ and $(\hat{\Delta}_T, \hat{\Psi}_T)$ are defined in (28) and (30).

Theorem 3.2 bounds the regret of Algorithm 2 without window-smoothing, based on the regularities in (30). We note that the average dynamic regret $\text{BL-Reg}_T/T \leq \mathcal{O}((d_1 + d_2)^{3/4} T^{-2/3} (\hat{\sigma}^2 + \hat{\Delta}_T) + (d_1 + d_2)^{3/2} T^{-1/3} \hat{\Psi}_T)$ remains sublinear under suitable conditions on $\hat{\Delta}_T, \hat{\Psi}_T$, and $\hat{\sigma}$.

Remark 3.3 (Regret Guarantee for Zeroth Order OBO). Theorem 3.2 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 3.2 aligns with optimal results for simpler offline min-max problems [39]. 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 3.4 (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 [60], which is $\mathcal{O}(T^{-1/2} \sigma^2 \sqrt{d})$. [60] proposed a zeroth-order stochastic gradient descent algorithm for unconstrained, non-convex, time-varying objective functions, achieving a regret bound of $\mathcal{O}(T^{-1/2} \sigma^2 \sqrt{d W_T})$ using a two-point gradient estimator, where W_T bounds the nonstationarity. Additionally, [29] showed that the local regret for standard online stochastic gradient descent with the standard two-point gradient estimator [1] is $\mathcal{O}(T^{-1/2} d \sqrt{V_T})$.

4 Experimental Results

In this section, we present experimental results for two applications: online black-box attacks on deep neural networks and parametric loss tuning for imbalanced data. Code is available at [\[link\]](#). Additional experiments and details on hyperparameter tuning are provided in Appendix E.

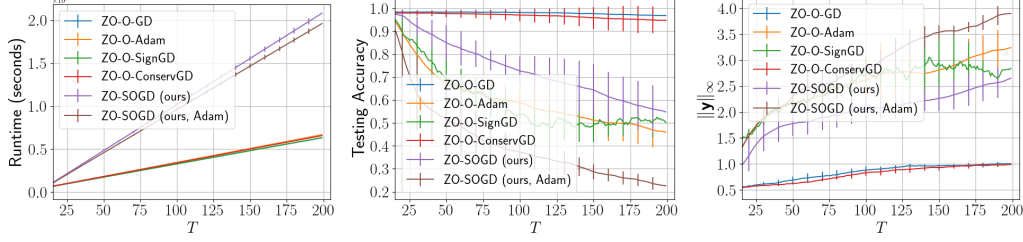


Figure 2: Performance comparison (mean $\pm$ std) of optimizers including ZO-O-GD, ZO-O-Adam, ZO-O-SignSGD, ZO-O-ConservSGD, ZO-SOGD, and ZO-SOGD (Adam) on **online adversarial attack** for MNIST data across five runs.

Bilevel Optimization for Black-Box Adversarial Attacks (BBAA) Deep neural networks are vulnerable to adversarial examples—inputs subtly perturbed to mislead classifiers. These examples can fool models without access to their internals, as in [11, 52, 13]. We first review the ZO single-level formulation for BBAA [11]. Let $(\mathbf{a}, b)$ be a clean image $\mathbf{a} \in \mathbb{R}^d$ with label $b \in \{1, \dots, J\}$, and define $\mathbf{a}' = \mathbf{a} + \mathbf{y}$, where $\mathbf{y}$ is the adversarial perturbation. Let $\mathcal{Y} := [-5, 5]^d$, and $\ell : \mathbb{R}^d \rightarrow \mathbb{R}$ be the black-box attack loss. For a given hyperparameter $\lambda > 0$, the BBAA problem is:

$$\min_{\mathbf{y} \in \mathcal{Y}} \frac{1}{m} \sum_{i=1}^m \ell(\mathbf{a}_i + \mathbf{y}) + \lambda \|\mathbf{y}\|^2. \quad (32)$$

To adapt (32) 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 $s \in [S]$:

$$\mathbf{x}_s^* \in \operatorname{argmin}_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^{T_s} f(\mathbf{y}_s^*(\mathbf{x}); \mathcal{D}_t^{\text{val}}) \quad \text{s.t.} \quad \mathbf{y}_s^*(\mathbf{x}) \in \operatorname{argmin}_{\mathbf{y} \in \mathcal{Y}} \sum_{t=1}^{T_s} g(\mathbf{x}, \mathbf{y}; \mathcal{D}_t^{\text{tr}}) \quad (33)$$

$$g(\mathbf{x}_t, \mathbf{y}_t; \mathcal{D}_t^{\text{tr}}) = \frac{1}{|\mathcal{D}_t^{\text{tr}}|} \sum_{i \in \mathcal{D}_t^{\text{tr}}} \ell(\mathbf{a}_t^{(i)} + \mathbf{y}_t) + \frac{1}{2} \sum_{\iota=1}^p e^{[\mathbf{x}_t]_{\iota}} [\mathbf{y}_t]_{\iota}^2, \quad (34a)$$

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

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 i th sample in that batch; and $[\mathbf{x}_t]_{\iota}$ and $[\mathbf{y}_t]_{\iota}$ denote the ι th 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 (34) 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. (33) 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 (32) enables finer control over model components, enhancing performance for complex models and heterogeneous data [53]. For a fair comparison with single-level BBAA, we replace λ with a fixed vector multiplied by each component of $\mathbf{y}$ in (32). 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 [59]; **ZO-O-Adam**, a single-level method that updates $\mathbf{y}_t$ with a fixed $\mathbf{x}$ at each timestep using ZO Adam [45, 13]; **ZO-O-SignSGD**, a single-level method that updates $\mathbf{y}_t$ with a fixed $\mathbf{x}$ at each timestep using ZO SignSGD [6]; and **ZO-O-ConservSGD**, a single-level method that updates $\mathbf{y}_t$ with a fixed $\mathbf{x}$ at each timestep using ZO Conservative SGD [44]. Note that ZO-SOGD (ours, Adam) is a variant of our algorithm with an adaptive stepsize, similar to that of [45].

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 a slower runtime than single-level baselines due to outer-level optimization on $\mathbf{x}$. The middle panel illustrates that accuracy decreases as the adversarial attack $\mathbf{y}$ strengthens, with ZO-SOGD outperforming ZO-O-GD and ZO-O-ConservGD, while ZO-SOGD (Adam) surpasses ZO-O-Adam and all baselines. The right panel indicates that the infinity norm of $\mathbf{y}_t$ increases over time for all methods, reducing accuracy. However, perturbations remain minor, with $\max \mathbf{y}_t$ not exceeding 4, demonstrating that ZO-SOGD achieves effective attacks with superior performance.

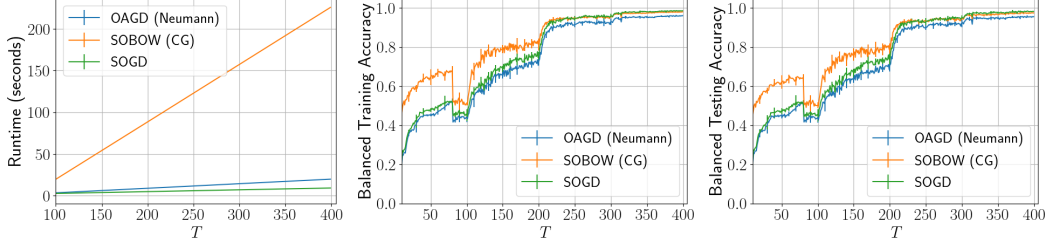


Figure 3: Performance (mean±std) on **online parametric loss tuning** with distribution shift on MNIST across five runs, comparing OGD [72], OAGD [67], SOBOW [50], and our SOGD.

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 [49]. We consider an optimization problem similar to that in (33). For a new sample $(\mathbf{a}_t, b_t)$, the follower and leader incur a parametric and balanced cross-entropy loss, respectively:

$$g(\mathbf{x}_t, \mathbf{y}_t; \mathcal{D}_t^{\text{tr}}) = -\log \frac{e^{\gamma_{b_t} [\mathbf{y}_t(\mathbf{a}_t)]_{b_t} + \Delta_{b_t}}}{\sum_{j=1}^J e^{\gamma_j [\mathbf{y}_t(\mathbf{a}_t)]_j + \Delta_j}}, \quad f(\mathbf{y}_t(\mathbf{x}_t); \mathcal{D}_t^{\text{val}}) = -u_{b_t} \log \frac{e^{[\mathbf{y}_t(\mathbf{a}_t)]_{b_t}}}{\sum_{j=1}^J e^{[\mathbf{y}_t(\mathbf{a}_t)]_j}}. \quad (35)$$

Here, $\mathbf{x}_t := (\Delta_j, \gamma_j)_{j=1}^J$ represents the logit 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 [49].

In (35), $\mathbf{y}_t(\mathbf{x}_t)$ is the follower conditioned on the leader, and $[\mathbf{y}_t(\mathbf{a}_t)]_{b_t}$ is the logit for class b_t on sample $\mathbf{a}_t$. The follower $\mathbf{y}_t$ uses a 4-layer CNN, inducing a nonconvex bilevel objective. We compare SOGD with **OAGD** [67], a static method using the Neumann series, and **SOBOW** [50], a dynamic method using conjugate gradients (CG). Experiments were conducted on MNIST [48] with batch size 64. We evaluated cumulative runtime, test accuracy, and balanced accuracy, defined as $\frac{1}{J} \sum_{j=1}^J \mathbb{P}_{\mathbf{a}_t \sim \mathcal{D}_j} [\arg\max_i ([\mathbf{y}_t(\mathbf{a}_t)]_i) = j]$, where $\mathcal{D}_j$ is the class- j sample distribution [49]. Learning rates were tuned as $\beta_t = \delta_t = \beta \in \{0.001, 0.005, 0.01, 0.05, 0.1\}$, $\alpha_t = \alpha \in \{0.0001, 0.0005, 0.001, 0.005, 0.01\}$, and $\gamma_t = \lambda_t = \eta_t = \gamma \in \{0.9, 0.99, 0.999\}$. Both OAGD and SOBOW used 5 iterations for their respective system solvers.

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.

5 Conclusion

This work introduced a novel online bilevel optimization framework that overcomes the limitations of existing algorithms, which often depend on extensive oracle information and incur high computational costs. Our method leverages 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 validate its effectiveness. A limitation of this study is that the results focus on nonconvex regret bounds, without extending guarantees to convex settings.

Acknowledgments and Disclosure of Funding

We thank the reviewers for their valuable comments. The work of DAT was supported by Samsung SDS Research America, Mountain View. The work of GM was supported in part by NSF grants DMS-2348640 and DMS-2319552.

References

- [1] Alekh Agarwal, Ofer Dekel, and Lin Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Colt*, pages 28–40. Citeseer, 2010.
- [2] Naman Agarwal, Alon Gonen, and Elad Hazan. Learning in non-convex games with an optimization oracle. In *Conference on Learning Theory*, pages 18–29. PMLR, 2019.
- [3] Alireza Aghasi and Saeed Ghadimi. Fully zeroth-order bilevel programming via gaussian smoothing. *arXiv preprint arXiv:2404.00158*, 2024.
- [4] Zeyuan Allen-Zhu and Yuanzhi Li. Neon2: Finding local minima via first-order oracles. *Advances in Neural Information Processing Systems*, 31, 2018.
- [5] Francis Bach and Vianney Perchet. Highly-smooth zero-th order online optimization. In *Conference on Learning Theory*, pages 257–283. PMLR, 2016.
- [6] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. Signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pages 560–569. PMLR, 2018.
- [7] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.
- [8] Jason Bohne, David Rosenberg, Gary Kazantsev, and Pawel Polak. Online nonconvex bilevel optimization with bregman divergences. *arXiv preprint arXiv:2409.10470*, 2024.
- [9] Jerome Bracken and James T McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1):37–44, 1973.
- [10] Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, and Rémi Munos. Online optimization in x-armed bandits. *Advances in Neural Information Processing Systems*, 21, 2008.
- [11] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 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*, pages 15–26, 2017.
- [12] Tianyi Chen, Yuejiao Sun, and Wotao Yin. Closing the gap: Tighter analysis of alternating stochastic gradient methods for bilevel problems. *Advances in Neural Information Processing Systems*, 34, 2021.
- [13] Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. *Advances in neural information processing systems*, 32, 2019.
- [14] Caroline Crockett, Jeffrey A Fessler, et al. Bilevel methods for image reconstruction. *Foundations and Trends® in Signal Processing*, 15(2-3):121–289, 2022.
- [15] Mathieu Dagréou, Pierre Ablin, Samuel Vaiter, and Thomas Moreau. A framework for bilevel optimization that enables stochastic and global variance reduction algorithms. *arXiv preprint arXiv:2201.13409*, 2022.
- [16] Stephan Dempe. *Foundations of bilevel programming*. Springer Science & Business Media, 2002.
- [17] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.
- [18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

- [19] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. In *International Conference on Machine Learning*, pages 1920–1930. PMLR, 2019.
- [20] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.
- [21] Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. In *International Conference on Machine Learning*, pages 1165–1173. PMLR, 2017.
- [22] Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.
- [23] Xiand Gao, Xiaobo Li, and Shuzhong Zhang. Online learning with non-convex losses and non-stationary regret. In *International Conference on Artificial Intelligence and Statistics*, pages 235–243. PMLR, 2018.
- [24] Xiang Gao, Bo Jiang, and Shuzhong Zhang. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76:327–363, 2018.
- [25] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.
- [26] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.
- [27] Saeed Ghadimi and Mengdi Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.
- [28] Gautam Goel, Yiheng Lin, Haoyuan Sun, and Adam Wierman. Beyond online balanced descent: An optimal algorithm for smoothed online optimization. *Advances in Neural Information Processing Systems*, 32, 2019.
- [29] Ziwei Guan, Yi Zhou, and Yingbin Liang. On the hardness of online nonconvex optimization with single oracle feedback. In *The Twelfth International Conference on Learning Representations*, 2023.
- [30] Ziwei Guan, Yi Zhou, and Yingbin Liang. Online nonconvex optimization with limited instantaneous oracle feedback. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3328–3355. PMLR, 2023.
- [31] Nadav Hallak, Panayotis Mertikopoulos, and Volkan Cevher. Regret minimization in stochastic non-convex learning via a proximal-gradient approach. In *International Conference on Machine Learning*, pages 4008–4017. PMLR, 2021.
- [32] Pierre Hansen, Brigitte Jaumard, and Gilles Savard. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on scientific and Statistical Computing*, 13(5):1194–1217, 1992.
- [33] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.
- [34] Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.
- [35] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.
- [36] Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*, pages 1433–1441. PMLR, 2017.

- [37] Amélie Héliou, Matthieu Martin, Panayotis Mertikopoulos, and Thibaud Rahier. Online non-convex optimization with imperfect feedback. *Advances in Neural Information Processing Systems*, 33:17224–17235, 2020.
- [38] Amélie Héliou, Matthieu Martin, Panayotis Mertikopoulos, and Thibaud Rahier. Zeroth-order non-convex learning via hierarchical dual averaging. In *International Conference on Machine Learning*, pages 4192–4202. PMLR, 2021.
- [39] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70, 2022.
- [40] Yu Huang, Yuan Cheng, Yingbin Liang, and Longbo Huang. Online min-max problems with non-convexity and non-stationarity. *Transactions on Machine Learning Research*, 2023.
- [41] Yu Huang, Yuan Cheng, Yingbin Liang, and Longbo Huang. Online min-max problems with non-convexity and non-stationarity. *Transactions on Machine Learning Research*, 2023.
- [42] Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International conference on machine learning*, pages 3100–3109. PMLR, 2019.
- [43] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. In *International Conference on Machine Learning*, pages 4882–4892. PMLR, 2021.
- [44] Bumsu Kim, HanQin Cai, Daniel McKenzie, and Wotao Yin. Curvature-aware derivative-free optimization. *arXiv preprint arXiv:2109.13391*, 2021.
- [45] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2014.
- [46] Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690, 2008.
- [47] Walid Krichene, Maximilian Balandat, Claire Tomlin, and Alexandre Bayen. The hedge algorithm on a continuum. In *International Conference on Machine Learning*, pages 824–832. PMLR, 2015.
- [48] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010.
- [49] Mingchen Li, Xuechen Zhang, Christos Thrampoulidis, Jiasi Chen, and Samet Oymak. Autobalance: Optimized loss functions for imbalanced data. *Advances in Neural Information Processing Systems*, 34:3163–3177, 2021.
- [50] Sen Lin, Daouda Sow, Kaiyi Ji, Yingbin Liang, and Ness Shroff. Non-convex bilevel optimization with time-varying objective functions. *Advances in Neural Information Processing Systems*, 36, 2024.
- [51] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.
- [52] Sijia Liu, Jie Chen, Pin-Yu Chen, and Alfred Hero. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In *International Conference on Artificial Intelligence and Statistics*, pages 288–297. PMLR, 2018.
- [53] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pages 1540–1552. PMLR, 2020.
- [54] Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33:20566–20577, 2020.

- [55] Yibing Lv, Tiesong Hu, Guangmin Wang, and Zhongping Wan. A penalty function method based on kuhn–tucker condition for solving linear bilevel programming. *Applied Mathematics and Computation*, 188(1):808–813, 2007.
- [56] Parvin Nazari, Ahmad Mousavi, Davoud Ataee Tarzanagh, and George Michailidis. A penalty-based method for communication-efficient decentralized bilevel programming. *Automatica*, 173:112039, 2025.
- [57] Parvin Nazari, Davoud Ataee Tarzanagh, and George Michailidis. Adaptive first-and zeroth-order methods for weakly convex stochastic optimization problems. *arXiv preprint arXiv:2005.09261*, 2020.
- [58] Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103:127–152, 2005.
- [59] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.
- [60] Abhishek Roy, Krishnakumar Balasubramanian, Saeed Ghadimi, and Prasant Mohapatra. Stochastic zeroth-order optimization under nonstationarity and nonconvexity. *Journal of Machine Learning Research*, 23(64):1–47, 2022.
- [61] Shai Shalev-Shwartz et al. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.
- [62] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.
- [63] Daouda Sow, Kaiyi Ji, and Yingbin Liang. On the convergence theory for hessian-free bilevel algorithms. *Advances in Neural Information Processing Systems*, 35:4136–4149, 2022.
- [64] Heinrich von Stackelberg. Theory of the market economy. *Oxford University Press*, 1952.
- [65] Bradly Stadie, Lunjun Zhang, and Jimmy Ba. Learning intrinsic rewards as a bi-level optimization problem. In *Conference on Uncertainty in Artificial Intelligence*, pages 111–120. PMLR, 2020.
- [66] Arun Sai Suggala and Praneeth Netrapalli. Online non-convex learning: Following the perturbed leader is optimal. In *Algorithmic Learning Theory*, pages 845–861. PMLR, 2020.
- [67] Davoud Ataee Tarzanagh, Parvin Nazari, Bojian Hou, Li Shen, and Laura Balzano. Online bilevel optimization: Regret analysis of online alternating gradient methods. In *International Conference on Artificial Intelligence and Statistics*, pages 2854–2862. PMLR, 2024.
- [68] Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*, 2020.
- [69] Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $\mathcal{O}(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. *arXiv preprint arXiv:2312.03807*, 2023.
- [70] Yan Zhang, Yi Zhou, Kaiyi Ji, and Michael M Zavlanos. Boosting one-point derivative-free online optimization via residual feedback. *arXiv preprint arXiv:2010.07378*, 2020.
- [71] Wei Zhou, Yiyang Li, Yongxin Yang, Huaimin Wang, and Timothy Hospedales. Online meta-critic learning for off-policy actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17662–17673, 2020.
- [72] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

Contents

1	Introduction	1
2	Stochastic OBO with Access to First- and Inner Second-Order Oracles	2
3	Stochastic OBO with Zeroth-Order Oracles	6
4	Experimental Results	8
5	Conclusion	10
A	Related Work	16
B	Additional Preliminaries and Notations	16
B.1	Preliminary Lemmas	16
B.2	Examples Illustrating Regularity Conditions	17
C	Proof of Regret Bounds for Simultaneous Online Gradient Descent (SOGD)	19
C.1	Proof of Lemma 2.1	19
C.2	Bounds on the Inner Decision Variable	19
C.3	Bounds on the Linear System Solution	24
C.4	Bounds on the Gradient Estimation Error of Outer Objective	30
C.5	Bounds on the Outer Objective and its Projected Gradient	33
C.6	Proof of Theorem 2.6	35
D	Proof of Regret Bounds for Zeroth Order SOGD (ZO-SOGD)	43
D.1	Auxiliary Lemmas for Proof of Theorem 3.2	43
D.2	Perturbation Bounds for OBO Objectives and Their Smoothing Variants	44
D.3	Bounds on the Zeroth-Order Inner Solution	47
D.4	Bounds on the Zeroth-Order System Solution	51
D.5	Bounds on the Zeroth-Order Estimation Error of Outer Objective	60
D.6	Bounds on the Zeroth-Order Objective Function and its Projected Gradients	64
D.7	Proof of Theorem 3.2	70
E	Hyperparameter Tuning Results	80

A Related Work

BO was introduced in game theory by [64] and modeled mathematically in [9]. Initial works [32, 55] reduced it to single-level optimization. Recently, gradient-based approaches have gained popularity for their simplicity and efficacy [21, 27, 43, 12, 56], though they assume offline objectives.

OBO was initiated by [67], proposing the OAGD method with regret bounds. [40] developed algorithms for online minimax optimization, special cases of OBO with local regret guarantees. [50] 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 [72, 34, 61]. Non-convex online optimization [36, 30, 29] poses greater challenges than its convex counterparts [61, 72, 35, 7]. Notable contributions in this field include adversarial multi-armed bandit algorithms [10, 37, 38, 47] and the Follow-the-Perturbed-Leader approach [2, 46, 66]. Hazan et al. [36] introduced window-smoothed local regret for gradient averaging in non-convex models, which Hallak et al. [31] extended to non-smooth, non-convex problems. Inspired by their work, we employ local regret for Online Bilevel Optimization (OBO) without window-smoothing.

Zeroth-Order Optimization. Single-Level ZO Optimization has been widely studied in both offline [25, 17, 1, 59, 57] and online settings [52, 29, 30, 70, 5]. We next review closely related work. Liu et al. [52] proposed ZOO-ADMM, a gradient-free online optimization algorithm utilizing ADMM. Guan et al. [30] studied online non-convex optimization with limited oracle feedback. Research on online non-convex optimization with bandit feedback includes work by Heliou et al. [37], which established bounds on global static and dynamic regret using dual averaging, further refined in [38]. Gao et al. [23] extended these ideas to ZO algorithms. Flaxman et al. [20] provided algorithms for bandit online optimization of convex functions using ZO gradient approximation. Our work closely relates to [63], 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 [3] 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 [26]). 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

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

where

$$\mathbf{x}^+ = \Pi_{\mathcal{X}}(\mathbf{x} - \alpha_t \mathbf{p}), \quad (36)$$

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

Lemma B.2. [28, Lemma 13] *If $f : \mathcal{X} \rightarrow \mathbb{R}$ 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 $\forall \mathbf{x} \in \mathcal{X}$,*

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

Lemma B.3. *Suppose $f(\mathbf{x})$ is L -smooth, and $\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$. 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:*

$$\frac{1}{2L} \|\nabla f(\mathbf{x})\|^2 \leq f(\mathbf{x}) - f(\mathbf{x}^*) \leq \frac{L}{2} \|\mathbf{x} - \mathbf{x}^*\|^2. \quad (37)$$

Lemma B.4. For any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the following holds for any $c > 0$:

$$\|\mathbf{x} + \mathbf{y}\|^2 \leq (1 + c)\|\mathbf{x}\|^2 + \left(1 + \frac{1}{c}\right)\|\mathbf{y}\|^2.$$

We also utilize a basic yet important property of the projected-gradient mapping.

Lemma B.5. [26, Proposition 1] Let $\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}; \mathbf{p})$ denote the projected gradient as defined in Definition B.1. For any $\mathbf{x}, \mathbf{p}_1, \mathbf{p}_2 \in \mathbb{R}^d$ and $\alpha_t > 0$, it holds that

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

Lemma B.6. [36, Proposition 2.4] Let $\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}; \mathbf{p})$ denote the projected gradient as defined in 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.7. 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 (36) 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 Illustrating Regularity Conditions

Theorem 2.6 achieves sublinear bilevel regret when the variations V_T and $H_{2,T}$ are $o(T^{2/3})$ and $o(T^{1/3})$, respectively. Below, we provide some examples of online optimization in both single-level and bilevel settings to illustrate when this occurs.

Example B.8. Consider function $f_t(\mathbf{x}) = \|\mathbf{A}_t \mathbf{x} - \mathbf{b}_t\|^2$, where $\mathbf{A}_t = [1, 0; 0, 1 + \frac{1}{t}]$, $\mathbf{b}_t = (1, 1)$. It follows from (11) that $V_T = \sum_{t=2}^T \max_{\mathbf{x}} |f_t(\mathbf{x}) - f_{t-1}(\mathbf{x})| = \sum_{t=2}^T \left| \left(\frac{1}{t}\right)^2 - \left(\frac{1}{t-1}\right)^2 \right|$, and

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

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 $T \rightarrow \infty$, V_T becomes bounded and approaches a constant value, indicating that V_T grows slower than T itself.

Example B.9. Let

$$f_t(\mathbf{x}) = \begin{cases} \left(-\frac{1}{t}, 0\right) & \text{if } t \text{ is even;} \\ \left(0, -\frac{1}{t}\right) & \text{if } t \text{ is odd.} \end{cases}$$

Then, $V_T = \sum_{t=2}^T \max_{\mathbf{x}} |f_t(\mathbf{x}) - f_{t-1}(\mathbf{x})| = \mathcal{O}(1)$.

Example B.10. 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

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

and

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

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

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

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 $V_T = o(T^{2/3})$ and $H_{2,T} = o(T^{1/3})$, 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 $\mathcal{O}(1/t^a)$ with $a > 1/3$. For example, if the coefficients in the functions change as $\mathcal{O}(1/t^a)$ with $a > 1/3$, then the cumulative sum over T will be $o(T^{2/3})$. 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, $\mathbf{d}_t^y$, and the true direction, $\nabla_y g_t(\mathbf{x}_t, \mathbf{y}_t)$, 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.6, which quantifies the error between the approximated direction of the momentum-based gradient estimator, $\mathbf{d}_t^y$, and the true direction, $\nabla_y^2 g_t(\mathbf{z}_t) \mathbf{v}_t + \nabla_y f_t(\mathbf{z}_t)$, 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.6. Moreover, we provide Lemma C.9, which quantifies the error between the approximated direction of the momentum-based hypergradient estimator, $\mathbf{d}_t^x$, and the true direction, $\nabla_x f_t(\mathbf{z}_t) + \nabla_{xy}^2 g_t(\mathbf{z}_t) \mathbf{v}_t$, 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 2.1

Proof. By letting $\nu = 1 - \eta$ for $\eta \in (0, 1)$, the window-smoothed gradient

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

is equivalent to

$$\nabla F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) = \frac{1}{W} \sum_{j=t-w+1}^t (1-\eta)^{t-j} \nabla f_j(\mathbf{x}_j, \mathbf{y}_j; \mathcal{B}_j). \quad (38)$$

Let $\mathbf{d}_t^x = \nabla F_{t,\nu}(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t)$. Then (38) is equivalent to

$$\mathbf{d}_t^x = \frac{1}{W} \nabla f_t(\mathbf{x}_t, \mathbf{y}_t; \mathcal{B}_t) + \frac{1}{W} \sum_{j=t-w+1}^{t-1} (1-\eta)^{t-j} \nabla f_j(\mathbf{x}_j, \mathbf{y}_j; \mathcal{B}_j).$$

Since

$$(1-\eta)\mathbf{d}_{t-1}^x = \frac{(1-\eta)}{W} \sum_{j=t-w}^{t-1} (1-\eta)^{t-1-j} \nabla f_j(\mathbf{x}_j, \mathbf{y}_j; \mathcal{B}_j),$$

we have

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

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

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

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

□

C.2 Bounds on the Inner Decision Variable

In the following, inspired by offline BO [69, 15] and OBO [67, 50], we provide a set of lemmas for the analysis of SOGD. We first present a lemma that characterizes the Lipschitz continuity of the approximate gradients, as well as the inner and system solutions.

Lemma C.1. Under Assumptions 2.2 and 2.3, for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and the search directions $\{\mathbf{d}_t^{\mathbf{x}}\}_{t=1}^T$ and $\{\mathbf{d}_t^{\mathbf{y}}\}_{t=1}^T$ generated by Algorithm 1, we have

$$\|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\|^2 \leq M_f^2 \left(\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 + \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \right), \quad (39a)$$

$$\|\mathbf{d}_t^{\mathbf{y}}\|^2 \leq M_v^2 \left(\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 + \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \right), \quad (39b)$$

$$\|\nabla f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) - \nabla f_t(\mathbf{x}', \mathbf{y}_t^*(\mathbf{x}'))\| \leq L_f \|\mathbf{x} - \mathbf{x}'\|, \quad (39c)$$

$$\|\mathbf{y}_t^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x}')\| \leq L_y \|\mathbf{x} - \mathbf{x}'\|, \quad (39d)$$

$$\|\mathbf{v}_t^*(\mathbf{x}) - \mathbf{v}_t^*(\mathbf{x}')\| \leq L_v \|\mathbf{x} - \mathbf{x}'\|, \quad (39e)$$

where M_f , M_v , and (L_y, L_v, L_f) are defined in (42), (43), and (44), respectively.

Proof. We first show (39a).

Using Assumptions 2.2 and 2.3, we have $\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \succeq \mu_g$, and

$$\|\mathbf{v}_t^*(\mathbf{x}_t)\| = \|(\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))^{-1} \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| \leq \frac{\ell_{f,0}}{\mu_g}. \quad (40)$$

Observe that

$$\begin{aligned} \|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| &\leq \|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| \\ &\quad + \|\mathbf{v}_t \nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{v}_t^*(\mathbf{x}_t) \nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| \\ &\leq \|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| \\ &\quad + \|\nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)\| \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\| \\ &\quad + \|\mathbf{v}_t^*(\mathbf{x}_t)\| \|\nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\| \\ &\leq \left(\ell_{f,1} + \frac{\ell_{g,2} \ell_{f,0}}{\mu_g} \right) \|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\| + \ell_{g,1} \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\| \\ &\leq M_f^2 (\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\| + \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|), \end{aligned} \quad (41)$$

where

$$M_f := \sqrt{2} \max \left\{ \ell_{f,1} + \frac{\ell_{g,2} \ell_{f,0}}{\mu_g}, \ell_{g,1} \right\}, \quad (42)$$

the third inequality is by Assumption 2.3, and the last inequality follows from (40).

Next, we establish (39b).

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

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

Then, from Assumption 2.3 and (40), we have

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

where

$$M_v := \sqrt{2} \max \left\{ \frac{\ell_{g,2} \ell_{f,0}}{\mu_g} + \ell_{f,1}, \ell_{g,1} \right\}. \quad (43)$$

The proofs of Eqs. (39c)-(39e) follow from [67, Lemma 17] by setting

$$\begin{aligned} L_{\mathbf{y}} &:= \frac{\ell_{g,1}}{\mu_g}, \\ L_{\mathbf{v}} &:= \ell_{f,1} + \frac{\ell_{g,1}\ell_{f,1}}{\mu_g} + \frac{\ell_{f,0}}{\mu_g} \left(\ell_{g,2} + \frac{\ell_{g,1}\ell_{g,2}}{\mu_g} \right), \\ L_f &:= \ell_{f,1} + \frac{\ell_{g,1}(\ell_{f,1} + M_f)}{\mu_g} + \frac{\ell_{f,0}}{\mu_g} \left(\ell_{g,2} + \frac{\ell_{g,1}\ell_{g,2}}{\mu_g} \right), \end{aligned} \quad (44)$$

where the other constants are defined in Assumption 2.3. $\square$

The following lemma is inspired by [69] and can be viewed as an extension of [69] to the online setting.

Lemma C.2. *Suppose Assumptions 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*

$$e_t^g := \mathbf{d}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t), \quad (45)$$

we have:

$$\begin{aligned} \mathbb{E}\|e_{t+1}^g\|^2 &\leq (1 - \gamma_{t+1})^2(1 + 48\ell_{g,1}^2\beta_t^2)\mathbb{E}\|e_t^g\|^2 + 2\gamma_{t+1}^2 \frac{\sigma_{g_{\mathbf{y}}}^2}{b} + 24(1 - \gamma_{t+1})^2\ell_{g,1}^2\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &\quad + 6(1 - \gamma_{t+1})^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{z}_{t+1})\|^2 \\ &\quad + 48(1 - \gamma_{t+1})^2\ell_{g,1}^2\beta_t^2\mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2. \end{aligned} \quad (46)$$

Proof. From Algorithm 1, we have

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

Then, we have

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

which implies that

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

where the second inequality follows from Cauchy–Schwartz inequality and Assumption C1.. Moreover, from Cauchy–Schwartz inequality, we have

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

From Assumption B3., we have

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

and

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

From the two inequalities above, we have

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

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

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

□

Lemma C.3. Suppose Assumptions 2.2, and B3. hold. Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$ generated by Algorithm 1, we have

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

where e_t^g defined in (45), $\mathbf{y}_t^*(\mathbf{x}_t)$ is defined in (1) and $a > 0$ is a constant.

Proof. From Lemma B.4, we have

$$\begin{aligned}
\mathbb{E} [\|\mathbf{y}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] & = \mathbb{E} [\|\mathbf{y}_t - \beta_t \mathbf{d}_t^{\mathbf{y}} - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] \\
& \leq (1 + a) \mathbb{E} [\|\mathbf{y}_t - \beta_t \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] \\
& \quad + (1 + \frac{1}{a}) \beta_t^2 \mathbb{E} [\|\mathbf{d}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2].
\end{aligned} \tag{47}$$

Next, we will bound the first term on the RHS of (47).
We have

$$\begin{aligned}
\mathbb{E} [\|\mathbf{y}_t - \beta_t \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] &= \mathbb{E} [\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] + \beta_t^2 \mathbb{E} [\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2] \\
&\quad - 2\beta_t \mathbb{E} [\langle \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t) \rangle] \\
&\leq \left(1 - 2\beta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right) \mathbb{E} [\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] \\
&\quad - \left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right) \mathbb{E} [\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2], \tag{48}
\end{aligned}$$

where the inequality results from the strong convexity of g_t by Assumption 2.2, 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 (48) into (47), gives the desired result. $\square$

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

$$\theta_t^{\mathbf{y}} := \|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2, \quad \text{and} \quad \theta_t^{\mathbf{v}} := \|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2. \tag{49}$$

The following lemma, inspired by the offline bilevel optimization framework in [69], characterizes the descent behavior of the iterates in the inner problem.

Lemma C.4. *Suppose Assumptions 2.2, and B3. hold. Let $\theta_t^{\mathbf{y}}$ be defined as in (49). Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$ generated by Algorithm 1, the following bound is guaranteed:*

$$\begin{aligned}
&\sum_{t=1}^T (\mathbb{E}[\theta_{t+1}^{\mathbf{y}}] - \mathbb{E}[\theta_t^{\mathbf{y}}]) \\
&\leq -\frac{L_{\mu_g}}{2} \sum_{t=1}^T \beta_t \mathbb{E}[\theta_t^{\mathbf{y}}] + \frac{2}{L_{\mu_g}} \sum_{t=1}^T \beta_t \mathbb{E}[\|e_t^g\|^2] + \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}} \sum_{t=1}^T \frac{1}{\beta_t} \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2] \\
&\quad + \frac{4}{L_{\mu_g}} \sum_{t=2}^T \frac{1}{\beta_t} \sup_{\mathbf{x} \in \mathcal{X}} \mathbb{E}[\|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2] + \sum_{t=1}^T \left(-\frac{2\beta_t}{\mu_g + \ell_{g,1}} + \beta_t^2\right) \mathbb{E}[\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2], \tag{50}
\end{aligned}$$

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 (44); $H_{2,T}$ is defined in (11). Moreover, e_t^g is defined in (45).

Proof. From Lemma B.4, we have for any $\hat{c} > 0$

$$\begin{aligned}
\mathbb{E} [\|\mathbf{y}_{t+1} - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\|^2] &= \mathbb{E} [\|\mathbf{y}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t) + \mathbf{y}_t^*(\mathbf{x}_t) - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\|^2] \\
&\leq (1 + \hat{c}) \mathbb{E} [\|\mathbf{y}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] \\
&\quad + \left(1 + \frac{1}{\hat{c}}\right) \mathbb{E} [\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2]. \tag{51}
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E} [\|\mathbf{y}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] &\leq (1 + a) \left(1 - 2\beta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right) \mathbb{E} [\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] \\
&\quad + \left(-(1 + a) \left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right) \mathbb{E} [\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2] \\
&\quad + \left(1 + \frac{1}{a}\right) \beta_t^2 \mathbb{E} [\|e_t^g\|^2]. \tag{52}
\end{aligned}$$

Substituting (52) into (51), we get

$$\begin{aligned}
& \mathbb{E} [\|\mathbf{y}_{t+1} - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\|^2] \\
& \leq (1 + \acute{c})(1 + a) \left(1 - 2\beta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right) \mathbb{E} [\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] \\
& + \left(-(1 + \acute{c})(1 + a) \left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right) \mathbb{E} [\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2] \\
& + (1 + \acute{c})(1 + \frac{1}{a}) \beta_t^2 \mathbb{E} [\|e_t^g\|^2] \\
& + \left(1 + \frac{1}{\acute{c}}\right) \mathbb{E} [\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2]. \tag{53}
\end{aligned}$$

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

$$\begin{aligned}
(1 + \acute{c})(1 + a) (1 - 2\beta_t L_{\mu_g}) &= 1 - \frac{\beta_t L_{\mu_g}}{2}, \\
(1 + a) (1 - 2\beta_t L_{\mu_g}) &= 1 - \beta_t L_{\mu_g}, \\
(1 + \acute{c}) (1 - \beta_t L_{\mu_g}) &= 1 - \frac{\beta_t L_{\mu_g}}{2}, \\
1 + \frac{1}{a} &\leq \frac{1}{\beta_t L_{\mu_g}}, \quad 1 + \frac{1}{\acute{c}} \leq \frac{2}{\beta_t L_{\mu_g}}. \tag{54}
\end{aligned}$$

Based on (53) and (54), we get

$$\begin{aligned}
& \mathbb{E} [\|\mathbf{y}_{t+1} - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\|^2] - \mathbb{E} [\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] \\
& \leq -\frac{\beta_t L_{\mu_g}}{2} \mathbb{E} [\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] + \left(-\left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right) \mathbb{E} [\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2] \\
& + \frac{2}{\beta_t L_{\mu_g}} \beta_t^2 \mathbb{E} [\|e_t^g\|^2] + \frac{2}{\beta_t L_{\mu_g}} \mathbb{E} [\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2]. \tag{55}
\end{aligned}$$

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

$$\begin{aligned}
& \mathbb{E} [\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] \\
& \leq 2 (\mathbb{E} [\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_{t+1}^*(\mathbf{x}_t)\|^2] + \mathbb{E} [\|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2]) \\
& \leq 2 (L_{\mathbf{y}}^2 \mathbb{E} [\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2] + \mathbb{E} [\|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2]), \tag{56}
\end{aligned}$$

where the second inequality is by Eq. (39d) in Lemma C.1.

Substituting (56) into (55) and summing over $t \in [T]$, give the desired result. $\square$

C.3 Bounds on the Linear System Solution

Lemma C.5. Suppose Assumptions 2.2 and B3. hold. Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 1, we have

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

for any $\acute{c} > 0$, where $\mathbf{v}_t^*(\mathbf{x}_t)$ is the solution of the system in Eq. (4), and e_t^v is defined in (60).

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

$$\begin{aligned}
\mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 &= \mathbb{E}\|\Pi_{\mathcal{Z}_p}[\mathbf{v}_t - \delta_t \mathbf{d}_t^y] - \Pi_{\mathcal{Z}_p}[\mathbf{v}_t^*(\mathbf{x}_t)]\|^2 \\
&\leq \mathbb{E}\|\mathbf{v}_t - \delta_t \mathbf{d}_t^y - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \\
&\leq (1 + \hat{c})\mathbb{E}\|\mathbf{v}_t - \delta_t \nabla P_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t), \mathbf{v}_t) - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \\
&\quad + (1 + \frac{1}{\hat{c}})\delta_t^2 \mathbb{E}\|\mathbf{d}_t^y - \nabla P_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t), \mathbf{v}_t)\|^2,
\end{aligned} \tag{57}$$

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

For the first term of Eq. (57) above, we have

$$\begin{aligned}
&\mathbb{E}\|\mathbf{v}_t - \delta_t \nabla P_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t), \mathbf{v}_t) - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \\
&= \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 - 2\delta_t \mathbb{E}\langle \mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t), \nabla P_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t), \mathbf{v}_t) \rangle + \delta_t^2 \mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t), \mathbf{v}_t)\|^2 \\
&\leq \left(1 - 2\delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right) \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 - (2\delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}} - \delta_t^2) \mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t), \mathbf{v}_t)\|^2 \\
&\leq \left(1 - 2\delta_t \frac{(\ell_{g,1} + \ell_{g,1}^3) \mu_g}{\mu_g + \ell_{g,1}} + \delta_t^2 \ell_{g,1}^2\right) \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2,
\end{aligned} \tag{58}$$

where the first inequality follows from the strong convexity of the function P_t , which is the gradient of the strongly convex quadratic program $\frac{1}{2} \mathbf{v}^\top \nabla_{\mathbf{y}}^2 g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) \mathbf{v} + \mathbf{v}^\top \nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))$. Then, we have

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

The second inequality is derived from the following inequality.

$$\begin{aligned}
\mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t), \mathbf{v}_t)\|^2 &= \mathbb{E}\|\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \mathbf{v}_t + \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\|^2 \\
&= \mathbb{E}\|\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) (\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t))\|^2 \\
&\leq \ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2,
\end{aligned} \tag{59}$$

where the second equality follows from (4).

Combining (57) and (58), we get the desired result. $\square$

Lemma C.6. 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_{t+1}^y defined as

$$e_t^y := \mathbf{d}_t^y - \nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t), \quad \text{where} \tag{60a}$$

$$\nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t) := \nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v}_t + \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t), \tag{60b}$$

we have:

$$\begin{aligned}
\mathbb{E}\|e_{t+1}^y\|^2 &\leq (1 - \lambda_{t+1})^2 (1 + 72\ell_{g,1}^2 \delta_t^2) \mathbb{E}\|e_t^y\|^2 + 4\lambda_{t+1}^2 \left(\frac{\sigma_{g_{yy}}^2}{b} p^2 + \frac{\sigma_{f_y}^2}{b}\right) \\
&\quad + 12p^2 (1 - \lambda_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
&\quad + 12(1 - \lambda_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
&\quad + 72(1 - \lambda_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E}\|e_t^g\|^2 + 2\beta_t^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
&\quad + 144(1 - \lambda_{t+1})^2 \ell_{g,1}^4 \delta_t^2 \mathbb{E}[\theta_t^y] + 288\ell_{g,1}^2 (p^2 \ell_{g,2}^2 + \ell_{f,1}^2) \delta_t^2 \mathbb{E}[\theta_t^y],
\end{aligned} \tag{61}$$

for all $t \in [T]$ and (θ_t^y, θ_t^x) and e_t^g are defined in (49) and (45), respectively.

Proof. Note that

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

where

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

From Algorithm 1, we have

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

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

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

which implies that

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

where the inequality follows from Cauchy–Schwartz inequality.

For the first term, from Assumptions C2. and C4., we have

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

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

$$\begin{aligned} \mathbb{E}\|e_{t+1}^{\mathbf{v}}\|^2 &\leq (1 - \lambda_{t+1})^2 \mathbb{E}\|e_t^{\mathbf{v}}\|^2 + 4\lambda_{t+1}^2 \left(\frac{\sigma_{g_{\mathbf{y}\mathbf{y}}}^2}{b} p^2 + \frac{\sigma_{f_{\mathbf{y}}}^2}{b}\right) \\ &\quad + 6(1 - \lambda_{t+1})^2 \mathbb{E}\|\nabla P_t(\mathbf{u}_t) - \nabla P_t(\mathbf{u}_{t+1})\|^2 \\ &\quad + 6(1 - \lambda_{t+1})^2 \mathbb{E}\|\nabla P_t(\mathbf{u}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_{t+1})\|^2 \\ &\quad + 6(1 - \lambda_{t+1})^2 \mathbb{E}\|\nabla P_{t+1}(\mathbf{u}_{t+1}; \mathcal{B}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_t; \mathcal{B}_{t+1})\|^2. \end{aligned} \quad (62)$$

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

$$\begin{aligned} &\mathbb{E}\|\nabla P_t(\mathbf{u}_{t+1}) - \nabla P_t(\mathbf{u}_t)\|^2 \\ &\leq 3\mathbb{E}\|\nabla P_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{v}_{t+1}) - \nabla P_t(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{v}_{t+1})\|^2 \\ &\quad + 3\mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{v}_{t+1}) - \nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_{t+1})\|^2 \\ &\quad + 3\mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_{t+1}) - \nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2 \\ &\leq 3\mathbb{E}\|(\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_{t+1})) \mathbf{v}_{t+1} + \nabla_{\mathbf{y}} f_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_{t+1})\|^2 \\ &\quad + 3\mathbb{E}\|(\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)) \mathbf{v}_{t+1} + \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &\quad + 3\mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_{t+1}) - \nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2 \\ &\leq 6(\ell_{g,2}^2 \mathbb{E}\|\mathbf{v}_{t+1}\|^2 + \ell_{f,1}^2 (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2)) + 3\ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2, \end{aligned} \quad (63)$$

where the last inequality follows from Assumptions B2., B3. and B4.;

From Eq. (63) and the inequality $\|a + b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$, we obtain

$$\begin{aligned}
& \mathbb{E}\|\nabla P_t(\mathbf{u}_{t+1}) - \nabla P_t(\mathbf{u}_t)\|^2 \\
& \leq 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \beta_t^2 \mathbb{E}\|\mathbf{d}_t^y\|^2) + 3\ell_{g,1}^2 \delta_t^2 \mathbb{E}\|\mathbf{d}_t^y\|^2 \\
& \leq 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E}\|e_t^g\|^2 + 2\beta_t^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
& \quad + 6\ell_{g,1}^2 \delta_t^2 (\mathbb{E}\|e_t^y\|^2 + \mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2) \\
& \leq 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E}\|e_t^g\|^2 + 2\beta_t^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
& \quad + 6\ell_{g,1}^2 \delta_t^2 (\mathbb{E}\|e_t^y\|^2 + 2\mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t), \mathbf{v}_t)\|^2 + 2\mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t) - \nabla P_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t), \mathbf{v}_t)\|^2) \\
& \leq 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E}\|e_t^g\|^2 + 2\beta_t^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
& \quad + 6\ell_{g,1}^2 \delta_t^2 (\mathbb{E}\|e_t^y\|^2 + 2\ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 + 4(p^2 \ell_{g,2}^2 + \ell_{f,1}^2) \mathbb{E}\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2), \tag{64}
\end{aligned}$$

where the last inequality follows from (59).

Similarly, we have

$$\begin{aligned}
& \mathbb{E}\|\nabla P_{t+1}(\mathbf{u}_{t+1}; \mathcal{B}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_t; \mathcal{B}_{t+1})\|^2 \\
& \leq 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E}\|e_t^g\|^2 + 2\beta_t^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
& \quad + 6\ell_{g,1}^2 \delta_t^2 (\mathbb{E}\|e_t^y\|^2 + 2\ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 + 4(p^2 \ell_{g,2}^2 + \ell_{f,1}^2) \mathbb{E}\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2). \tag{65}
\end{aligned}$$

Substituting (65) and (64) into (62), we have

$$\begin{aligned}
\mathbb{E}\|e_{t+1}^y\|^2 & \leq (1 - \lambda_{t+1})^2 (1 + 72\ell_{g,1}^2 \delta_t^2) \mathbb{E}\|e_t^y\|^2 + 4\lambda_{t+1}^2 \left(\frac{\sigma_{g_{yy}}^2}{b} p^2 + \frac{\sigma_{f_y}^2}{b} \right) \\
& \quad + 6(1 - \lambda_{t+1})^2 \mathbb{E}\|\nabla P_t(\mathbf{u}_{t+1}) - \nabla P_{t+1}(\mathbf{u}_{t+1})\|^2 \\
& \quad + 72(1 - \lambda_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E}\|e_t^g\|^2 + 2\beta_t^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
& \quad + 144(1 - \lambda_{t+1})^2 \ell_{g,1}^4 \delta_t^2 \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 + 288\ell_{g,1}^2 (p^2 \ell_{g,2}^2 + \ell_{f,1}^2) \delta_t^2 \mathbb{E}\|\mathbf{y}_t - \mathbf{y}_t^*(\mathbf{x}_t)\|^2.
\end{aligned}$$

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

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

This completes the proof. $\square$

As demonstrated in Lemma C.6, the gradient estimation error e_{t+1}^y for the linear system consists of four key components: (1) an iteratively refined error term $(1 - \lambda_{t+1})^2 (1 + 72\ell_{g,1}^2 \delta_t^2) \mathbb{E}\|e_t^y\|^2$, which depends on the stepsize δ_t ; (2) the error arising from the variation in the Hessian of the lower-level objective; (3) the error resulting from the variation in the gradient of the upper-level objective, and (4) approximation error terms of order $\mathcal{O}(\delta_t^2 \mathbb{E}[\theta_t^y])$ and $\mathcal{O}(\delta_t^2 \mathbb{E}[\theta_t^y])$ associated with solving the linear system and the iterates in the inner problem, respectively.

Lemma C.7. *Suppose Assumptions 2.2, B1., B2. and B4. hold. Let $\mathbf{v}_t^*(\mathbf{x})$ is a solution of Subproblem (4). Then, we have*

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

where $\nu := \ell_{f,1} + \frac{\ell_{g,2} \ell_{f,0}}{\mu_g}$.

Proof. Based on (4), we have that

$$\begin{aligned}
& \|\mathbf{v}_t^*(\mathbf{x}_t) - \mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\|^2 \\
&= \|(\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))^{-1} \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \\
&\quad - (\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})))^{-1} \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\|^2 \\
&\leq 2 \left\| \left((\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))^{-1} - (\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})))^{-1} \right) \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \right\|^2 \quad (66a) \\
&\quad + 2 \left\| (\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})))^{-1} (\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))) \right\|^2. \quad (66b)
\end{aligned}$$

In the following steps, we bound the terms (66a) and (66b), respectively.

For (66a), we have:

$$\begin{aligned}
& \left\| (\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))^{-1} - (\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})))^{-1} \right\|^2 \\
&= \|(\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))^{-1} (\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))) \\
&\quad - \nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) (\nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})))^{-1}\|^2 \\
&\leq \frac{1}{\mu_g^2} \|\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - \nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\|^2 \\
&\leq \frac{\ell_{g,2}}{\mu_g^2} \|(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - (\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\|^2 \\
&\leq \frac{\ell_{g,2}}{\mu_g^2} (\|\mathbf{y}_t^*(\mathbf{x}_t) - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\|^2 + \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2), \quad (67)
\end{aligned}$$

where the 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 inequalities are obtained from Assumptions 2.2 and B4..

Thus, from (67) and Assumption B1., we get

$$(66a) \leq \frac{\ell_{f,0}\ell_{g,2}}{\mu_g^2} (\|\mathbf{y}_t^*(\mathbf{x}_t) - \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1})\|^2 + \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2). \quad (68)$$

For (66b), we have

$$\begin{aligned}
(66b) &\leq \frac{1}{\mu_g} \|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\|^2 \\
&\leq \frac{\ell_{f,1}}{\mu_g} \|(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - (\mathbf{x}_{t+1}, \mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}))\|^2 \\
&\leq \frac{\ell_{f,1}}{\mu_g} (\|\mathbf{y}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2). \quad (69)
\end{aligned}$$

Combining (68) and (69), we have

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

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. $\square$

The following lemma characterizes the decrease in $\theta_t^{\mathbf{y}}$ defined in (49) and can be viewed as an extension of the offline BO result in [69] to the OBO setting.

Lemma C.8. Suppose Assumptions 2.2 and 2.3 hold. Let $\theta_t^{\mathbf{y}}$ be defined in (49). Then, for any positive choice of step size δ_t as

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

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

$$\begin{aligned} \sum_{t=1}^T (\mathbb{E}[\theta_{t+1}^{\mathbf{v}}] - \mathbb{E}[\theta_t^{\mathbf{v}}]) &\leq -\frac{\delta_t \dot{L}_{\mu_g}}{4} \sum_{t=1}^T \mathbb{E}[\theta_t^{\mathbf{v}}] + \frac{4}{\dot{L}_{\mu_g}} \delta_t \sum_{t=1}^T \mathbb{E}\|e_t^{\mathbf{v}}\|^2 \\ &\quad + \frac{16\nu^2}{\dot{L}_{\mu_g} \mu_g^2 \delta_t} \sum_{t=1}^T \mathbb{E}\|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 \\ &\quad + \frac{8\nu^2}{\dot{L}_{\mu_g} \mu_g^2 \delta_t} (1 + 2L_{\mathbf{y}}^2) \sum_{t=1}^T \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2, \end{aligned} \quad (70)$$

where $e_t^{\mathbf{v}}$ is defined in (60), ν and $L_{\mathbf{y}}$, are defined in Lemmas C.7 and C.4, respectively.

Proof. By Lemma B.4, for any $a > 0$, we have

$$\begin{aligned} \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\|^2 &= \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t^*(\mathbf{x}_t) + \mathbf{v}_t^*(\mathbf{x}_t) - \mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\|^2 \\ &\leq (1+a) \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \\ &\quad + \left(1 + \frac{1}{a}\right) \mathbb{E}\|\mathbf{v}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{v}_t^*(\mathbf{x}_t)\|^2. \end{aligned} \quad (71)$$

From Lemma C.5, we have for any $\hat{c} > 0$:

$$\begin{aligned} \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 &\leq (1+\hat{c}) \left(1 - 2\delta_t \frac{(\ell_{g,1} + \ell_{g,1}^3)\mu_g}{\mu_g + \ell_{g,1}} + \delta_t^2 \ell_{g,1}^2\right) \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \\ &\quad + (1 + \frac{1}{\hat{c}}) \delta_t^2 \mathbb{E}\|e_t^{\mathbf{v}}\|^2. \end{aligned} \quad (72)$$

Substituting (72) into (71), we get

$$\begin{aligned} \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\|^2 &\leq (1+a) (1+\hat{c}) \left(1 - 2\delta_t \frac{(\ell_{g,1} + \ell_{g,1}^3)\mu_g}{\mu_g + \ell_{g,1}} + \delta_t^2 \ell_{g,1}^2\right) \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \\ &\quad + (1+a) (1 + \frac{1}{\hat{c}}) \delta_t^2 \mathbb{E}\|e_t^{\mathbf{v}}\|^2 \\ &\quad + \left(1 + \frac{1}{a}\right) \mathbb{E}\|\mathbf{v}_{t+1}^*(\mathbf{x}_{t+1}) - \mathbf{v}_t^*(\mathbf{x}_t)\|^2. \end{aligned} \quad (73)$$

In the following, we provide a bound for the third term on the right-hand side of (73). To this end, we have from Lemma C.7:

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

where the last inequality follows from Lemma C.1.

Combining this result with (73) gives

$$\begin{aligned} \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_{t+1}^*(\mathbf{x}_{t+1})\|^2 &\leq (1+a) (1+\hat{c}) \left(1 - 2\delta_t \frac{(\ell_{g,1} + \ell_{g,1}^3)\mu_g}{\mu_g + \ell_{g,1}} + \delta_t^2 \ell_{g,1}^2\right) \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2 \\ &\quad + (1+a) (1 + \frac{1}{\hat{c}}) \delta_t^2 \mathbb{E}\|e_t^{\mathbf{v}}\|^2 + 4 \left(1 + \frac{1}{a}\right) \frac{\nu^2}{\mu_g^2} \mathbb{E}\|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 \\ &\quad + 2 \left(1 + \frac{1}{a}\right) \frac{\nu^2}{\mu_g^2} (1 + 2L_{\mathbf{y}}^2) \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2. \end{aligned} \quad (74)$$

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

$$\begin{aligned} 1 - 2\delta_t \frac{(\ell_{g,1} + \ell_{g,1}^3)\mu_g}{\mu_g + \ell_{g,1}} + \delta_t^2 \ell_{g,1}^2 &= 1 - 2\delta_t \dot{L}_{\mu_g} + \delta_t^2 \ell_{g,1}^2 \\ &\leq 1 - \delta_t \dot{L}_{\mu_g}, \end{aligned} \quad (75)$$

where the last inequality follows from $\delta_t \leq \frac{\dot{L}_{\mu_g}}{\ell_{g,1}^2}$.

Choose $a = \frac{\delta_t \dot{L}_{\mu_g}/4}{1 - \frac{\delta_t \dot{L}_{\mu_g}}{2}}$ and $\hat{c} = \frac{\delta_t \dot{L}_{\mu_g}/2}{1 - \delta_t \dot{L}_{\mu_g}}$. Then, from (75), we have

$$\begin{aligned} (1+a)(1+\hat{c}) &\left(1 - 2\delta_t \frac{(\ell_{g,1} + \ell_{g,1}^3)\mu_g}{\mu_g + \ell_{g,1}} + \delta_t^2 \ell_{g,1}^2\right) \\ &\leq (1+a)(1+\hat{c}) \left(1 - \delta_t \dot{L}_{\mu_g}\right) = 1 - \frac{\delta_t \dot{L}_{\mu_g}}{4}, \\ (1+a) \left(1 + \frac{1}{\hat{c}}\right) &\leq \frac{4}{\delta_t \dot{L}_{\mu_g}}, \\ 1 + \frac{1}{\hat{c}} &\leq \frac{2}{\delta_t \dot{L}_{\mu_g}}, \quad 1 + \frac{1}{a} \leq \frac{4}{\delta_t \dot{L}_{\mu_g}}. \end{aligned} \quad (76)$$

Thus, from (74) and (76) we have

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

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

C.4 Bounds on the Gradient Estimation Error of Outer Objective

The following lemma, inspired by [69], provides a characterization of the descent of the gradient estimation error for the outer-level function.

Lemma C.9. *Suppose Assumptions B2., B3., B4., 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*

$$e_t^f := \mathbf{d}_t^{\mathbf{x}} - \tilde{\mathbf{d}}_t(\mathbf{z}_t, \mathbf{v}_t), \quad \text{where} \quad \tilde{\mathbf{d}}_t(\mathbf{z}_t, \mathbf{v}_t) = \nabla_{\mathbf{x}} f_t(\mathbf{z}_t) + \nabla_{\mathbf{xy}}^2 g_t(\mathbf{z}_t) \mathbf{v}_t, \quad (77)$$

we have:

$$\begin{aligned} \mathbb{E} \|e_{t+1}^f\|^2 &\leq (1 - \eta_{t+1})^2 \mathbb{E} \|e_t^f\|^2 + 4\eta_{t+1}^2 \left(\frac{\sigma_{g_{\mathbf{xy}}}^2}{b} p^2 + \frac{\sigma_{f_{\mathbf{x}}}^2}{b}\right) \\ &\quad + 12p^2 (1 - \eta_{t+1})^2 \mathbb{E} \|\nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{xy}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\ &\quad + 12(1 - \eta_{t+1})^2 \mathbb{E} \|\nabla_{\mathbf{x}} f_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\ &\quad + 72(1 - \eta_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E} \|e_t^g\|^2 + 2\beta_t^2 \mathbb{E} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\ &\quad + 72\ell_{g,1}^2 (1 - \eta_{t+1})^2 \delta_t^2 \mathbb{E} \|e_t^{\mathbf{v}}\|^2 + 72(1 - \eta_{t+1})^2 \ell_{g,1}^4 \delta_t^2 \mathbb{E} [\theta_t^{\mathbf{v}}], \end{aligned} \quad (78)$$

for all $t \in [T]$, $\theta_t^{\mathbf{y}}$, $e_t^{\mathbf{y}}$ and e_t^g are defined in (49), (60) and (45), respectively.

Proof. Note that

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

where

$$\tilde{\mathbf{d}}_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{v}_{t+1}) = \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) + \nabla_{\mathbf{xy}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) \mathbf{v}_{t+1}. \quad (79)$$

From Algorithm 1, we have

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

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

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

which implies that

$$\begin{aligned} &\mathbb{E}\|e_{t+1}^f\|^2 \\ &= (1 - \eta_{t+1})^2 \mathbb{E}\|e_t^f\|^2 + \mathbb{E}\|\eta_{t+1}(\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1}; \mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})) \\ &\quad - (1 - \eta_{t+1})(\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_t; \mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1}; \mathcal{B}_{t+1}) + \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1}) - \tilde{\mathbf{d}}_t(\mathbf{u}_t))\|^2 \\ &\leq (1 - \eta_{t+1})^2 \mathbb{E}\|e_t^f\|^2 + 2\eta_{t+1}^2 \mathbb{E}\|\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1}; \mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})\|^2 \\ &\quad + 2(1 - \eta_{t+1})^2 \mathbb{E}\|\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1}; \mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_t; \mathcal{B}_{t+1}) + \tilde{\mathbf{d}}_t(\mathbf{u}_t)\|^2, \end{aligned} \quad (80)$$

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 (80). Based on (79), we have

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

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 (80) and using $\|a + b + c\|^2 \leq 3(\|a\|^2 + \|b\|^2 + \|c\|^2)$, we obtain

$$\begin{aligned} \mathbb{E}\|e_{t+1}^f\|^2 &\leq (1 - \eta_{t+1})^2 \mathbb{E}\|e_t^f\|^2 + 4\lambda_{t+1}^2 \left(\frac{\sigma_{g_{\mathbf{xy}}}^2}{b} p^2 + \frac{\sigma_{f_{\mathbf{x}}}^2}{b}\right) \\ &\quad + 6(1 - \eta_{t+1})^2 \mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{u}_t) - \tilde{\mathbf{d}}_t(\mathbf{u}_{t+1})\|^2 \\ &\quad + 6(1 - \eta_{t+1})^2 \mathbb{E}\|\tilde{\mathbf{d}}_t(\mathbf{u}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1})\|^2 \\ &\quad + 6(1 - \eta_{t+1})^2 \mathbb{E}\|\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1}; \mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_t; \mathcal{B}_{t+1})\|^2. \end{aligned} \quad (81)$$

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 \\
& \stackrel{(i)}{\leq} 3\mathbb{E}\|(\nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_{t+1}))\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_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_{t+1}) - \nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t))\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 \\
& \stackrel{(ii)}{\leq} 6(\ell_{g,2}^2 \mathbb{E}\|\mathbf{v}_{t+1}\|^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2) + 3\ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 \\
& \stackrel{(iii)}{\leq} 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \beta_t^2 \mathbb{E}\|\mathbf{d}_t^y\|^2) + 3\ell_{g,1}^2 \delta_t^2 \mathbb{E}\|\mathbf{d}_t^y\|^2 \\
& \stackrel{(iv)}{\leq} 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\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) \\
& + 6\ell_{g,1}^2 \delta_t^2 (\mathbb{E}\|e_t^y\|^2 + \mathbb{E}\|\nabla P_t(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2) \\
& \stackrel{(vi)}{\leq} 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\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) \\
& + 6\ell_{g,1}^2 \delta_t^2 (\mathbb{E}\|e_t^y\|^2 + \ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2), \tag{82}
\end{aligned}$$

where the (i) follows from (79); (ii) follows from Assumptions B2., B3. and B4.; (iii) follows from (8); (iv) follows from (45) and (60); (vi) follows from (59).

Similarly, we have

$$\begin{aligned}
& \mathbb{E}\|\tilde{\mathbf{d}}_{t+1}(\mathbf{u}_{t+1}; \mathcal{B}_{t+1}) - \tilde{\mathbf{d}}_{t+1}(\mathbf{u}_t; \mathcal{B}_{t+1})\|^2 \\
& \leq 6(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E}\|e_t^g\|^2 + 2\beta_t^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
& + 6\ell_{g,1}^2 \delta_t^2 (\mathbb{E}\|e_t^y\|^2 + \ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_t - \mathbf{v}_t^*(\mathbf{x}_t)\|^2). \tag{83}
\end{aligned}$$

Substituting (83) and (82) into (81), we have

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

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

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

This completes the proof. $\square$

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 objective; (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 $\mathcal{O}(\delta_t^2 \mathbb{E}\|e_t^y\|^2 + 72(1 - \eta_{t+1})^2 \ell_{g,1}^4 \delta_t^2 \mathbb{E}[\theta_t^y])$, which is introduced by the one-step momentum update in solving the linear system problem.

C.5 Bounds on the Outer Objective and its Projected Gradient

Lemma C.10. *Let Assumption 2.4 holds. Then, for the sequence of functions $\{f_t\}_{t=1}^T$, we have*

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

where M is defined in Assumption 2.4; V_T is defined in (11).

Proof. Note that, we have

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

where the inequality follows from Assumption 2.4. $\square$

Lemma C.11. *Let $\{f_t\}_{t=1}^T$ denote the sequence of functions presented to Algorithm 1, satisfying Assumptions 2.2, 2.3 and 2.4. Let $\mathcal{P}_{\mathcal{X}, \alpha_t}$ be defined as in Definition B.1. For any positive step size α_t such that $\alpha_t \leq 1/4L_f$ for all $t \in [T]$, Algorithm 1 ensures the following bound:*

$$\begin{aligned} & \sum_{t=1}^T (\alpha_t - L_f \alpha_t^2) \mathbb{E} \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))\|^2 \\ &\leq 8M + 4V_T + 2M_f^2 \sum_{t=1}^T (2\alpha_t - L_f \alpha_t^2) (\mathbb{E}[\theta_t^y] + \mathbb{E}[\theta_t^x]) \\ &+ 2 \sum_{t=1}^T (2\alpha_t - L_f \alpha_t^2) \mathbb{E} \|e_t^f\|^2. \end{aligned} \tag{84}$$

Here, θ_t^y and θ_t^x are defined in (49); V_T , M , M_f and e_t^f are defined in (11), Assumption 2.4, Eq. (42), and (77).

Proof. It follows from Lemma C.1 that

$$\begin{aligned} & f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1})) - f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \\ &\leq \langle \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L_f}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= -\alpha_t \langle \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)), \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^x) \rangle + \frac{L_f \alpha_t^2}{2} \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^x)\|^2. \end{aligned} \tag{85}$$

For the first term on the right hand side of (85), we have that

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

where the inequality follows from Lemma B.7.

Let $\tilde{\mathbf{d}}_t(\mathbf{z}_t, \mathbf{v}_t) = \nabla_{\mathbf{x}} f_t(\mathbf{z}_t) + \nabla_{\mathbf{y}}^2 g_t(\mathbf{z}_t) \mathbf{v}_t$. Then, from Lemma C.1, we have

$$\begin{aligned} \|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\|^2 &= \left\| \mathbf{d}_t^{\mathbf{x}} - \tilde{\mathbf{d}}_t(\mathbf{z}_t, \mathbf{v}_t) + \tilde{\mathbf{d}}_t(\mathbf{z}_t, \mathbf{v}_t) - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \right\|^2 \\ &\leq 2 \left\| \mathbf{d}_t^{\mathbf{x}} - \tilde{\mathbf{d}}_t(\mathbf{z}_t, \mathbf{v}_t) \right\|^2 + 2 \left\| \tilde{\mathbf{d}}_t(\mathbf{z}_t, \mathbf{v}_t) - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \right\|^2 \\ &\leq 2 \left\| e_t^f \right\|^2 + 2 \left\| \tilde{\mathbf{d}}_t(\mathbf{z}_t, \mathbf{v}_t) - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \right\|^2 \\ &\leq 2 \left\| e_t^f \right\|^2 + M_f^2 (\theta_t^{\mathbf{y}} + \theta_t^{\mathbf{v}}), \end{aligned} \quad (86)$$

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

$$\begin{aligned} & - \langle \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)), \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}}) \rangle \\ & \leq -\frac{1}{2} \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}})\|^2 + 2 \left\| e_t^f \right\|^2 + M_f^2 (\theta_t^{\mathbf{y}} + \theta_t^{\mathbf{v}}). \end{aligned} \quad (87)$$

Plugging the bound (87) into (85), we have that

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

which can be rearranged into

$$\begin{aligned} & (\alpha_t - L_f \alpha_t^2) \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}})\|^2 \\ & \leq 2f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1})) + 4\alpha_t \left\| e_t^f \right\|^2 + 2M_f^2 (\theta_t^{\mathbf{y}} + \theta_t^{\mathbf{v}}) \alpha_t. \end{aligned} \quad (88)$$

In addition, we have

$$\begin{aligned} & \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))\|^2 \\ & \leq 2 \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}}) - \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))\|^2 + 2 \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}})\|^2 \\ & \leq 2 \|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\|^2 + 2 \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}})\|^2 \\ & \leq 4 \left\| e_t^f \right\|^2 + 4M_f^2 (\theta_t^{\mathbf{y}} + \theta_t^{\mathbf{v}}) + 4 \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}})\|^2, \end{aligned} \quad (89)$$

where the second inequality follows from non-expansiveness of the projection operator and the last inequality follows from (86).

Combining (88) and (89), we have

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

where the second inequality is due to Lemma C.10. $\square$

Lemma C.12. Let Assumptions 2.2, and 2.3 hold. Let $\{\mathbf{x}_t\}_{t=1}^T$ be generated according to Algorithm 1. Then, we have

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

where $\theta_t^{\mathbf{y}}$ and $\theta_t^{\mathbf{v}}$ are defined in (49), M_f is defined in (42).

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

$$\begin{aligned}
\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 &= \alpha_t^2 \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}})\|^2 \\
&\leq 2\alpha_t^2 \left(\|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))\|^2 \right. \\
&\quad \left. + \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \mathbf{d}_t^{\mathbf{x}}) - \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))\|^2 \right) \\
&\leq 2\alpha_t^2 \left(\|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))\|^2 \right. \\
&\quad \left. + \|\mathbf{d}_t^{\mathbf{x}} - \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))\|^2 \right) \\
&\leq 2\alpha_t^2 \left(\|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))\|^2 + M_f^2 (\theta_t^{\mathbf{y}} + \theta_t^{\mathbf{v}}) \right), \tag{90}
\end{aligned}$$

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. (39a) in Lemma C.1. $\square$

C.6 Proof of Theorem 2.6

Proof. **Bounding $\mathbb{E}\|e_t^f\|^2$ in (78).** From (78), we have

$$\begin{aligned}
\frac{\mathbb{E}\|e_{t+1}^f\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^f\|^2}{\alpha_{t-1}} &\leq \left(\frac{(1-\eta_{t+1})^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \right) \mathbb{E}\|e_t^f\|^2 + \frac{4\eta_{t+1}^2}{\alpha_t} \left(\frac{\sigma_{g_{\mathbf{x}\mathbf{y}}}^2}{b} p^2 + \frac{\sigma_{f_{\mathbf{x}}}^2}{b} \right) \\
&+ \frac{12p^2}{\alpha_t} (1-\eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}\mathbf{y}}^2 g_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{x}\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
&+ \frac{12}{\alpha_t} (1-\eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
&+ \frac{72}{\alpha_t} (1-\eta_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E}\|e_t^g\|^2 + 2\beta_t^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
&+ \frac{72}{\alpha_t} \ell_{g,1}^2 (1-\eta_{t+1})^2 \delta_t^2 \mathbb{E}\|e_t^{\mathbf{v}}\|^2 + \frac{72}{\alpha_t} (1-\eta_{t+1})^2 \ell_{g,1}^4 \delta_t^2 \mathbb{E}[\theta_t^{\mathbf{v}}]. \tag{91}
\end{aligned}$$

With respect to the coefficient of the first term on the right-hand side of Eq. (91), it is important to note that we have:

$$\frac{(1-\eta_{t+1})^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{1}{\alpha_t} - \frac{\eta_{t+1}}{\alpha_t} - \frac{1}{\alpha_{t-1}}. \tag{92}$$

Using the definition of α_t in (16), we have

$$\begin{aligned}
\frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} &= (c+t)^{1/3} - (c+t-1)^{1/3} \stackrel{(i)}{\leq} \frac{1}{3(c+t-1)^{2/3}} \stackrel{(ii)}{\leq} \frac{1}{3(\frac{c}{2}+t)^{2/3}} \\
&= \frac{2^{2/3}}{3(c+2t)^{2/3}} \stackrel{(iii)}{\leq} \frac{2^{2/3}}{3(c+t)^{2/3}} \stackrel{(iv)}{\leq} \frac{2^{2/3}}{3} \alpha_t^2 \stackrel{(vi)}{\leq} \frac{\alpha_t}{6L_f}, \tag{93}
\end{aligned}$$

where the (i) follows from $(a+b)^{1/3} - a^{1/3} \leq b/(3a^{2/3})$; (ii) follows from $c \geq 2$ in (107); (iii) follows from (16); (iv) follows from $\alpha_t \leq 1/4L_f$ in (107).

Substituting (93) into (92) and using $\delta_t = c_\delta \alpha_t$ and $\eta_{t+1} = c_\eta \alpha_t^2$, we have

$$\frac{(1-\eta_{t+1})^2}{\alpha_t} - \frac{1}{\alpha_{t-1}} \leq \frac{\alpha_t}{6L_f} - \frac{\eta_{t+1}}{\alpha_t} = \frac{\alpha_t}{6L_f} - c_\eta \alpha_t \leq -5\Omega \alpha_t, \tag{94}$$

where the inequalities follow from $c_\eta = \frac{1}{6L_f} + 5\Omega$ in (106).

Then, substituting (94) into (91) yields

$$\begin{aligned}
& \frac{1}{\Omega} \mathbb{E} \left(\frac{\|e_{t+1}^f\|^2}{\alpha_t} - \frac{\|e_t^f\|^2}{\alpha_{t-1}} \right) \leq -5\alpha_t \mathbb{E} \|e_t^f\|^2 + \frac{4\eta_{t+1}^2}{\Omega\alpha_t} \left(\frac{\sigma_{g_{xy}}^2}{b} p^2 + \frac{\sigma_{f_x}^2}{b} \right) \\
& + \frac{12p^2}{\Omega\alpha_t} (1 - \eta_{t+1})^2 \mathbb{E} \|\nabla_{\mathbf{xy}}^2 g_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{xy}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
& + \frac{12}{\Omega\alpha_t} (1 - \eta_{t+1})^2 \mathbb{E} \|\nabla_{\mathbf{x}} f_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
& + \frac{72}{\Omega\alpha_t} (1 - \eta_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E} \|e_t^g\|^2 + 2\beta_t^2 \mathbb{E} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
& + \frac{72}{\Omega\alpha_t} \ell_{g,1}^2 (1 - \eta_{t+1})^2 \delta_t^2 \mathbb{E} \|e_t^v\|^2 + \frac{72}{\Omega\alpha_t} (1 - \eta_{t+1})^2 \ell_{g,1}^4 \delta_t^2 \mathbb{E} [\theta_t^v]. \tag{95}
\end{aligned}$$

Bounding $\mathbb{E} \|e_t^g\|^2$ in (46) .

From (46), we have

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

Let us examine the coefficient of the first term on the right-hand side of Eq. (96). Specifically, for $\gamma_{t+1} = c_\gamma \alpha_t^2$ and $\beta_t = c_\beta \alpha_t$, we have:

$$\begin{aligned}
\frac{1}{\alpha_t} (1 - \gamma_{t+1})^2 (1 + 48\ell_{g,1}^2 \beta_t^2) - \frac{1}{\alpha_{t-1}} & \leq \frac{1}{\alpha_t} (1 - \gamma_{t+1}) (1 + 48\ell_{g,1}^2 \beta_t^2) - \frac{1}{\alpha_{t-1}} \\
& = \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - \frac{\gamma_{t+1}}{\alpha_t} + \frac{1 - \gamma_{t+1}}{\alpha_t} 48\ell_{g,1}^2 \beta_t^2 \\
& = \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - c_\gamma \alpha_t + \left(\frac{1}{\alpha_t} - c_\gamma \alpha_t \right) 48\ell_{g,1}^2 c_\beta^2 \alpha_t^2 \\
& \leq \frac{\alpha_t}{6L_f} + 48\ell_{g,1}^2 c_\beta^2 \alpha_t - c_\gamma \alpha_t, \tag{97}
\end{aligned}$$

where the last inequality follows from (93).

From the selected c_γ in (107) and the definition of Φ in (106), we have

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

Combined this with Eq. (97) yields

$$\frac{1}{\alpha_t} (1 - \gamma_{t+1})^2 (1 + 48\ell_{g,1}^2 \beta_t^2) - \frac{1}{\alpha_{t-1}} \leq -\hbar\Phi\alpha_t. \tag{98}$$

Substituting Eq. (98) into Eq. (96) yields

$$\begin{aligned}
\frac{1}{\Phi} \left(\frac{\mathbb{E} \|e_{t+1}^g\|^2}{\alpha_t} - \frac{\mathbb{E} \|e_t^g\|^2}{\alpha_{t-1}} \right) & \leq -\hbar\alpha_t \mathbb{E} \|e_t^g\|^2 \\
& + 2 \frac{\gamma_{t+1}^2}{\Phi\alpha_t} \frac{\sigma_{g_y}^2}{b} + \frac{24}{\Phi\alpha_t} (1 - \gamma_{t+1})^2 \ell_{g,1}^2 \mathbb{E} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
& + \frac{6}{\Phi\alpha_t} (1 - \gamma_{t+1})^2 \mathbb{E} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
& + 48(1 - \gamma_{t+1})^2 \ell_{g,1}^2 \frac{\beta_t^2}{\Phi\alpha_t} \mathbb{E} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2. \tag{99}
\end{aligned}$$

Bounding $\mathbb{E}\|e_t^{\mathbf{v}}\|^2$ in (61) .

From (61), we get

$$\begin{aligned}
\frac{\mathbb{E}\|e_{t+1}^{\mathbf{v}}\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^{\mathbf{v}}\|^2}{\alpha_{t-1}} &\leq \left(\frac{1}{\alpha_t}(1 - \lambda_{t+1})^2(1 + 72\ell_{g,1}^2\delta_t^2) - \frac{1}{\alpha_{t-1}} \right) \mathbb{E}\|e_t^{\mathbf{v}}\|^2 \\
&+ 4\frac{\lambda_{t+1}^2}{\alpha_t} \left(\frac{\sigma_{g\mathbf{y}\mathbf{y}}^2}{b} p^2 + \frac{\sigma_{f\mathbf{y}}^2}{b} \right) + \frac{12p^2}{\alpha_t} (1 - \lambda_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
&+ \frac{12}{\alpha_t} (1 - \lambda_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
&+ \frac{72}{\alpha_t} (1 - \lambda_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E}\|e_t^g\|^2 + 2\beta_t^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
&+ \frac{144}{\alpha_t} (1 - \lambda_{t+1})^2 \ell_{g,1}^4 \delta_t^2 \mathbb{E}[\theta_t^{\mathbf{y}}] + \frac{288}{\alpha_t} \ell_{g,1}^2 (p^2 \ell_{g,2}^2 + \ell_{f,1}^2) \delta_t^2 \mathbb{E}[\theta_t^{\mathbf{y}}]. \tag{100}
\end{aligned}$$

Let us examine the coefficient of the first term on the right-hand side of equation (100). Specifically, for $\lambda_{t+1} = c_\lambda \alpha_t^2$ and $\delta_t = c_\delta \alpha_t$, we have:

$$\begin{aligned}
\frac{1}{\alpha_t}(1 - \lambda_{t+1})^2(1 + 72\ell_{g,1}^2\delta_t^2) - \frac{1}{\alpha_{t-1}} &\leq \frac{1}{\alpha_t}(1 - \lambda_{t+1})(1 + 72\ell_{g,1}^2\delta_t^2) - \frac{1}{\alpha_{t-1}} \\
&= \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - \frac{\lambda_{t+1}}{\alpha_t} + \frac{1 - \lambda_{t+1}}{\alpha_t} 72\ell_{g,1}^2\delta_t^2 \\
&= \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}} - c_\lambda \alpha_t + \left(\frac{1}{\alpha_t} - c_\lambda \alpha_t \right) 72\ell_{g,1}^2 c_\delta^2 \alpha_t^2 \\
&\leq \frac{\alpha_t}{6L_f} + 72\ell_{g,1}^2 c_\delta^2 \alpha_t - c_\lambda \alpha_t, \tag{101}
\end{aligned}$$

where the last inequality follows from (93).

From the selected c_γ in (107) and the definition of Ψ in (106), we have

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

Combined this with Eq. (101) yields

$$\frac{1}{\alpha_t}(1 - \lambda_{t+1})^2(1 + 72\ell_{g,1}^2\delta_t^2) - \frac{1}{\alpha_{t-1}} \leq -j\Psi\alpha_t. \tag{102}$$

Substituting Eq. (102) into Eq. (100) yields

$$\begin{aligned}
\frac{1}{\Psi} \left(\frac{\mathbb{E}\|e_{t+1}^{\mathbf{v}}\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^{\mathbf{v}}\|^2}{\alpha_{t-1}} \right) &\leq -j\alpha_t \mathbb{E}\|e_t^{\mathbf{v}}\|^2 \\
&+ 4\frac{\lambda_{t+1}^2}{\Psi\alpha_t} \left(\frac{\sigma_{g\mathbf{y}\mathbf{y}}^2}{b} p^2 + \frac{\sigma_{f\mathbf{y}}^2}{b} \right) + \frac{12p^2}{\Psi\alpha_t} (1 - \lambda_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}}^2 g_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
&+ \frac{12}{\Psi\alpha_t} (1 - \lambda_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1})\|^2 \\
&+ \frac{72}{\Psi\alpha_t} (1 - \lambda_{t+1})^2 (\ell_{g,2}^2 p^2 + \ell_{f,1}^2) (\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2\beta_t^2 \mathbb{E}\|e_t^g\|^2 + 2\beta_t^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2) \\
&+ \frac{72}{\Psi\alpha_t} (1 - \lambda_{t+1})^2 \ell_{g,1}^4 \delta_t^2 \mathbb{E}[\theta_t^{\mathbf{y}}] + \frac{288}{\Psi\alpha_t} \ell_{g,1}^2 (p^2 \ell_{g,2}^2 + \ell_{f,1}^2) \delta_t^2 \mathbb{E}[\theta_t^{\mathbf{y}}]. \tag{103}
\end{aligned}$$

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

$$\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \leq 2\alpha_t^2 \left(\|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))\|^2 + M_f^2 (\theta_t^{\mathbf{y}} + \theta_t^{\mathbf{v}}) \right). \tag{104}$$

Let

$$\begin{aligned}
\Lambda &:= \Gamma \sum_{t=1}^T (\mathbb{E}[\theta_{t+1}^{\mathbf{y}}] - \mathbb{E}[\theta_t^{\mathbf{y}}]) + \Upsilon \sum_{t=1}^T (\mathbb{E}[\theta_{t+1}^{\mathbf{v}}] - \mathbb{E}[\theta_t^{\mathbf{v}}]) + \frac{1}{\Phi} \sum_{t=1}^T \left(\frac{\mathbb{E}\|e_{t+1}^g\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^g\|^2}{\alpha_{t-1}} \right) \\
&+ \frac{1}{\Psi} \sum_{t=1}^T \left(\frac{\mathbb{E}\|e_{t+1}^{\mathbf{v}}\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^{\mathbf{v}}\|^2}{\alpha_{t-1}} \right) + \frac{1}{\Omega} \sum_{t=1}^T \left(\frac{\mathbb{E}\|e_{t+1}^f\|^2}{\alpha_t} - \frac{\mathbb{E}\|e_t^f\|^2}{\alpha_{t-1}} \right). \tag{105}
\end{aligned}$$

Here

$$\begin{aligned}
\Gamma &= \frac{11M_f^2}{L_{\mu_g}c_\beta}, \quad \Upsilon = \frac{22M_f^2}{\dot{L}_{\mu_g}c_\delta}, \quad \Phi \geq \max \left\{ 480\ell_{g,1}^2, 192\ell_{g,1}^2 \frac{(\mu_g + \ell_{g,1})}{\Gamma} c_\beta \right\}, \\
\Psi &= \max \left\{ 144(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) \left(10 + \frac{L_{\mu_g}^2 c_\beta^2}{M_f^2} \right), \frac{288\ell_{g,1}^4}{M_f^2} c_\delta^2, \right. \\
&\quad \left. 576(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) \frac{(\mu_g + \ell_{g,1})}{\Gamma} c_\beta, \frac{576}{M_f^2} \ell_{g,1}^2 (p^2 \ell_{g,2}^2 + \ell_{f,1}^2) c_\delta^2 \right\}, \\
\Omega &= \max \left\{ 144(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) \left(10 + \frac{L_{\mu_g}^2 c_\beta^2}{M_f^2} \right), \frac{288\ell_{g,1}^4}{M_f^2} c_\delta^2, \right. \\
&\quad \left. 576(\ell_{g,2}^2 p^2 + \ell_{f,1}^2) \frac{(\mu_g + \ell_{g,1})}{\Gamma} c_\beta, \frac{72\ell_{g,1}^2 \dot{L}_{\mu_g}^2}{M_f^2} c_\delta^2 \right\}, \tag{106}
\end{aligned}$$

where $L_{\mu_g} = \mu_g \ell_{g,1} / (\mu_g + \ell_{g,1})$ and $\dot{L}_{\mu_g} = (\ell_{g,1} + \ell_{g,1}^3) \mu_g / (\mu_g + \ell_{g,1})$. Here, we have

$$\begin{aligned}
c &\geq \max \{ 4L_f, c_\beta(\mu_g + \ell_{g,1}), 2 \}, \\
c_\beta &= \sqrt{880} \frac{L_{\mathbf{y}} M_f}{L_{\mu_g}}, \\
c_\delta &= \sqrt{3520(1 + 2L_{\mathbf{y}}^2)} \frac{\nu M_f}{\dot{L}_{\mu_g} \mu_g}, \quad \text{where } \nu = \ell_{f,1} + \frac{\ell_{g,2} \ell_{f,0}}{\mu_g}, \quad L_{\mathbf{y}} = \frac{\ell_{g,1}}{\mu_g}, \\
c_\gamma &= \frac{1}{6L_f} + 48\ell_{g,1}^2 c_\beta^2 + \hbar \Phi, \quad \text{where } \hbar := 25 \frac{M_f^2}{L_{\mu_g}^2}, \\
c_\eta &= \frac{1}{6L_f} + 5\Omega, \\
c_\lambda &= \frac{1}{6L_f} + 72\ell_{g,1}^2 c_\delta^2 + j\Psi, \quad \text{where } j = 90 \frac{M_f^2}{\dot{L}_{\mu_g}^2}. \tag{107}
\end{aligned}$$

Using (103), (99), (95), (84), (70), and (50), along with (104) and the fact that α_t decreases with respect to t , we obtain:

$$\begin{aligned}
&\sum_{t=1}^T A(\alpha_t, \beta_t, \delta_t) \mathbb{E} \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))\|^2 + \Lambda \\
&\leq 8M + 4V_T + \sum_{t=1}^T B(\alpha_t, \beta_t, \delta_t) \mathbb{E}[\theta_t^{\mathbf{y}}] + \sum_{t=1}^T C(\alpha_t, \beta_t, \delta_t) \mathbb{E}[\theta_t^{\mathbf{y}}] \tag{108a}
\end{aligned}$$

$$+ \sum_{t=1}^T D(\alpha_t) \mathbb{E} \|e_t^f\|^2 + \sum_{t=1}^T F(\alpha_t, \beta_t) \mathbb{E} \|e_t^g\|^2 + \sum_{t=1}^T I(\alpha_t, \delta_t) \mathbb{E} \|e_t^{\mathbf{v}}\|^2 \tag{108b}$$

$$+ \sum_{t=1}^T L(\alpha_t, \beta_t) \mathbb{E} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \sum_{t=2}^T N(\beta_t, \delta_t) \sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2 \tag{108c}$$

$$+ \frac{\sigma_{g_{\mathbf{y}}}^2}{b} \frac{2}{\Phi} \sum_{t=1}^T \frac{\gamma_{t+1}^2}{\alpha_t} + \frac{4}{\Psi} \left(\frac{\sigma_{g_{\mathbf{y}\mathbf{y}}}^2}{b} p^2 + \frac{\sigma_{f_{\mathbf{y}}}^2}{b} \right) \sum_{t=1}^T \frac{\lambda_{t+1}^2}{\alpha_t} + \frac{4}{\Omega} \left(\frac{\sigma_{g_{\mathbf{x}\mathbf{y}}}^2}{b} p^2 + \frac{\sigma_{f_{\mathbf{x}}}^2}{b} \right) \sum_{t=1}^T \frac{\eta_{t+1}^2}{\alpha_t} \tag{108d}$$

$$+ \frac{6}{\Phi \alpha_T} G_{\mathbf{y}, T} + \frac{12p^2}{\Omega \alpha_T} G_{\mathbf{x}\mathbf{y}, T} + \frac{12p^2}{\Psi \alpha_T} G_{\mathbf{y}\mathbf{y}, T} + \frac{12\ell_{f,1}^2}{\Psi \alpha_T} D_{\mathbf{y}, T} + \frac{12\ell_{f,1}^2}{\Omega \alpha_T} D_{\mathbf{x}, T}. \tag{108e}$$

Here, M is defined in Assumption 2.4, V_T and $H_{2,T}$ are defined in (11). Moreover, $G_{\mathbf{y},T}$, $G_{\mathbf{xy},T}$, and $G_{\mathbf{yy},T}$ are defined in (13). Let

$$\begin{aligned}
E(\alpha_t, \beta_t, \delta_t) &:= \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}\beta_t}\Gamma + \frac{8\nu^2}{\dot{L}_{\mu_g}\mu_g^2\delta_t}(1 + 2L_{\mathbf{y}}^2)\Upsilon + 72(1 - \eta_{t+1})^2(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\frac{1}{\Omega\alpha_t} \\
&\quad + 24(1 - \gamma_{t+1})^2\ell_{g,1}^2\frac{1}{\Phi\alpha_t} + 72(1 - \lambda_{t+1})^2(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\frac{1}{\Psi\alpha_t}, \\
A(\alpha_t, \beta_t, \delta_t) &:= \alpha_t - (L_f + 2E(\alpha_t, \beta_t, \delta_t))\alpha_t^2, \\
B(\alpha_t, \beta_t, \delta_t) &:= -\frac{\dot{L}_{\mu_g}\Upsilon}{4}\delta_t + 4M_f^2\alpha_t - 2M_f^2L_f\alpha_t^2 + 2M_f^2E(\alpha_t, \beta_t, \delta_t)\alpha_t^2 \\
&\quad + 72(1 - \lambda_{t+1})^2\ell_{g,1}^4\delta_t^2\frac{1}{\Psi\alpha_t} + 72(1 - \eta_{t+1})^2\ell_{g,1}^4\delta_t^2\frac{1}{\Omega\alpha_t}, \\
C(\alpha_t, \beta_t, \delta_t) &:= -\frac{L_{\mu_g}\Gamma}{2}\beta_t + 4M_f^2\alpha_t - 2L_fM_f^2\alpha_t^2 + 2M_f^2E(\alpha_t, \beta_t, \delta_t)\alpha_t^2 \\
&\quad + \frac{288}{\Psi\alpha_t}\ell_{g,1}^2(p^2\ell_{g,2}^2 + \ell_{f,1}^2)\delta_t^2, \\
D(\alpha_t) &:= 2(2\alpha_t - L_f\alpha_t^2) - 5\alpha_t, \\
F(\alpha_t, \beta_t) &:= \frac{2\Gamma}{L_{\mu_g}}\beta_t - \hbar\alpha_t + 144(1 - \lambda_{t+1})^2(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\frac{\beta_t^2}{\Psi\alpha_t} \\
&\quad + 144(1 - \eta_{t+1})^2(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\frac{\beta_t^2}{\Omega\alpha_t}, \\
I(\alpha_t, \delta_t) &:= \frac{4\Upsilon}{\dot{L}_{\mu_g}}\delta_t - \jmath\alpha_t + 72\ell_{g,1}^2(1 - \eta_{t+1})^2\frac{\delta_t^2}{\Omega\alpha_t}.
\end{aligned} \tag{109}$$

Moreover, we have

$$\begin{aligned}
L(\alpha_t, \beta_t) &:= -\frac{2\Gamma}{\mu_g + \ell_{g,1}}\beta_t + \Gamma\beta_t^2 + 48(1 - \gamma_{t+1})^2\ell_{g,1}^2\frac{\beta_t^2}{\Phi\alpha_t} \\
&\quad + 144(1 - \lambda_{t+1})^2(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\frac{\beta_t^2}{\Psi\alpha_t} + 144(1 - \eta_{t+1})^2(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\frac{\beta_t^2}{\Omega\alpha_t}, \\
N(\beta_t, \delta_t) &:= \frac{4}{L_{\mu_g}\beta_t}\Gamma + \frac{16\nu^2}{\dot{L}_{\mu_g}\mu_g^2\delta_t}\Upsilon.
\end{aligned} \tag{110}$$

Note that, we have

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

which together with $\beta_t = c_\beta\alpha_t$ and $\delta_t = c_\delta\alpha_t$ in Eq. (16), we have

$$\begin{aligned}
\alpha_t^2 E(\alpha_t, \beta_t, \delta_t) &= \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}}\Gamma\frac{\alpha_t^2}{\beta_t} + \frac{8\nu^2}{\dot{L}_{\mu_g}\mu_g^2}(1 + 2L_{\mathbf{y}}^2)\Upsilon\frac{\alpha_t^2}{\delta_t} + 24(1 - \gamma_{t+1})^2\ell_{g,1}^2\frac{\alpha_t}{\Phi} \\
&\quad + 72(1 - \eta_{t+1})^2(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\frac{\alpha_t}{\Omega} + 72(1 - \lambda_{t+1})^2(\ell_{g,2}^2p^2 + \ell_{f,1}^2)\frac{\alpha_t}{\Psi} \\
&\leq \frac{44L_{\mathbf{y}}^2}{L_{\mu_g}^2}M_f^2\frac{\alpha_t}{c_\beta^2} + \frac{176\nu^2}{\dot{L}_{\mu_g}^2\mu_g^2}(1 + 2L_{\mathbf{y}}^2)M_f^2\frac{\alpha_t}{c_\delta^2} \\
&\quad + 24\ell_{g,1}^2\frac{\alpha_t}{\Phi} + 72(\ell_{g,2}^2p^2 + \ell_{f,1}^2)(\frac{1}{\Omega} + \frac{1}{\Psi})\alpha_t \\
&\leq \frac{\alpha_t}{4},
\end{aligned} \tag{111}$$

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 (106); the last inequality follows from $c_\beta = \sqrt{880\frac{L_y^2M_f^2}{L_{\mu_g}^2}}$, $c_\delta = \sqrt{3520\frac{\nu^2M_f^2}{L_{\mu_g}^2\mu_g^2}(1+2L_y^2)}$, in (107) and $\Phi \geq 480\ell_{g,1}^2$, and $\Omega, \Psi \geq 1440(\ell_{g,2}^2p^2 + \ell_{f,1}^2)$ in (106). Moreover, we have

$$\begin{aligned} A(\alpha_t, \beta_t, \delta_t) &= \alpha_t - L_f\alpha_t^2 - 2E(\alpha_t, \beta_t, \delta_t)\alpha_t^2 \\ &\geq \alpha_t - L_f\alpha_t^2 - \frac{\alpha_t}{2} \\ &\geq \frac{\alpha_t}{4}, \end{aligned} \tag{112}$$

where the last inequality follows from $\alpha_t \leq 1/4L_f$ in (107) since $\alpha_t = 1/(c+t)^{1/3}$ in (16).

Bounding (108a).

From (109), we have

$$\begin{aligned} B(\alpha_t, \beta_t, \delta_t) &= -\frac{\dot{L}_{\mu_g}\Upsilon}{4}\delta_t + 4M_f^2\alpha_t - 2M_f^2L_f\alpha_t^2 + 2M_f^2E(\alpha_t, \beta_t, \delta_t)\alpha_t^2 \\ &\quad + 72(1-\lambda_{t+1})^2\ell_{g,1}^4\delta_t^2\frac{1}{\Psi\alpha_t} + 72(1-\eta_{t+1})^2\ell_{g,1}^4\delta_t^2\frac{1}{\Omega\alpha_t} \\ &\leq -\frac{\dot{L}_{\mu_g}\Upsilon}{4}\delta_t + 4M_f^2\alpha_t - 2M_f^2L_f\alpha_t^2 + \frac{M_f^2}{2}\alpha_t + 72\ell_{g,1}^4\left(\frac{1}{\Psi} + \frac{1}{\Omega}\right)\frac{\delta_t^2}{\alpha_t} \\ &= \left(-\frac{\dot{L}_{\mu_g}}{4}\Upsilon c_\delta + \frac{9}{2}M_f^2 + 72\ell_{g,1}^4\left(\frac{1}{\Psi} + \frac{1}{\Omega}\right)c_\delta^2\right)\alpha_t \\ &\leq -\frac{1}{2}M_f^2\alpha_t, \end{aligned} \tag{113}$$

where the first inequality follows from $\beta_t = c_\beta\alpha_t$, $\delta_t = c_\delta\alpha_t$ in (16), and Eq. (111); 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 (106); the last inequality follows from in (106). Moreover, using Eq. (109) together with $\beta_t = c_\beta\alpha_t$ and $\delta_t = c_\delta\alpha_t$ in Eq. (16), we have

$$\begin{aligned} C(\alpha_t, \beta_t, \delta_t) &= -\frac{L_{\mu_g}\Gamma}{2}\beta_t + 4M_f^2\alpha_t - 2L_fM_f^2\alpha_t^2 + 2M_f^2E(\alpha_t, \beta_t, \delta_t)\alpha_t^2 \\ &\quad + \frac{288}{\Psi\alpha_t}\ell_{g,1}^2(p^2\ell_{g,2}^2 + \ell_{f,1}^2)\delta_t^2 \\ &\leq -\frac{L_{\mu_g}}{2}\Gamma c_\beta\alpha_t + \frac{9}{2}M_f^2\alpha_t + \frac{288}{\Psi}\ell_{g,1}^2(p^2\ell_{g,2}^2 + \ell_{f,1}^2)c_\delta^2\alpha_t \\ &\leq -\frac{M_f^2}{2}\alpha_t, \end{aligned} \tag{114}$$

where the first inequality follows from (111); the last inequality follows from $\Gamma = \frac{11M_f^2}{L_{\mu_g}c_\beta}$ and $\Psi \geq \frac{576}{M_f^2}\ell_{g,1}^2(p^2\ell_{g,2}^2 + \ell_{f,1}^2)c_\delta^2$ in (106).

Thus, from (113) and (114), we get

$$(108a) \leq \mathcal{O}(V_T). \tag{115}$$

Bounding (108b).

From (109), we also have

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

From Eq. (109), $\beta_t = c_\beta \alpha_t$, and $\Gamma = \frac{11M_f^2}{L_{\mu_g} c_\beta}$, we obtain

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

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 (106); and the last equality is by $\hbar = 25 \frac{M_f^2}{L_{\mu_g}^2}$.

From $\delta_t = c_\delta \alpha_t$ in (16), we obtain

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

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

Thus, we get

$$(108b) \leq 0. \quad (116)$$

Bounding (108c).

From $\beta_t = c_\beta \alpha_t$ in (16) and Eq. (110), we have

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

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 (106); the last inequality follows from $\alpha_t \leq 1/c_\beta (\mu_g + \ell_{g,1})$ in (107).

From $\beta_t = c_\beta \alpha_t$, $\delta_t = c_\delta \alpha_t$ in (16) and Eq. (110), we obtain

$$N(\beta_t, \delta_t) = \frac{4}{L_{\mu_g} \beta_t} \Gamma + \frac{16\nu^2}{\tilde{L}_{\mu_g} \mu_g^2 \delta_t} \Upsilon = \frac{4}{L_{\mu_g} c_\beta \alpha_t} \Gamma + \frac{16\nu^2}{\tilde{L}_{\mu_g} \mu_g^2 c_\delta \alpha_t} \Upsilon.$$

Thus, we get

$$\begin{aligned} (108c) &= \sum_{t=1}^T L(\alpha_t, \beta_t) \mathbb{E} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \sum_{t=2}^T N(\beta_t, \delta_t) \sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2 \\ &\leq \mathcal{O} \left(\frac{H_{2,T}}{\alpha_T} \right). \end{aligned} \quad (117)$$

Bounding (108d).

From $\eta_{t+1} = c_\eta \alpha_t^2$, $\gamma_{t+1} = c_\gamma \alpha_t^2$, $\lambda_{t+1} = c_\lambda \alpha_t^2$ in Eq. (16), we obtain

$$\begin{aligned} (108d) &= \frac{\sigma_{g_y}^2}{\bar{b}} \frac{2}{\Phi} \sum_{t=1}^T \frac{\gamma_{t+1}^2}{\alpha_t} + \frac{4}{\Psi} \left(\frac{\sigma_{g_{yy}}^2}{\bar{b}} p^2 + \frac{\sigma_{f_y}^2}{\bar{b}} \right) \sum_{t=1}^T \frac{\lambda_{t+1}^2}{\alpha_t} + \frac{4}{\Omega} \left(\frac{\sigma_{g_{xy}}^2}{\bar{b}} p^2 + \frac{\sigma_{f_x}^2}{\bar{b}} \right) \sum_{t=1}^T \frac{\eta_{t+1}^2}{\alpha_t} \\ &\leq \mathcal{O} \left(\left(\frac{\sigma_{g_y}^2}{\bar{b}} + \frac{\sigma_{g_{yy}}^2}{\bar{b}} + \frac{\sigma_{f_y}^2}{\bar{b}} + \frac{\sigma_{g_{xy}}^2}{\bar{b}} + \frac{\sigma_{f_x}^2}{\bar{b}} \right) \sum_{t=1}^T \alpha_t^3 \right). \end{aligned} \quad (118)$$

Bounding (108e).

We also have

$$\begin{aligned} (108e) &= \frac{6}{\Phi \alpha_T} G_{\mathbf{y},T} + \frac{12p^2}{\Omega \alpha_T} G_{\mathbf{xy},T} + \frac{12p^2}{\Psi \alpha_T} G_{\mathbf{yy},T} + \frac{12\ell_{f,1}^2}{\Psi \alpha_T} D_{\mathbf{y},T} + \frac{12\ell_{f,1}^2}{\Omega \alpha_T} D_{\mathbf{x},T} \\ &\leq \mathcal{O} \left(\frac{1}{\alpha_T} (G_{\mathbf{y},T} + G_{\mathbf{xy},T} + G_{\mathbf{yy},T} + D_{\mathbf{y},T} + D_{\mathbf{x},T}) \right). \end{aligned} \quad (119)$$

From Eq. (16), we have $b = \bar{b} = 1$. Moreover, by (10), $\sigma^2 = \sigma_{g_y}^2 + \sigma_{g_{yy}}^2 + \sigma_{f_y}^2 + \sigma_{g_{xy}}^2 + \sigma_{f_x}^2$. From (15), we also have

$$\begin{aligned} G_T &= G_{\mathbf{y},T} + G_{\mathbf{xy},T} + G_{\mathbf{yy},T}, \\ D_T &= D_{\mathbf{y},T} + D_{\mathbf{x},T}. \end{aligned}$$

Then, by inequalities (112), (115), (116), (117), (118), (119), we have

$$\begin{aligned} &\sum_{t=1}^T \frac{\alpha_t}{2} \mathbb{E} \|\mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)))\|^2 + \Lambda \\ &\leq \mathcal{O} \left(V_T + \frac{H_{2,T}}{\alpha_T} + \frac{\sigma^2}{\bar{b}} \sum_{t=1}^T \alpha_t^3 + \frac{G_T}{\alpha_T} + \frac{D_T}{\alpha_T} \right). \end{aligned} \quad (120)$$

From the definition of Λ in (105), we have

$$\begin{aligned} -\Lambda &= \Gamma \sum_{t=1}^T (\mathbb{E}[\theta_t^y] - \mathbb{E}[\theta_{t+1}^y]) + \Upsilon \sum_{t=1}^T (\mathbb{E}[\theta_t^v] - \mathbb{E}[\theta_{t+1}^v]) + \frac{1}{\Phi} \sum_{t=1}^T \left(\frac{\mathbb{E}\|e_t^g\|^2}{\alpha_{t-1}} - \frac{\mathbb{E}\|e_{t+1}^g\|^2}{\alpha_t} \right) \\ &\quad + \frac{1}{\Psi} \sum_{t=1}^T \left(\frac{\mathbb{E}\|e_t^y\|^2}{\alpha_{t-1}} - \frac{\mathbb{E}\|e_{t+1}^y\|^2}{\alpha_t} \right) + \frac{1}{\Omega} \sum_{t=1}^T \left(\frac{\mathbb{E}\|e_t^f\|^2}{\alpha_{t-1}} - \frac{\mathbb{E}\|e_{t+1}^f\|^2}{\alpha_t} \right) \\ &\leq \Gamma \theta_1^y + \Upsilon \theta_1^v + \frac{\sigma_{g_y}^2}{\Phi \alpha_0} + \frac{\sigma_{g_{yy}}^2 + \sigma_{f_y}^2}{\Psi \alpha_0} + \frac{\sigma_{g_{xy}}^2 + \sigma_{f_x}^2}{\Omega \alpha_0}. \end{aligned} \quad (121)$$

Using (121), we get

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

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

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

This completes the proof. $\square$

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

Proof Roadmap. We provide Lemma D.7, which quantifies the error between the approximated direction of the momentum-based gradient estimator, $\hat{\mathbf{d}}_t^{\mathbf{y}}$ and the true direction, $\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t)$, at each iteration. Lemma D.9 assesses the convergence of the iterative solutions $\{\mathbf{y}_t\}_{t=1}^T$, specifically the gap $\mathbb{E}[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2]$, while accounting for the error introduced in Lemma D.7. To establish Lemma D.13, which quantifies the error between the approximated direction of the momentum-based gradient estimator, $\hat{\mathbf{d}}_t^{\mathbf{y}}$, and the true direction, $\nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) + \nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v}_t$, we need to present Lemma D.11. This lemma quantifies the error between $\hat{\mathbf{d}}_t^{\mathbf{y}}$ and $\nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) + \frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t))$. Then, Lemma D.15 captures the error of the system solution of Problem (18), i.e., gap $\mathbb{E}[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2]$, based on these errors. To establish Lemma D.19, which quantifies the error between the approximated direction of the momentum-based hypergradient estimator, $\hat{\mathbf{d}}_t^{\mathbf{x}}$, and the true direction, $\nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) + \nabla_{\mathbf{x}\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v}_t$, it is necessary to introduce Lemma D.17. This lemma quantifies the error between $\hat{\mathbf{d}}_t^{\mathbf{x}}$ and $\nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) + \frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t))$. Then, Lemma D.20 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 3.2

Lemma D.1. [4, Lemma A.1.] Suppose Assumption B4. holds. Then, for any $\mathbf{x}, \mathbf{v} \in \mathcal{X}$, we have:

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

Lemma D.2. Suppose that Assumptions 2.2 and 2.3 hold for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, and $t \in [T]$, and that $\mathbf{d}_{t,\rho}^{\mathbf{x}}$ and $\mathbf{d}_{t,\rho}^{\mathbf{y}}$ are defined in (22). Then, we have

$$\|\mathbf{d}_{t,\rho}^{\mathbf{x}} - \nabla f_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))\|^2 \leq M_f^2 \left(\|\mathbf{y} - \hat{\mathbf{y}}_t^*(\mathbf{x})\|^2 + \|\mathbf{v} - \hat{\mathbf{v}}_t^*(\mathbf{x})\|^2 \right), \quad (122a)$$

$$\|\mathbf{d}_{t,\rho}^{\mathbf{y}}\|^2 \leq M_{\mathbf{v}}^2 \left(\|\mathbf{y} - \hat{\mathbf{y}}_t^*(\mathbf{x})\|^2 + \|\mathbf{v} - \hat{\mathbf{v}}_t^*(\mathbf{x})\|^2 \right), \quad (122b)$$

$$\|\nabla f_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) - \nabla f_{t,\rho}(\mathbf{x}', \hat{\mathbf{y}}_t^*(\mathbf{x}'))\| \leq L_f \|\mathbf{x} - \mathbf{x}'\|, \quad (122c)$$

$$\|\hat{\mathbf{y}}_t^*(\mathbf{x}) - \hat{\mathbf{y}}_t^*(\mathbf{x}')\| \leq L_{\mathbf{y}} \|\mathbf{x} - \mathbf{x}'\|, \quad (122d)$$

$$\|\hat{\mathbf{v}}_t^*(\mathbf{x}) - \hat{\mathbf{v}}_t^*(\mathbf{x}')\| \leq L_{\mathbf{v}} \|\mathbf{x} - \mathbf{x}'\|. \quad (122e)$$

Here, $\hat{\mathbf{y}}_t^*(\mathbf{x})$, $f_{t,\rho}$ and $\hat{\mathbf{y}}_t^*(\mathbf{x})$ are defined in (20), (18), and (19), respectively. Moreover, the constants M_f , $M_{\mathbf{v}}$, and $(L_{\mathbf{y}}, L_{\mathbf{v}}, L_f)$ are defined as in (42), (43), and (44), respectively.

Proof. We first show Eq. (122a).

Using Assumptions 2.2 and B1., we have $\nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) \succeq \mu_g$, and

$$\|\hat{\mathbf{v}}_t^*(\mathbf{x})\| = \|(\nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})))^{-1} \nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))\| \leq \frac{\ell_{f,0}}{\mu_g}. \quad (123)$$

Observe that we have

$$\begin{aligned} \|\mathbf{d}_{t,\rho}^{\mathbf{x}} - \nabla f_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))\| &\leq \|\nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))\| \\ &\quad + \|\mathbf{v} \nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}, \mathbf{y}) - \hat{\mathbf{v}}_t^*(\mathbf{x}) \nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))\| \\ &\leq \|\nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))\| \\ &\quad + \|\nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}, \mathbf{y})\| \|\mathbf{v} - \hat{\mathbf{v}}_t^*(\mathbf{x})\| \\ &\quad + \|\hat{\mathbf{v}}_t^*(\mathbf{x})\| \|\nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))\| \\ &\leq \left(\ell_{f,1} + \frac{\ell_{g,2}\ell_{f,0}}{\mu_g} \right) \|\mathbf{y} - \hat{\mathbf{y}}_t^*(\mathbf{x})\| + \ell_{g,1} \|\mathbf{v} - \hat{\mathbf{v}}_t^*(\mathbf{x})\| \\ &\leq M_f^2 (\|\mathbf{y} - \hat{\mathbf{y}}_t^*(\mathbf{x})\| + \|\mathbf{v} - \hat{\mathbf{v}}_t^*(\mathbf{x})\|), \end{aligned} \quad (124)$$

where M_f is defined as in (42); the third inequality is by Assumption 2.3 and the last inequality is by Eq. (123).

We now show Eq. (122b).

Since $\mathbf{d}_{t,\rho}^{\mathbf{v}*} := \nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) + \nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) \hat{\mathbf{v}}_t^*(\mathbf{x}) = 0$, we have

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

Then, from Assumption 2.3 and Eq. (123), we have

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

where $M_{\mathbf{v}}$ is defined as in (43).

The proofs of Eqs. (122c)-(122e) follow from [67, Lemma 17] by setting $(L_{\mathbf{y}}, L_{\mathbf{v}}, L_f)$ as in (44). $\square$

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

The following two lemmas are inspired by [24].

Lemma D.3. *Given $\rho = (\rho_{\mathbf{s}}, \rho_{\mathbf{r}})$ as positive smoothing parameters, let $g_{t,\rho}(\mathbf{x}, \mathbf{y})$ and $f_{t,\rho}(\mathbf{x}, \mathbf{y})$ be the functions defined by (18).*

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

$$|g_{t,\rho}(\mathbf{x}, \mathbf{y}) - g_t(\mathbf{x}, \mathbf{y})| \leq \frac{\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{2}. \quad (125)$$

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

$$|f_{t,\rho}(\mathbf{x}, \mathbf{y}) - f_t(\mathbf{x}, \mathbf{y})| \leq \frac{\ell_{f,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{2}. \quad (126)$$

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

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

Thus, we get

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

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 $\rho = (\rho_{\mathbf{s}}, \rho_{\mathbf{r}})$ as positive smoothing parameters, let $g_{t,\rho}(\mathbf{x}, \mathbf{y})$ and $f_{t,\rho}(\mathbf{x}, \mathbf{y})$ be the functions defined by (18).

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

$$\|\nabla g_{t,\rho}(\mathbf{x}, \mathbf{y}) - \nabla g_t(\mathbf{x}, \mathbf{y})\| \leq \frac{\ell_{g,1}(\rho_{\mathbf{s}}d_1 + \rho_{\mathbf{r}}d_2)}{2}. \quad (127)$$

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

$$\|\nabla f_t(\mathbf{x}, \mathbf{y}) - \nabla f_{t,\rho}(\mathbf{x}, \mathbf{y})\| \leq \frac{\ell_{f,1}(\rho_{\mathbf{s}}d_1 + \rho_{\mathbf{r}}d_2)}{2}. \quad (128)$$

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.

$$\begin{aligned} & \|\nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y})\| \\ &= \left\| \frac{1}{S(d_1)} \left(\frac{d_1}{\rho_{\mathbf{s}}} \int_{U_{B_1}} g_t(\mathbf{x} + \rho_{\mathbf{s}}\mathbf{s}, \mathbf{y}) d\mathbf{s} \right) - \nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}) \right\| \\ &= \left\| \frac{1}{S(d_1)} \left(\frac{d_1}{\rho_{\mathbf{s}}} \int_{U_{B_1}} g_t(\mathbf{x} + \rho_{\mathbf{s}}\mathbf{s}, \mathbf{y}) d\mathbf{s} - \int_{U_{B_1}} \frac{d_1}{\rho_{\mathbf{s}}} g_t(\mathbf{x}, \mathbf{y}) d\mathbf{s} \right. \right. \\ &\quad \left. \left. - \int_{U_{B_1}} \frac{d_1}{\rho_{\mathbf{s}}} \langle \nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}), \rho_{\mathbf{s}}\mathbf{s} \rangle d\mathbf{s} \right) \right\| \\ &\leq \frac{d_1}{S(d_1)\rho_{\mathbf{s}}} \int_{U_{B_1}} \left| g_t(\mathbf{x} + \rho_{\mathbf{s}}\mathbf{s}, \mathbf{y}) - g_t(\mathbf{x}, \mathbf{y}) - \langle \nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}), \rho_{\mathbf{s}}\mathbf{s} \rangle \right| \|\mathbf{s}\| d\mathbf{s} \\ &\leq \frac{d_1}{S(d_1)\rho_{\mathbf{s}}} \cdot \frac{\ell_{g,1}\rho_{\mathbf{s}}^2}{2} \int_{U_{B_1}} \|\mathbf{s}\|^3 d\mathbf{s} \\ &= \frac{\rho_{\mathbf{s}}d_1\ell_{g,1}}{2}, \end{aligned} \quad (129)$$

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.

$$\begin{aligned}
& \|\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y})\| \\
&= \left\| \frac{1}{S(d_2)} \left(\frac{d_2}{\rho_{\mathbf{r}}} \int_{U_{B_2}} g_t(\mathbf{x}, \mathbf{y} + \rho_{\mathbf{r}} \mathbf{r}) d\mathbf{r} \right) - \nabla_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}) \right\| \\
&= \left\| \frac{1}{S(d_2)} \left(\frac{d_2}{\rho_{\mathbf{r}}} \int_{U_{B_2}} g_t(\mathbf{x}, \mathbf{y} + \rho_{\mathbf{r}} \mathbf{r}) d\mathbf{r} - \int_{U_{B_2}} \frac{d_2}{\rho_{\mathbf{r}}} g_t(\mathbf{x}, \mathbf{y}) d\mathbf{r} \right. \right. \\
&\quad \left. \left. - \int_{U_{B_2}} \frac{d_2}{\rho_{\mathbf{r}}} \langle \nabla_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}), \rho_{\mathbf{r}} \mathbf{r} \rangle d\mathbf{r} \right) \right\| \\
&\leq \frac{d_2}{S(d_2) \rho_{\mathbf{r}}} \int_{U_{B_2}} |g_t(\mathbf{x}, \mathbf{y} + \rho_{\mathbf{r}} \mathbf{r}) - g_t(\mathbf{x}, \mathbf{y}) - \langle \nabla_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}), \rho_{\mathbf{r}} \mathbf{r} \rangle| \|\mathbf{r}\| d\mathbf{r} \\
&\leq \frac{d_2}{S(d_2) \rho_{\mathbf{r}}} \cdot \frac{\ell_{g,1} \rho_{\mathbf{r}}^2}{2} \int_{U_{B_2}} \|\mathbf{r}\|^3 d\mathbf{r} \\
&= \frac{\rho_{\mathbf{r}} d_2 \ell_{g,1}}{2},
\end{aligned} \tag{130}$$

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

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

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

$$\|\nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y})\| \leq \frac{\rho_{\mathbf{s}} d_1 \ell_{f,1}}{2}, \tag{131}$$

and

$$\|\nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y})\| \leq \frac{\rho_{\mathbf{r}} d_2 \ell_{f,1}}{2}, \tag{132}$$

which implies

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

□

Lemma D.5. Suppose Assumption B3. holds. Let $\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathbf{B}}_t)$ and $\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathbf{B}}_t)$ be defined as in (24a) and (24b), 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

$$\mathbb{E}_{(\mathbf{r}, \bar{\mathbf{B}}_t)} \left[\|\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathbf{B}}_t) - \hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathbf{B}}_t)\|^2 \right] \leq 3d_2 \ell_{g,1}^2 \|\mathbf{y} - \mathbf{y}\|^2 + \frac{3\ell_{g,1}^2 d_2^2 \rho_{\mathbf{r}}^2}{2}, \tag{133a}$$

$$\mathbb{E}_{(\mathbf{s}, \bar{\mathbf{B}}_t)} \left[\|\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathbf{B}}_t) - \hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathbf{B}}_t)\|^2 \right] \leq 3d_1 \ell_{g,1}^2 \|\mathbf{x} - \mathbf{x}\|^2 + \frac{3\ell_{g,1}^2 d_1^2 \rho_{\mathbf{s}}^2}{2}, \tag{133b}$$

for all $\mathbf{y} \in \mathbb{R}^{d_2}$ and $\mathbf{x} \in \mathbb{R}^{d_1}$.

Proof. The proof is similar to that of Lemma 5 in [42].

□

Lemma D.6. Suppose Assumptions 2.2 and B3. hold. Let (ρ_s, ρ_r) be positive smoothing parameters. Let $\mathbf{y}_t^*(\mathbf{x})$ and $\hat{\mathbf{y}}_t^*(\mathbf{x})$ be defined in (1) and (19), respectively. Then, we have

$$\mathbb{E} \left[\|\hat{\mathbf{y}}_t^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2 \right] \leq \frac{\ell_{g,1}(\rho_s^2 + \rho_r^2)}{\mu_g}. \quad (134)$$

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 2.2, $g_t(\mathbf{x}, \mathbf{y})$ is μ_g -strongly convex with respect to $\mathbf{y}$, it follows from Lemma B.2 that

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

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

$$\|\hat{\mathbf{y}}_t^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2 \leq \frac{2}{\mu_g} (g_t(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) - g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))). \quad (135)$$

Similarly, from (19), we have

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

where $\rho = (\rho_s, \rho_r)$. By Assumption 2.2, $g_{t,\rho}(\mathbf{x}, \mathbf{y})$ is μ_g -strongly convex with respect to $\mathbf{y}$. Hence, according to Lemma B.2, we obtain

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

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

$$\|\mathbf{y}_t^*(\mathbf{x}) - \hat{\mathbf{y}}_t^*(\mathbf{x})\|^2 \leq \frac{2}{\mu_g} (g_{t,\rho}(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) - g_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))). \quad (136)$$

Summing up (135) and (136), we get

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

which implies

$$\begin{aligned} \|\mathbf{y}_t^*(\mathbf{x}) - \hat{\mathbf{y}}_t^*(\mathbf{x})\|^2 &\leq \frac{1}{\mu_g} |g_{t,\rho}(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) - g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))| \\ &\quad + \frac{1}{\mu_g} |g_t(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) - g_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))| \\ &\leq \frac{\ell_{g,1}(\rho_s^2 + \rho_r^2)}{\mu_g}, \end{aligned}$$

where the last inequality is by Eq. (125). $\square$

D.3 Bounds on the Zeroth-Order Inner Solution

Recall that $\mathbf{s} \in \mathbb{R}^{d_1}$ and $\mathbf{r} \in \mathbb{R}^{d_2}$ are vectors uniformly sampled from the unit balls B_1 and B_2 , respectively. Let

$$\begin{aligned} U_b^s &= \{\mathbf{s}_i \in \mathbb{R}^{d_1}\}_{i=1}^b, & U_b^r &= \{\mathbf{r}_i \in \mathbb{R}^{d_2}\}_{i=1}^b, \\ U_b^s &= \{\mathbf{s}_i \in \mathbb{R}^{d_1}\}_{i=1}^{\bar{b}}, & U_b^r &= \{\mathbf{r}_i \in \mathbb{R}^{d_2}\}_{i=1}^{\bar{b}}, \end{aligned}$$

be generated from the uniform distributions over the unit spheres (U_{B_1}, U_{B_2}) . Here, (U_{B_1}, U_{B_2}) denote the uniform distributions over the (d_1, d_2) -dimensional unit Euclidean balls (B_1, B_2) , respectively.

Then, similar to (23), we have

$$\begin{aligned} \mathbb{E}_{(U_b^s, \mathcal{B}_t)} [\hat{\nabla}_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}; \mathcal{B}_t)] &= \nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}, \mathbf{y}), & \mathbb{E}_{(U_b^s, \mathcal{B}_t)} [\hat{\nabla}_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}; \mathcal{B}_t)] &= \nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}, \mathbf{y}), \\ \mathbb{E}_{(U_b^r, \bar{\mathcal{B}}_t)} [\hat{\nabla}_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathcal{B}}_t)] &= \nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}, \mathbf{y}), & \mathbb{E}_{(U_b^r, \bar{\mathcal{B}}_t)} [\hat{\nabla}_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}; \bar{\mathcal{B}}_t)] &= \nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{x}, \mathbf{y}). \end{aligned} \quad (137)$$

Lemma D.7. 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

$$e_t^{g\rho} := \nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \hat{\mathbf{d}}_t^{\mathbf{y}}. \quad (138)$$

Then, we have

$$\begin{aligned} \mathbb{E}\|e_{t+1}^{g\rho}\|^2 &\leq (1 - \gamma_{t+1})^2 \mathbb{E}\|e_t^{g\rho}\|^2 + 12(1 - \gamma_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_{t-1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &\quad + 9d_2^2 \ell_{g,1}^2 (1 - \gamma_{t+1})^2 \rho_{\mathbf{r}}^2 + 24d_2 \ell_{g,1}^2 (1 - \gamma_{t+1})^2 \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &\quad + 24d_2 \ell_{g,1}^2 (1 - \gamma_{t+1})^2 \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 + 2 \frac{\hat{\sigma}_{g\mathbf{y}}^2}{b} \gamma_{t+1}^2. \end{aligned} \quad (139)$$

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

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

Then, we have

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

From (137), we have

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

then, we have

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

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 D1., we have

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

where the second inequality follows from Young's inequality and Lemma D.5. From Eq. (130), we have

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

Finally, we get

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

□

Lemma D.8. Suppose Assumptions 2.2 and B3. hold. Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$ generated by Algorithm 2, we have

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

where $a > 0$ is a constant, $e_t^{g\rho}$ is defined in (138), and $\hat{\mathbf{y}}_t^*(\mathbf{x}_t)$ is defined in (19).

Proof. From Lemma B.4, we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] & = \mathbb{E}[\|\mathbf{y}_t - \beta_t \hat{\mathbf{d}}_t^{\mathbf{y}} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] \\
& \leq (1 + a) \mathbb{E}[\|\mathbf{y}_t - \beta_t \nabla_{\mathbf{y}}g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] \\
& + (1 + \frac{1}{a})\beta_t^2 \mathbb{E}[\|\hat{\mathbf{d}}_t^{\mathbf{y}} - \nabla_{\mathbf{y}}g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2].
\end{aligned} \tag{140}$$

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

$$\begin{aligned}
\mathbb{E} [\|\mathbf{y}_t - \beta_t \nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] &= \mathbb{E} [\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] + \beta_t^2 \mathbb{E} [\|\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2] \\
&\quad - 2\beta_t \mathbb{E} [\langle \nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t) \rangle] \\
&\leq \left(1 - 2\beta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right) \mathbb{E} [\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] \\
&\quad - \left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right) \mathbb{E} [\|\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2], \quad (141)
\end{aligned}$$

where the inequality results from the strong convexity of $g_{t,\rho}$ by Assumption 2.2, which implies

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

Substituting (141) into (140), gives the desired result. $\square$

For notational brevity in the analysis, we define

$$\hat{\theta}_t^{\mathbf{y}} := \|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2, \quad \hat{\theta}_t^{\mathbf{v}} := \|\mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2, \quad (142)$$

where $\hat{\mathbf{y}}_t^*(\mathbf{x})$ and $\hat{\mathbf{v}}_t^*(\mathbf{x})$ are defined in (19) and (20), respectively.

Lemma D.9. Suppose Assumptions 2.2 and B3. hold. Let $\hat{\theta}_t^{\mathbf{y}}$ be defined in (142). Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t)\}_{t=1}^T$ generated by Algorithm 2 guarantees the following bound:

$$\begin{aligned}
&\sum_{t=1}^T \left(\mathbb{E}[\hat{\theta}_{t+1}^{\mathbf{y}}] - \mathbb{E}[\hat{\theta}_t^{\mathbf{y}}] \right) \\
&\leq \left(-\frac{L_{\mu_g}}{2} \sum_{t=1}^T \mathbb{E}[\hat{\theta}_t^{\mathbf{y}}] + \frac{2}{L_{\mu_g}} \sum_{t=1}^T \mathbb{E} [\|e_t^{g_\rho}\|^2] \right) \beta_t + \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}} \sum_{t=1}^T \mathbb{E} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \frac{1}{\beta_t} \\
&\quad + \sum_{t=1}^T \left(\frac{24\ell_{g,1}}{L_{\mu_g} \mu_g} (\rho_s^2 + \rho_r^2) + \frac{12}{L_{\mu_g}} \sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2 \right) \frac{1}{\beta_t} \\
&\quad + \sum_{t=1}^T \left(-\frac{2\beta_t}{\mu_g + \ell_{g,1}} + \beta_t^2 \right) \mathbb{E} [\|\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2], \quad (143)
\end{aligned}$$

where $L_{\mathbf{y}} = \frac{\ell_{g,1}}{\mu_g}$ is defined as in (44) and $L_{\mu_g} = \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}$.

Proof. From Lemma B.4, we have for any $c > 0$

$$\begin{aligned}
\mathbb{E} [\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1})\|^2] &= \mathbb{E} [\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t) + \hat{\mathbf{y}}_t^*(\mathbf{x}_t) - \hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1})\|^2] \\
&\leq (1+c) \mathbb{E} [\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] \\
&\quad + \left(1 + \frac{1}{c}\right) \mathbb{E} [\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2]. \quad (144)
\end{aligned}$$

From Lemma D.8, we have for any $a > 0$

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

Substituting (145) into (144), we get

$$\begin{aligned}
& \mathbb{E} [\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1})\|^2] \\
& \leq (1+c)(1+a) \left(1 - 2\beta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}\right) \mathbb{E} [\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] \\
& + \left(-(1+c)(1+a) \left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right) \mathbb{E} [\|\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2] \\
& + (1+c)(1+\frac{1}{a})\beta_t^2 \mathbb{E} [\|e_t^{g\rho}\|^2] + \left(1 + \frac{1}{c}\right) \mathbb{E} [\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2]. \tag{146}
\end{aligned}$$

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.

$$\begin{aligned}
(1+c)(1+a) (1 - 2\beta_t L_{\mu_g}) &= 1 - \frac{\beta_t L_{\mu_g}}{2}, \\
(1+a) (1 - 2\beta_t L_{\mu_g}) &= 1 - \beta_t L_{\mu_g}, \\
(1+c) (1 - \beta_t L_{\mu_g}) &= 1 - \frac{\beta_t L_{\mu_g}}{2}, \\
1 + \frac{1}{a} &\leq \frac{1}{\beta_t L_{\mu_g}}, \quad 1 + \frac{1}{c} \leq \frac{2}{\beta_t L_{\mu_g}},
\end{aligned} \tag{147}$$

where $L_{\mu_g} = \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}$. Based on (146) and (147), we get

$$\begin{aligned}
& \mathbb{E} [\|\mathbf{y}_{t+1} - \hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1})\|^2] - \mathbb{E} [\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] \\
& \leq -\frac{\beta_t L_{\mu_g}}{2} \mathbb{E} [\|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] + \left(-\left(\frac{2\beta_t}{\mu_g + \ell_{g,1}} - \beta_t^2\right)\right) \mathbb{E} [\|\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2] \\
& + \frac{2}{\beta_t L_{\mu_g}} \beta_t^2 \mathbb{E} [\|e_t^{g\rho}\|^2] + \frac{2}{\beta_t L_{\mu_g}} \mathbb{E} [\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2]. \tag{148}
\end{aligned}$$

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

$$\begin{aligned}
& \mathbb{E} [\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] \\
& \leq 2 (\mathbb{E} [\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_t)\|^2] + \mathbb{E} [\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2]) \\
& \leq 2 (L_{\mathbf{y}}^2 \mathbb{E} [\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2] + \mathbb{E} [\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2]), \tag{149}
\end{aligned}$$

where the second inequality is by Lemma D.2.

Moreover, from Lemma D.6, we get

$$\begin{aligned}
\mathbb{E} [\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] &\leq 3\mathbb{E} [\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_{t+1}^*(\mathbf{x}_t)\|^2] \\
&+ 3\mathbb{E} [\|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] + 3\mathbb{E} [\|\mathbf{y}_t^*(\mathbf{x}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] \\
&\leq 3\mathbb{E} [\|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] + \frac{6\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g}. \tag{150}
\end{aligned}$$

Combining (149) and (150) yields

$$\begin{aligned}
& \mathbb{E} [\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2] \\
& \leq 2 \left(L_{\mathbf{y}}^2 \mathbb{E} [\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2] + 3\mathbb{E} [\|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2] + \frac{6\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g} \right). \tag{151}
\end{aligned}$$

Substituting (151) into (148) and summing over $t \in [T]$, give the desired result. $\square$

D.4 Bounds on the Zeroth-Order System Solution

Lemma D.10. Suppose Assumptions B2. and B3. hold. Let

$$\vartheta := \mathbb{E} [\|\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) + \hat{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t; \mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\|^2],$$

where $\hat{\nabla}_{\mathbf{y}} f_t$ and $\hat{\nabla}_{\mathbf{y}}^2 g_t$ are defined in (25a) and (26a), respectively. Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 2, we have

$$\begin{aligned} \vartheta &\leq (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2\mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\ &\quad + \frac{9}{2}d_2\ell_{g,1}^2\mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + (3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_2\rho_{\mathbf{r}}^2. \end{aligned}$$

Proof. From Lemma D.5, we have

$$\begin{aligned} &\|\hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_t; \mathcal{B}_{t+1})\|^2 \\ &\leq 3d_2\ell_{f,1}^2\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 + \frac{3}{2}\ell_{f,1}^2d_2^2\rho_{\mathbf{r}}^2 \\ &\leq 6d_2\ell_{f,1}^2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6d_2\ell_{f,1}^2\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 + \frac{3}{2}\ell_{f,1}^2d_2^2\rho_{\mathbf{r}}^2. \end{aligned} \quad (152)$$

Moreover, from (26a), we have

$$\begin{aligned} &\|\hat{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\|^2 \\ &= \frac{1}{4\rho_{\mathbf{v}}^2}\|\hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1} + \rho_{\mathbf{v}}\mathbf{v}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t; \bar{\mathcal{B}}_{t+1})\|^2 \\ &\leq \frac{3}{4\rho_{\mathbf{v}}^2}d_2\ell_{g,1}^2\|(\mathbf{x}_{t+1}, \mathbf{y}_{t+1} + \rho_{\mathbf{v}}\mathbf{v}_{t+1}) - (\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 + \frac{3}{8\rho_{\mathbf{v}}^2}\ell_{g,1}^2d_2^2\rho_{\mathbf{r}}^2 \\ &\leq \frac{9}{4\rho_{\mathbf{v}}^2}d_2\ell_{g,1}^2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \frac{9}{4\rho_{\mathbf{v}}^2}d_2\ell_{g,1}^2\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\ &\quad + \frac{9}{4}d_2\ell_{g,1}^2\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + \frac{3}{8\rho_{\mathbf{v}}^2}\ell_{g,1}^2d_2^2\rho_{\mathbf{r}}^2, \end{aligned} \quad (153)$$

where the first inequality follows from Lemma D.5.

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

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

where the second inequality follows from (152) and (153). $\square$

Lemma D.11. 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

$$e_{t+1}^M := \nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) + \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) - \hat{\mathbf{d}}_{t+1}^{\mathbf{v}}, \quad \text{where} \quad (154)$$

$$\begin{aligned} \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) &= \frac{1}{2\rho_{\mathbf{v}}}(\nabla_{\mathbf{y}} g_{t+1, \rho}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1} + \rho_{\mathbf{v}}\mathbf{v}_{t+1}) \\ &\quad - \nabla_{\mathbf{y}} g_{t+1, \rho}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1} - \rho_{\mathbf{v}}\mathbf{v}_{t+1})). \end{aligned} \quad (155)$$

Then, we have

$$\begin{aligned}
\mathbb{E}\|e_{t+1}^M\|^2 &\leq (1 - \lambda_{t+1})^2 \mathbb{E}\|e_t^M\|^2 + 36\mathbb{E}\|\nabla_{\mathbf{y}}f_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}}f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\quad + \left(18d_2^2\ell_{f,1}^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_2^2\right)\rho_{\mathbf{r}}^2 + 18d_2^2\ell_{g,1}^2\frac{\rho_{\mathbf{r}}^2}{\rho_{\mathbf{v}}^2} \\
&\quad + \frac{18}{\rho_{\mathbf{v}}^2}\mathbb{E}\|\nabla_{\mathbf{y}}g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}}g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&\quad + \frac{18}{\rho_{\mathbf{v}}^2}\mathbb{E}\|\nabla_{\mathbf{y}}g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{y}}g_t(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\
&\quad + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2\mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\
&\quad + 27d_2\ell_{g,1}^2\mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + 3(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b})\lambda_{t+1}^2.
\end{aligned} \tag{156}$$

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

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

Then we have

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

Since

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

then, we have

$$\begin{aligned}
& \mathbb{E} \|\nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_{t+1}^{\mathbf{v}}\|^2 \\
&= (1 - \lambda_{t+1})^2 \mathbb{E} \|\nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{z}_t) - \hat{\mathbf{d}}_t^{\mathbf{v}}\|^2 \\
&+ \|\lambda_{t+1}(\nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1})) \\
&+ (1 - \lambda_{t+1}) \left(\nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{z}_t) \right. \\
&+ \nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t) \\
&- \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{y}}^2 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}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1}) \left. \right) \|^2 \\
&\leq (1 - \lambda_{t+1})^2 \mathbb{E} \|\nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{z}_t) - \hat{\mathbf{d}}_t^{\mathbf{v}}\|^2 \\
&+ 3(1 - \lambda_{t+1})^2 \mathbb{E} \|\nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{z}_t) \\
&+ \nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t) \\
&- \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{y}}^2 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}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1}) \|^2 \\
&+ 3\lambda_{t+1}^2 \mathbb{E} \|\nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_{t+1}) - \hat{\nabla}_{\mathbf{y}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1})\|^2 \\
&+ 3\lambda_{t+1}^2 \mathbb{E} \|\tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1})\|^2, \tag{157}
\end{aligned}$$

where the second inequality holds by Cauchy-Schwarz inequality.

Note that, for the last term on the right-hand side of (157), from (26a) and (155), we have

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

where the last inequality follows from Assumption D1..

Then, from $\mathbb{E}\|a - \mathbb{E}[a]\|^2 = \mathbb{E}\|a\|^2 - \|\mathbb{E}[a]\|^2$ and Assumptions D1. and D3., we have

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

Then, from Young's inequality and Lemma D.10, we obtain

$$\begin{aligned}
& \mathbb{E} \|\nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_{t+1}^{\mathbf{v}}\|^2 \\
&\leq (1 - \lambda_{t+1})^2 \mathbb{E} \|\nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{z}_t) - \hat{\mathbf{d}}_t^{\mathbf{v}}\|^2 \\
&+ 12(1 - \lambda_{t+1})^2 \mathbb{E} \|\nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{z}_t) - \nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{z}_t)\|^2 \\
&+ 12(1 - \lambda_{t+1})^2 \mathbb{E} \|\tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{z}_t)\|^2 \\
&+ 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_2 \|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\
&+ 27d_2\ell_{g,1}^2 \|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) d_2^2 \rho_{\mathbf{r}}^2 + 3\lambda_{t+1}^2 \left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b} \right). \tag{158}
\end{aligned}$$

For the third term on the right-hand side of (158), based on (155), we have

$$\begin{aligned} & \|\tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ & \leq \frac{1}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{y}} g_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \end{aligned} \quad (159a)$$

$$+ \frac{1}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2. \quad (159b)$$

For (159a), we get

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

where the last inequality follows from Eq. (130).

Similary, for (159b), we have

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

Substituting the above inequalities in (159), we have

$$\begin{aligned} & \|\tilde{\nabla}_{\mathbf{y}}^2 g_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ & \leq \frac{3}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\ & + \frac{3}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 + \frac{3\rho_{\mathbf{r}}^2 d_2^2 \ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}. \end{aligned} \quad (160)$$

For the second term on the right-hand side of (158), we have

$$\begin{aligned} & \|\nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ & \leq 3 \|\nabla_{\mathbf{y}} f_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ & + 3 \|\nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ & + 3 \|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ & \leq 3 \|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \frac{3\rho_{\mathbf{r}}^2 d_2^2 \ell_{f,1}^2}{2}, \end{aligned} \quad (161)$$

where the last inequality follows from Eq. (132).

From (160), (161) and (158), we get

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

□

Lemma D.12. Suppose Assumption B4. holds. Let

$$e_t^H := \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v}_t, \quad (162a)$$

$$e_t^J := \tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v}_t, \quad (162b)$$

where

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

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

(a)

$$\mathbb{E} [\|e_t^H\|^2] \leq \ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4. \quad (163a)$$

(b)

$$\mathbb{E} [\|e_t^J\|^2] \leq \ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4. \quad (163b)$$

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

$$\begin{aligned} \mathbb{E} [\|e_t^H\|] &= \mathbb{E} [\|\tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v}_t\|] \\ &\leq \frac{1}{2\rho_{\mathbf{v}}} \mathbb{E} [\|\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \rho_{\mathbf{v}} \mathbf{v}_t\|] \\ &\quad + \frac{1}{2\rho_{\mathbf{v}}} \mathbb{E} [\|\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \rho_{\mathbf{v}} \mathbf{v}_t\|] \\ &\leq \ell_{g,2} \rho_{\mathbf{v}} \mathbb{E} [\|\mathbf{v}_t\|^2] \\ &\leq \ell_{g,2} \rho_{\mathbf{v}} p^2, \end{aligned} \quad (164)$$

where the last inequality follows from (8).

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

$$\begin{aligned} \mathbb{E} [\|e_t^J\|] &= \mathbb{E} [\|\tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v}_t\|] \\ &\leq \frac{1}{2\rho_{\mathbf{v}}} \mathbb{E} [\|\nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \rho_{\mathbf{v}} \mathbf{v}_t\|] \\ &\quad + \frac{1}{2\rho_{\mathbf{v}}} \mathbb{E} [\|\nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \rho_{\mathbf{v}} \mathbf{v}_t\|] \\ &\leq \ell_{g,2} \rho_{\mathbf{v}} \mathbb{E} [\|\mathbf{v}_t\|^2] \\ &\leq \ell_{g,2} \rho_{\mathbf{v}} p^2, \end{aligned} \quad (165)$$

where the last inequality follows from (8). $\square$

Lemma D.13. Suppose Assumption B4. holds. Then, for the directions $\hat{\mathbf{d}}_t^{\mathbf{y}}$ and $\hat{\mathbf{d}}_t^{\mathbf{x}}$ provided to Algorithm 2, and

(a) for $\mathbf{d}_{t,\rho}^{\mathbf{y}}$ defined in (22b), we have

$$\mathbb{E} [\|\hat{\mathbf{d}}_t^{\mathbf{y}} - \mathbf{d}_{t,\rho}^{\mathbf{y}}\|^2] \leq 2\mathbb{E} [\|e_t^M\|^2] + 2\ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4 =: B_t, \quad (166a)$$

where $e_t^M = \nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) + \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \hat{\mathbf{d}}_t^{\mathbf{y}}$ is defined as in (154).

(b) and for $\mathbf{d}_{t,\rho}^{\mathbf{x}}$ defined in (22c), we have

$$\mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \mathbf{d}_{t,\rho}^{\mathbf{x}} \right\|^2 \right] \leq 2\mathbb{E} \left[\left\| e_t^L \right\|^2 \right] + 2\ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4, \quad (166b)$$

where

$$e_t^L := \nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) + \tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \hat{\mathbf{d}}_t^{\mathbf{x}}, \quad (166c)$$

with $\tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)$ is defined in (172).

Proof. For part (a): Let

$$\tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) = \frac{1}{2\rho_{\mathbf{v}}} (\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)). \quad (167)$$

According to the definition of $\mathbf{d}_{t,\rho}^{\mathbf{y}}$ in (22b), we have

$$\begin{aligned} \mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{y}} - \mathbf{d}_{t,\rho}^{\mathbf{y}} \right\|^2 \right] &= \mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v} \right\|^2 \right] \\ &\leq 2\mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) \right\|^2 \right] \end{aligned} \quad (168a)$$

$$+ 2\mathbb{E} \left[\left\| \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v} \right\|^2 \right]. \quad (168b)$$

Next, we separately bound (168a) and (168b) on the RHS of the above inequality.

Bounding (168a). We have

$$2\mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) \right\|^2 \right] := 2\mathbb{E} \left[\left\| e_t^M \right\|^2 \right]. \quad (169)$$

Bounding (168b). From Lemmas D.1 and D.12, we have

$$(168b) = \mathbb{E} \left[\left\| e_t^H \right\|^2 \right] \leq 3\ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4. \quad (170)$$

Combining (169) and (170) yields

$$\mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{y}} - \mathbf{d}_{t,\rho}^{\mathbf{y}} \right\|^2 \right] \leq 2\mathbb{E} \left[\left\| e_t^M \right\|^2 \right] + 2\ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4. \quad (171)$$

For part (b): Let

$$\tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) = \frac{1}{2\rho_{\mathbf{v}}} (\nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)). \quad (172)$$

According to the definition of $\mathbf{d}_{t,\rho}^{\mathbf{x}}$ in (22c), we have

$$\begin{aligned} \mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \mathbf{d}_{t,\rho}^{\mathbf{x}} \right\|^2 \right] &= \mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v} \right\|^2 \right] \\ &\leq 2\mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) \right\|^2 \right] \end{aligned} \quad (173a)$$

$$+ 2\mathbb{E} \left[\left\| \tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{xy}}^2 g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) \mathbf{v} \right\|^2 \right]. \quad (173b)$$

Next, we separately bound (173a) and (173b) on the RHS of the above inequality.

Bounding (173a). We have

$$2\mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla_{\mathbf{x}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t) \right\|^2 \right] := 2\mathbb{E} \left[\left\| e_t^L \right\|^2 \right]. \quad (174)$$

Bounding (173b). From Lemmas D.1 and D.12, we have

$$(173b) = \mathbb{E} \left[\left\| e_t^J \right\|^2 \right] \leq 2\ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4. \quad (175)$$

Combining (174)–(175) yields

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

□

Lemma D.14. Suppose Assumptions 2.2, B1., B3. and B4. hold. Set the step size δ_t and the parameter p in (8), as

$$\delta_t \leq \left(2 + \frac{1}{\ell_{g,1}^2}\right) \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}}, \quad \forall t \in [T], \quad \text{and} \quad p = \frac{\ell_{f,0}}{\mu_g}. \quad (176)$$

Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 2 and $\hat{\mathbf{v}}_t^*(\mathbf{x}_t)$ in (20), we have

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

for some $\acute{a} > 0$, where $\hat{\theta}_t^{\mathbf{v}}$ and B_t are defined in Eq. (142) and Lemma D.13, respectively.

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

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \right] &= \mathbb{E} \left[\left\| \Pi_{\mathcal{Z}_p} \left[\mathbf{v}_t - \delta_t \hat{\mathbf{d}}_t^{\mathbf{v}} \right] - \Pi_{\mathcal{Z}_p} \left[\hat{\mathbf{v}}_t^*(\mathbf{x}_t) \right] \right\|^2 \right] \\ &\leq \mathbb{E} \left[\|\mathbf{v}_t - \delta_t \hat{\mathbf{d}}_t^{\mathbf{v}} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \right] \\ &\leq (1 + \acute{a}) \underbrace{\mathbb{E} \left[\|\mathbf{v}_t - \delta_t \nabla P_t(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t), \mathbf{v}_t) - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \right]}_{I_t} \\ &\quad + \underbrace{\left(1 + \frac{1}{\acute{a}} \right) \delta_t^2 \mathbb{E} \left[\|\hat{\mathbf{d}}_t^{\mathbf{v}} - \nabla P_t(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t), \mathbf{v}_t)\|^2 \right]}_{K_t}, \end{aligned} \quad (177)$$

where $\nabla P_t(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t), \mathbf{v}_t) := \nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) \mathbf{v}_t + \nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))$; the first inequality follows from non-expansiveness property of a projection operator.

We next bound the I_t , and K_t terms in (177), respectively.

Bounding I_t . We have

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

where the inequality holds since ∇P_t is the gradient of the strongly convex quadratic program $\frac{1}{2} \mathbf{v}^\top \nabla_{\mathbf{y}}^2 g_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x})) \mathbf{v} + \mathbf{v}^\top \nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}, \hat{\mathbf{y}}_t^*(\mathbf{x}))$.

Thus, we have

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

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

$$\begin{aligned} I_t &\leq \left(1 - 2\delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}} \right) \mathbb{E} \left[\|\mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \right] + \frac{1}{\ell_{g,1}^2} \left(\frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}} \delta_t \right) \mathbb{E} \left[\|\nabla P_t(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t), \mathbf{v}_t)\|^2 \right] \\ &\leq \left(1 - \delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}} \right) \mathbb{E} \left[\|\mathbf{v}_t - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \right], \end{aligned} \quad (178)$$

where the second inequality holds since from (20), we have

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

where the second inequality follows from Assumption B3..

Bounding K_t . Let

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

From Lemma D.13, we have

$$K_t = \mathbb{E} \left[\|\hat{\mathbf{d}}_t^{\mathbf{v}} - \mathbf{d}_{t,\rho}^{\mathbf{v}}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\|^2 \right] \leq B_t. \quad (179)$$

Putting (178), and (179) together with Eq. (177) yields the desired result.

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

□

Lemma D.15. Suppose Assumptions 2.2 and 2.3 hold. Let $\hat{\theta}_t^{\mathbf{v}}$ be defined in (142). 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:

$$\begin{aligned} \sum_{t=1}^T \left(\mathbb{E}[\hat{\theta}_{t+1}^{\mathbf{v}}] - \mathbb{E}[\hat{\theta}_t^{\mathbf{v}}] \right) &\leq \sum_{t=1}^T \left(-\frac{L_{\mu_g}}{4} \mathbb{E}[\hat{\theta}_t^{\mathbf{v}}] + \frac{4}{L_{\mu_g}} B_t \right) \delta_t \\ &\quad + \frac{16\nu^2}{L_{\mu_g} \mu_g^2} (2L_{\mathbf{y}}^2 + 1) \sum_{t=1}^T \mathbb{E} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \frac{1}{\delta_t} \\ &\quad + \sum_{t=1}^T \left(\frac{96\ell_{g,1}\nu^2}{L_{\mu_g} \mu_g^3} (\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2) + \frac{48\nu^2}{L_{\mu_g} \mu_g^2} \sup_{\mathbf{x} \in \mathcal{X}} \|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\|^2 \right) \frac{1}{\delta_t}, \end{aligned} \quad (180)$$

where B_t , ν and $(L_{\mu_g}, L_{\mathbf{y}})$ are defined in Lemmas D.13, C.7 and D.9, respectively.

Proof. From Lemma B.4, we have, for any $\acute{c} > 0$

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1})\|^2 \right] &= \mathbb{E} \left[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t) + \hat{\mathbf{v}}_t^*(\mathbf{x}_t) - \hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1})\|^2 \right] \\ &\leq (1 + \acute{c}) \mathbb{E} \left[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \right] \\ &\quad + \left(1 + \frac{1}{\acute{c}} \right) \mathbb{E} \left[\|\hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \right]. \end{aligned} \quad (181)$$

From Lemma D.14, we have, for any $\acute{a} > 0$

$$\mathbb{E} \left[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \right] \leq (1 + \acute{a}) \left(1 - \delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}} \right) \hat{\theta}_t^{\mathbf{v}} + \left(1 + \frac{1}{\acute{a}} \right) \delta_t^2 B_t. \quad (182)$$

Substituting (182) into (181), we get

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1})\|^2 \right] &\leq (1 + \acute{c}) (1 + \acute{a}) \left(1 - \delta_t \frac{\mu_g \ell_{g,1}}{\mu_g + \ell_{g,1}} \right) \hat{\theta}_t^{\mathbf{v}} \\ &\quad + (1 + \acute{c}) \left(1 + \frac{1}{\acute{a}} \right) \delta_t^2 B_t \\ &\quad + \left(1 + \frac{1}{\acute{c}} \right) \mathbb{E} \left[\|\hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \right]. \end{aligned} \quad (183)$$

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.

$$\begin{aligned} (1 + \acute{c})(1 + \acute{a})(1 - \delta_t L_{\mu_g}) &= 1 - \frac{\delta_t L_{\mu_g}}{4}, \\ (1 + \acute{c}) \left(1 + \frac{1}{\acute{a}}\right) &\leq \frac{4}{\delta_t L_{\mu_g}}, \\ 1 + \frac{1}{\acute{a}} &\leq \frac{2}{\delta_t L_{\mu_g}}, \quad 1 + \frac{1}{\acute{c}} \leq \frac{4}{\delta_t L_{\mu_g}}, \end{aligned} \tag{184}$$

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

Thus, we have

$$\begin{aligned} \mathbb{E} \left[\|\mathbf{v}_{t+1} - \hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1})\|^2 \right] &\leq \left(1 - \frac{\delta_t L_{\mu_g}}{4}\right) \hat{\theta}_t^{\mathbf{v}} + \frac{4}{L_{\mu_g}} \delta_t B_t \\ &\quad + \frac{4}{L_{\mu_g}} \frac{1}{\delta_t} \mathbb{E} \left[\|\hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \right]. \end{aligned} \tag{185}$$

We now bound the last term on the right-hand side of (185). By Lemma C.7, we have:

$$\begin{aligned} &\|\hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \\ &\leq 2 \frac{\nu^2}{\mu_g^2} \left(\|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \right) \\ &\leq 2 \frac{\nu^2}{\mu_g^2} \left(2 \|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_t)\|^2 \right. \\ &\quad \left. + 2 \|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \right) \\ &\leq 2 \frac{\nu^2}{\mu_g^2} \left(2L_{\mathbf{y}}^2 \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 2 \|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2 + \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \right), \end{aligned} \tag{186}$$

where the last inequality follows from Lemma D.2.

From (150), we have

$$\begin{aligned} \|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2 &\leq 3 \|\hat{\mathbf{y}}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_{t+1}^*(\mathbf{x}_t)\|^2 \\ &\quad + 3 \|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 + 3 \|\mathbf{y}_t^*(\mathbf{x}_t) - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2 \\ &\leq 3 \|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 + \frac{6\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g}. \end{aligned} \tag{187}$$

Plugging (187) into (186), we get

$$\begin{aligned} &\|\hat{\mathbf{v}}_{t+1}^*(\mathbf{x}_{t+1}) - \hat{\mathbf{v}}_t^*(\mathbf{x}_t)\|^2 \\ &\leq 4 \frac{\nu^2}{\mu_g^2} (2L_{\mathbf{y}}^2 + 1) \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &\quad + 4 \frac{\nu^2}{\mu_g^2} \left(3 \|\mathbf{y}_{t+1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\|^2 + \frac{6\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g} \right). \end{aligned} \tag{188}$$

Then, substituting (188) into (185), rearranging the resulting inequality and summing over $t \in [T]$, we obtain the desired result. $\square$

D.5 Bounds on the Zeroth-Order Estimation Error of Outer Objective

Lemma D.16. Suppose Assumptions B2. and B3. hold. Let

$$\varpi := \|\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) + \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t; \mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\|^2,$$

where $\hat{\nabla}_{\mathbf{x}} f_{t+1}$ and $\hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}$ are defined in (25b) and (26b), respectively. Then, for the sequence $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ generated by Algorithm 2, we have

$$\begin{aligned} \varpi &\leq (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\ &\quad + \frac{9}{2}d_1\ell_{g,1}^2\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + (3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_1^2\rho_{\mathbf{s}}^2. \end{aligned}$$

Proof. From Lemma D.5, we have

$$\begin{aligned} &\|\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t; \mathcal{B}_{t+1})\|^2 \\ &\leq 3d_1\ell_{g,1}^2\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 + \frac{3}{2}\ell_{f,1}^2d_1^2\rho_{\mathbf{s}}^2 \\ &\leq 6d_1\ell_{f,1}^2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6d_1\ell_{f,1}^2\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 + \frac{3}{2}\ell_{f,1}^2d_1^2\rho_{\mathbf{s}}^2. \end{aligned} \quad (189)$$

Moreover, from (26a), we have

$$\begin{aligned} &\|\hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\|^2 \\ &= \frac{1}{4\rho_{\mathbf{v}}^2} \|\hat{\nabla}_{\mathbf{x}} g_{t+1}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1} + \rho_{\mathbf{v}}\mathbf{v}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t; \bar{\mathcal{B}}_{t+1})\|^2 \\ &\leq \frac{3}{4\rho_{\mathbf{v}}^2} d_1\ell_{g,1}^2\|(\mathbf{x}_{t+1}, \mathbf{y}_{t+1} + \rho_{\mathbf{v}}\mathbf{v}_{t+1}) - (\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 + \frac{3}{8\rho_{\mathbf{v}}^2} \ell_{g,1}^2 d_1^2 \rho_{\mathbf{s}}^2 \\ &\leq \frac{9}{4\rho_{\mathbf{v}}^2} d_1\ell_{g,1}^2\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + \frac{9}{4\rho_{\mathbf{v}}^2} d_1\ell_{g,1}^2\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\ &\quad + \frac{9}{4}d_1\ell_{g,1}^2\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + \frac{3}{8\rho_{\mathbf{v}}^2} \ell_{g,1}^2 d_1^2 \rho_{\mathbf{s}}^2, \end{aligned} \quad (190)$$

where the first inequality follows from Lemma D.5.

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

$$\begin{aligned} \varpi &\leq 2\|\hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathcal{B}}_{t+1}) - \hat{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_t; \bar{\mathcal{B}}_{t+1})\|^2 \\ &\quad + 2\|\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t; \mathcal{B}_{t+1})\|^2 \\ &\leq (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + (12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\ &\quad + \frac{9}{2}d_1\ell_{g,1}^2\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + (3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_1^2\rho_{\mathbf{s}}^2, \end{aligned}$$

where the second inequality follows from (189) and (190). $\square$

Lemma D.17. 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. For e_t^L defined in (166c), we have

$$\begin{aligned} \mathbb{E}\|e_{t+1}^L\|^2 &\leq (1 - \eta_{t+1})^2\mathbb{E}\|e_t^L\|^2 + 36\mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &\quad + \left(18d_1^2\ell_{f,1}^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_1^2\right)\rho_{\mathbf{s}}^2 + 18d_1^2\ell_{g,1}^2\frac{\rho_{\mathbf{s}}^2}{\rho_{\mathbf{v}}^2} \\ &\quad + \frac{18}{\rho_{\mathbf{v}}^2}\mathbb{E}\|\nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\ &\quad + \frac{18}{\rho_{\mathbf{v}}^2}\mathbb{E}\|\nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}}\mathbf{v}_t)\|^2 \\ &\quad + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1\mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\ &\quad + 27d_1\ell_{g,1}^2\mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + 3(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{b\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b})\eta_{t+1}^2. \end{aligned} \quad (191)$$

where the second inequality holds by Cauchy-Schwarz inequality.

Note that for the last term on the right-hand side of (192), using (172) and (26b), we have

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

where the last inequality follows from Assumption D2..

Then, from $\mathbb{E}\|a - \mathbb{E}[a]\|^2 = \mathbb{E}\|a\|^2 - \|\mathbb{E}[a]\|^2$ and Assumption D4., we have

$$\begin{aligned}
& \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1, \rho}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_{t+1}^{\mathbf{x}}\|^2 \\
& \leq (1 - \eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t, \rho}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_t(\mathbf{z}_t) - \hat{\mathbf{d}}_t^{\mathbf{x}}\|^2 \\
& + 6(1 - \eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1, \rho}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t) - \nabla_{\mathbf{x}} f_{t, \rho}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_t(\mathbf{z}_t)\|^2 \\
& + 6(1 - \eta_{t+1})^2 \mathbb{E}\|\hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_{t+1}; \bar{\mathbf{B}}_{t+1}) + \hat{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}; \bar{\mathbf{B}}_{t+1}) \\
& + \hat{\nabla}_{\mathbf{x}} f_{t+1}(\mathbf{z}_t; \bar{\mathbf{B}}_{t+1}) + \hat{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t; \bar{\mathbf{B}}_{t+1})\|^2 + 3\eta_{t+1}^2 \left(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{b\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b}\right). \tag{193}
\end{aligned}$$

Then, from Young's inequality and Lemma D.16, we have

$$\begin{aligned}
& \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1, \rho}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_{t+1}^{\mathbf{x}}\|^2 \\
& \leq (1 - \eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t, \rho}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_t(\mathbf{z}_t) - \hat{\mathbf{d}}_t^{\mathbf{x}}\|^2 \\
& + 12(1 - \eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1, \rho}(\mathbf{z}_t) - \nabla_{\mathbf{x}} f_{t, \rho}(\mathbf{z}_t)\|^2 \\
& + 12(1 - \eta_{t+1})^2 \mathbb{E}\|\tilde{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_{t+1}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_t(\mathbf{z}_t)\|^2 \\
& + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\
& + 27d_1\ell_{g,1}^2\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_1^2\rho_{\mathbf{s}}^2 + 3\eta_{t+1}^2 \left(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{b\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b}\right). \tag{194}
\end{aligned}$$

For the third term on the right-hand side of (193), we have

$$\begin{aligned}
& \|\tilde{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}_{\mathbf{x}\mathbf{y}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
& \leq \frac{1}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{x}} g_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \tag{195a}
\end{aligned}$$

$$+ \frac{1}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{x}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2. \tag{195b}$$

For (195a), we get

$$\begin{aligned}
& \|\nabla_{\mathbf{x}} g_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
& \leq 3\|\nabla_{\mathbf{x}} g_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
& + 3\|\nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
& + 3\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
& \leq 3\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 + \frac{3\rho_{\mathbf{s}}^2 d_1^2 \ell_{g,1}^2}{2},
\end{aligned}$$

where the last inequality follows from Eq. (130).

Similary, for (195b), we have

$$\begin{aligned}
& \|\nabla_{\mathbf{x}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
& \leq 3\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 + \frac{3\rho_{\mathbf{s}}^2 d_1^2 \ell_{g,1}^2}{2}.
\end{aligned}$$

Substituting these inequalities in (195), we have

$$\begin{aligned}
& \|\tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
& \leq \frac{3}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
& \quad + \frac{3}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 + \frac{3\rho_{\mathbf{s}}^2 d_1^2 \ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}. \tag{196}
\end{aligned}$$

For the second term on the right-hand side of (193), we have

$$\begin{aligned}
& \|\nabla_{\mathbf{x}} f_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
& \leq 3\|\nabla_{\mathbf{x}} f_{t+1, \rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
& \quad + 3\|\nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
& \quad + 3\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
& \leq 3\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t)\|^2 + \frac{3\rho_{\mathbf{s}}^2 d_1^2 \ell_{f,1}^2}{2}, \tag{197}
\end{aligned}$$

where the last inequality follows from Eq. (132).

From (196), (197) and (194), we get

$$\begin{aligned}
& \mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1, \rho}(\mathbf{z}_{t+1}) + \tilde{\nabla}_{\mathbf{xy}}^2 g_{t+1}(\mathbf{z}_{t+1}) - \hat{\mathbf{d}}_{t+1}^{\mathbf{x}}\|^2 \\
& \leq (1 - \eta_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{x}} f_{t, \rho}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{xy}}^2 g_t(\mathbf{z}_t) - \hat{\mathbf{d}}_t^{\mathbf{x}}\|^2 \\
& \quad + 36\mathbb{E}\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t)\|^2 + 18\rho_{\mathbf{s}}^2 d_1^2 \ell_{f,1}^2 \\
& \quad + \frac{18}{\rho_{\mathbf{v}}^2} \mathbb{E}\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\
& \quad + \frac{18}{\rho_{\mathbf{v}}^2} \mathbb{E}\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 + \frac{18\rho_{\mathbf{s}}^2 d_1^2 \ell_{g,1}^2}{\rho_{\mathbf{v}}^2} \\
& \quad + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) d_1 \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\
& \quad + 27d_1 \ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) d_1^2 \rho_{\mathbf{s}}^2 + 3\eta_{t+1}^2 (\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{b\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b}).
\end{aligned}$$

□

D.6 Bounds on the Zeroth-Order Objective Function and its Projected Gradients

Lemma D.18. Suppose Assumptions 2.2, B2., B3., and 2.4 hold. Then, for the sequence of functions $\{f_{t, \rho}\}_{t=1}^T$ defined in Eq. (18), we have

$$\begin{aligned}
& \sum_{t=1}^T (f_{t, \rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_{t, \rho}(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))) \\
& \leq 2M + V_T + \ell_{f,1} \left(1 + 2\frac{\ell_{g,1}}{\mu_g}\right) T (\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2).
\end{aligned}$$

Here, V_T is defined in (11); and M is defined in Assumption 2.4.

Proof. Note that, we have

$$\begin{aligned} & \sum_{t=1}^T (f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_{t,\rho}(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))) \\ &= \sum_{t=1}^T (f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))) \end{aligned} \quad (198)$$

$$+ \sum_{t=1}^T (f_t(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))) \quad (199)$$

$$+ \sum_{t=1}^T (f_t(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1})) - f_{t,\rho}(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))). \quad (200)$$

From (126), we have

$$(198) \leq T \frac{\ell_{f,1}(\rho_s^2 + \rho_r^2)}{2}, \quad (201)$$

and

$$(200) \leq T \frac{\ell_{f,1}(\rho_s^2 + \rho_r^2)}{2}. \quad (202)$$

Moreover, from Lemma D.6, we have

$$\begin{aligned} (199) &= \sum_{t=1}^T (f_t(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \\ &+ \sum_{t=1}^T (f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1}))) \\ &+ \sum_{t=1}^T (f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1})) - f_t(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))) \\ &\leq \ell_{f,1} \sum_{t=1}^T \|\hat{\mathbf{y}}_t^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\| + \ell_{f,1} \sum_{t=1}^T \|\hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}) - \mathbf{y}_t^*(\mathbf{x}_{t+1})\| \\ &+ \sum_{t=1}^T (f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1}))) \\ &\leq 2T\ell_{f,1} \frac{\ell_{g,1}(\rho_s^2 + \rho_r^2)}{\mu_g} + \sum_{t=1}^T (f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - f_t(\mathbf{x}_{t+1}, \mathbf{y}_t^*(\mathbf{x}_{t+1}))). \end{aligned} \quad (203)$$

For the last term of the above inequality, we have

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

which implies that

$$(199) \leq 2T\ell_{f,1} \frac{\ell_{g,1}(\rho_s^2 + \rho_r^2)}{\mu_g} + 2M + V_T. \quad (204)$$

From (201), (202), and (204), we get the desired result. $\square$

Lemma D.19. Suppose that Assumptions 2.2 and 2.3 hold. Let $f_{t,\rho}$ be defined as in (18). Then, for $\hat{\mathbf{d}}_t^{\mathbf{x}}$ generated by Algorithm 2, for all $t \in [T]$, we have

$$\begin{aligned} \mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) \right\|^2 \right] &\leq 4\mathbb{E} \left[\|e_t^L\|^2 \right] + 4\ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4 \\ &\quad + 2M_f^2 \left(\mathbb{E}[\hat{\theta}_t^{\mathbf{y}}] + \mathbb{E}[\hat{\theta}_t^{\mathbf{v}}] \right) := A_t, \end{aligned} \quad (205)$$

where e_t^L is defined in Lemma D.13, and $\hat{\theta}_t^{\mathbf{y}}, \hat{\theta}_t^{\mathbf{v}}$ are as defined in (142). Additionally, M_f is given in Lemma D.2.

Proof. From $\|a + b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$, we get

$$\begin{aligned} &\mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) \right\|^2 \right] \\ &\leq 2\mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \mathbf{d}_{t,\rho}^{\mathbf{x}} \right\|^2 \right] \end{aligned} \quad (206a)$$

$$+ 2\mathbb{E} \left[\left\| \mathbf{d}_{t,\rho}^{\mathbf{x}} - \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) \right\|^2 \right], \quad (206b)$$

where $\mathbf{d}_{t,\rho}^{\mathbf{x}}$ is defined in (22c). From Lemma D.13, we have

$$(206a) \leq 4\mathbb{E} \left[\|e_t^L\|^2 \right] + 4\ell_{g,2}^2 \rho_{\mathbf{v}}^2 p^4. \quad (207)$$

Moreover, from Eq. (122a), we get

$$(206b) \leq 2M_f^2 \left(\mathbb{E}[\hat{\theta}_t^{\mathbf{y}}] + \mathbb{E}[\hat{\theta}_t^{\mathbf{v}}] \right). \quad (208)$$

Substituting (207) and (208) into (206), we conclude the desired result. $\square$

Lemma D.20. Suppose Assumptions 2.2, 2.3, and 2.4 hold. Let the sequence of functions $\{f_{t,\rho}\}_{t=1}^T$ be defined in (18), and let $\mathcal{P}_{\mathcal{X},\alpha_t}$ be given in Definition B.1. Then, for any positive choice of step sizes satisfying $\alpha_t \leq 1/4L_f$, for all $t \in [T]$, Algorithm 2 guarantees the following bound:

$$\begin{aligned} &\sum_{t=1}^T (\alpha_t - L_f \alpha_t^2) \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\ &\leq 12M + 6V_T + \sum_{t=1}^T (6\alpha_t - 3L_f \alpha_t^2) A_t \\ &\quad + \sum_{t=1}^T \left(6\ell_{f,1} \left(1 + 2\frac{\ell_{g,1}}{\mu_g} \right) + \frac{3\ell_{f,1}\ell_{g,1}}{\mu_g} (\alpha_t - L_f \alpha_t^2) \right) (\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2), \end{aligned} \quad (209)$$

where V_T and A_t are respectively defined in Eq. (11) and Lemma D.19.

Proof. Due to the L_f -smoothness of the function f_t by Eq. (39c) in Lemma C.1, $f_{t,\rho}$ is also L_f -smooth. Hence,

$$\begin{aligned} &f_{t,\rho}(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1})) - f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) \\ &\leq \langle \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L_f}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &= -\alpha_t \left\langle \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)), \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\rangle + \frac{L_f \alpha_t^2}{2} \left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\|^2. \end{aligned} \quad (210)$$

For the first term on the R.H.S of Eq. (210), we have that

$$\begin{aligned}
& -\mathbb{E} \left\langle \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)), \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\rangle \\
& = -\mathbb{E} \left\langle \hat{\mathbf{d}}_t^{\mathbf{x}}, \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\rangle \\
& = -\mathbb{E} \left\langle \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - \hat{\mathbf{d}}_t^{\mathbf{x}}, \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\rangle \\
& \leq -\frac{1}{2} \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\|^2 \right] + \frac{1}{2} \mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) \right\|^2 \right] \\
& \leq -\frac{1}{2} \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\|^2 \right] + \frac{A_t}{2},
\end{aligned} \tag{211}$$

where the first inequality follows from Lemma B.7; the last inequality follows from Lemma D.19.

Plugging the bound (211) into (210), we have that

$$\begin{aligned}
& \mathbb{E} [f_{t,\rho}(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1})) - f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))] \\
& \leq \frac{(L_f \alpha_t^2 - \alpha_t)}{2} \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\|^2 \right] + \frac{\alpha_t A_t}{2},
\end{aligned}$$

which can be rearranged into

$$\begin{aligned}
& (\alpha_t - L_f \alpha_t^2) \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\|^2 \right] \\
& \leq 2 \mathbb{E} [f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_{t,\rho}(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))] + \alpha_t A_t.
\end{aligned} \tag{212}$$

In addition, we have

$$\begin{aligned}
& \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \nabla f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\
& \leq 3 \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) - \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\
& + 3 \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))) - \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \nabla f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\
& + 3 \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\|^2 \right] \\
& \leq 3 \mathbb{E} \left[\left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) \right\|^2 \right] \\
& + 3 \mathbb{E} \left[\left\| \nabla f_{t,\rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - \nabla f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \right\|^2 \right] \\
& + 3 \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\|^2 \right],
\end{aligned}$$

where the second inequality follows from non-expansiveness of the projection operator.

Then, from Lemma D.19 and Assumption B2., we have

$$\begin{aligned}
& \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \nabla f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\
& \leq 3A_t + 3\ell_{f,1} \mathbb{E} \left[\left\| \hat{\mathbf{y}}_t^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t) \right\|^2 \right] + 3 \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\|^2 \right] \\
& \leq 3A_t + 3\ell_{f,1} \frac{\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g} + 3 \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X},\alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\|^2 \right],
\end{aligned} \tag{213}$$

where the last inequality is by Lemma D.6.

Combining (212) and (213) and summing over $t = 1$ to T , we have

$$\begin{aligned}
& \sum_{t=1}^T (\alpha_t - L_f \alpha_t^2) \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\
& \leq 6 \sum_{t=1}^T (f_{t, \rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - f_{t, \rho}(\mathbf{x}_{t+1}, \hat{\mathbf{y}}_t^*(\mathbf{x}_{t+1}))) \\
& \quad + \frac{3\ell_{f,1}\ell_{g,1}}{\mu_g} (\rho_s^2 + \rho_r^2) \sum_{t=1}^T (\alpha_t - L_f \alpha_t^2) + 3 \sum_{t=1}^T (2\alpha_t - L_f \alpha_t^2) A_t \\
& \leq 12M + 6V_T + 6\ell_{f,1} \left(1 + 2\frac{\ell_{g,1}}{\mu_g} \right) T (\rho_s^2 + \rho_r^2) \\
& \quad + \frac{3\ell_{f,1}\ell_{g,1}}{\mu_g} (\rho_s^2 + \rho_r^2) \sum_{t=1}^T (\alpha_t - L_f \alpha_t^2) + 3 \sum_{t=1}^T (2\alpha_t - L_f \alpha_t^2) A_t,
\end{aligned}$$

where the second inequality is due to Lemma D.18. $\square$

Lemma D.21. *Let the sequence $\{(\mathbf{x}_t, \mathbf{y}_t, \mathbf{v}_t)\}_{t=1}^T$ be generated by Algorithm 2.*

(a) *Then, we have*

$$\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \leq 2\beta_t^2 \|e_t^{g\rho}\|^2 + 2\beta_t^2 \|\nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2,$$

where $e_t^{g\rho}$ is defined in (138).

(b) *Suppose Assumptions 2.2, B2. and B3. hold. Then, we have*

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 & \leq 4\alpha_t^2 \left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \\
& \quad + \frac{4\ell_{f,1}\ell_{g,1}\alpha_t^2(\rho_s^2 + \rho_r^2)}{\mu_g} + 2A_t\alpha_t^2,
\end{aligned} \tag{214}$$

where A_t is defined in (205).

(c) *Suppose Assumptions B1., B2. and B3. hold. Then, we have*

$$\begin{aligned}
\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 & \leq 2\delta_t^2 \|e_t^M\|^2 + 3d_2^2 \ell_{f,1}^2 \delta_t^2 \rho_r^2 \\
& \quad + (12\ell_{f,0}^2 + 6\ell_{g,1}^2 p^2) \delta_t^2 + 6\ell_{g,1}^2 \frac{\delta_t^2}{\rho_{\mathbf{v}^2}} \hat{\theta}_t^{\mathbf{y}},
\end{aligned}$$

where e_t^M and $\hat{\theta}_t^{\mathbf{y}}$ are defined in (154) and (142), respectively.

Proof. **For part (a):** From Algorithm 2, we have

$$\begin{aligned}
\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 & = \beta_t^2 \|\hat{\mathbf{d}}_t^{\mathbf{y}}\|^2 \\
& \leq 2\beta_t^2 \|\hat{\mathbf{d}}_t^{\mathbf{y}} - \nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2 + 2\beta_t^2 \|\nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
& = 2\beta_t^2 \|e_t^{g\rho}\|^2 + 2\beta_t^2 \|\nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2.
\end{aligned} \tag{215}$$

For part (b):

From the update rule in Algorithm 2, we obtain

$$\begin{aligned}
\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 &= \alpha_t^2 \left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) \right\|^2 \\
&\leq 2\alpha_t^2 \left(\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))) \right\|^2 \right. \\
&\quad \left. + \left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \hat{\mathbf{d}}_t^{\mathbf{x}}) - \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))) \right\|^2 \right) \\
&\leq 2\alpha_t^2 \left(\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))) \right\|^2 \right. \\
&\quad \left. + \left\| \hat{\mathbf{d}}_t^{\mathbf{x}} - \nabla f_{t, \rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) \right\|^2 \right) \\
&\leq 2\alpha_t^2 \left(\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))) \right\|^2 + A_t \right), \tag{216}
\end{aligned}$$

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 Lemma D.19.

The first term in the above inequality can be bounded as

$$\begin{aligned}
&\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))) \right\|^2 \\
&\leq 2 \left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t))) - \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \\
&\quad + 2 \left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \\
&\leq 2 \left\| \nabla f_{t, \rho}(\mathbf{x}_t, \hat{\mathbf{y}}_t^*(\mathbf{x}_t)) - \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \right\|^2 \\
&\quad + 2 \left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \\
&\leq 2\ell_{f,1} \left\| \hat{\mathbf{y}}_t^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t) \right\|^2 + 2 \left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \\
&\leq 2\ell_{f,1} \frac{\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g} + 2 \left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2, \tag{217}
\end{aligned}$$

where the last inequality follows from Lemma D.6.

Based on (217) and (216), we get

$$\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \leq 2\alpha_t^2 \left(2 \left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 + \frac{2\ell_{f,1}\ell_{g,1}(\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2)}{\mu_g} + A_t \right).$$

For part (c): From the nonexpansiveness of projection, we have

$$\begin{aligned}
\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 &= \|\Pi_{\mathcal{Z}_p}[\mathbf{v}_t - \delta_t \hat{\mathbf{d}}_t^{\mathbf{v}}] - \Pi_{\mathcal{Z}_p}[\mathbf{v}_t]\|^2 \\
&\leq \delta_t^2 \|\hat{\mathbf{d}}_t^{\mathbf{v}}\|^2 \\
&\leq 2\delta_t^2 \|\hat{\mathbf{d}}_t^{\mathbf{v}} - \nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{z}_t) - \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{z}_t)\|^2 + 2\delta_t^2 \|\nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{z}_t) + \tilde{\nabla}_{\mathbf{y}}^2 g_t(\mathbf{z}_t)\|^2 \\
&= 2\delta_t^2 \|e_t^M\|^2 \\
&\quad + 2\delta_t^2 \|\nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t) + \frac{1}{2\rho_{\mathbf{v}}} (\nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t))\|^2 \\
&\leq 2\delta_t^2 \|e_t^M\|^2 + 6\delta_t^2 \|\nabla_{\mathbf{y}} f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\
&\quad + \frac{3\delta_t^2}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 + \frac{3\delta_t^2}{2\rho_{\mathbf{v}}^2} \|\nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2, \tag{218}
\end{aligned}$$

where the second equality follows from (154).

From Assumption B3., Lemma B.3 and (8), we have

$$\begin{aligned}
\|\nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 &\leq \ell_{g,1}^2 \|\mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2 \\
&\leq 2\ell_{g,1}^2 \|\rho_{\mathbf{v}} \mathbf{v}_t\|^2 + 2\ell_{g,1}^2 \|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2 \\
&\leq 2\ell_{g,1}^2 \rho_{\mathbf{v}}^2 p^2 + 2\ell_{g,1}^2 \|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2. \tag{219}
\end{aligned}$$

Similarly, we get

$$\|\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t - \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \leq 2\ell_{g,1}^2 \rho_{\mathbf{v}}^2 p^2 + 2\ell_{g,1}^2 \|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2. \quad (220)$$

Moreover, from Eq. (132) and Assumption B1., we have

$$\begin{aligned} \|\nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2 &\leq 2\|\nabla_{\mathbf{y}} f_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 + 2\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &\leq \frac{d_2^2 \ell_{f,1}^2 \rho_{\mathbf{r}}^2}{2} + 2\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &\leq \frac{d_2^2 \ell_{f,1}^2 \rho_{\mathbf{r}}^2}{2} + 2\ell_{f,0}^2. \end{aligned} \quad (221)$$

Substituting (219), (220) and (221), into (218), we get

$$\begin{aligned} \|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 &\leq 2\delta_t^2 \|e_t^M\|^2 + 3d_2^2 \ell_{f,1}^2 \delta_t^2 \rho_{\mathbf{r}}^2 \\ &\quad + (12\ell_{f,0}^2 + 6\ell_{g,1}^2 p^2) \delta_t^2 + \frac{6\ell_{g,1}^2}{\rho_{\mathbf{v}}^2} \delta_t^2 \|\mathbf{y}_t - \hat{\mathbf{y}}_t^*(\mathbf{x}_t)\|^2. \end{aligned}$$

□

D.7 Proof of Theorem 3.2

Proof. Since $(1 - \gamma_{t+1})^2 \leq 1 - \gamma_{t+1}$ and $\gamma_{t+1} = c_{\gamma} \alpha_t$ in (31), from (139), we have

$$\begin{aligned} \mathbb{E}\|e_{t+1}^{g_{\rho}}\|^2 - \mathbb{E}\|e_t^{g_{\rho}}\|^2 &\leq -c_{\gamma} \alpha_t \mathbb{E}\|e_t^{g_{\rho}}\|^2 \\ &\quad + 12(1 - \gamma_{t+1})^2 \mathbb{E}\|\nabla_{\mathbf{y}} g_{t-1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &\quad + 9d_2^2 \ell_{g,1}^2 (1 - \gamma_{t+1})^2 \rho_{\mathbf{r}}^2 + 24d_2 \ell_{g,1}^2 (1 - \gamma_{t+1})^2 \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\ &\quad + 24d_2 \ell_{g,1}^2 (1 - \gamma_{t+1})^2 \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 + 2\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b} \gamma_{t+1}^2. \end{aligned} \quad (222)$$

Since $(1 - \eta_{t+1})^2 \leq 1 - \eta_{t+1}$ and $\eta_{t+1} = c_{\eta} \alpha_t$ in (31), from (191), we have

$$\begin{aligned} \mathbb{E}\|e_{t+1}^L\|^2 - \mathbb{E}\|e_t^L\|^2 &\leq -c_{\eta} \alpha_t \mathbb{E}\|e_t^L\|^2 + 36\mathbb{E}\|\nabla_{\mathbf{x}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &\quad + \left(18d_1^2 \ell_{f,1}^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_1^2\right) \rho_{\mathbf{s}}^2 + 18d_1^2 \ell_{g,1}^2 \frac{\rho_{\mathbf{s}}^2}{\rho_{\mathbf{v}}^2} \\ &\quad + \frac{36}{\rho_{\mathbf{v}}^2} \mathbb{E}\|\nabla_{\mathbf{x}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\ &\quad + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1 \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_1 \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\ &\quad + 27\ell_{g,1}^2 d_1 \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + 3(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{b\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b}) \eta_{t+1}^2. \end{aligned} \quad (223)$$

Since $(1 - \lambda_{t+1})^2 \leq 1 - \lambda_{t+1}$ and $\lambda_{t+1} = c_{\lambda} \alpha_t$ in (31), from (156), we have

$$\begin{aligned} \mathbb{E}\|e_{t+1}^M\|^2 - \mathbb{E}\|e_t^M\|^2 &\leq -c_{\lambda} \alpha_t \mathbb{E}\|e_t^M\|^2 + 36\mathbb{E}\|\nabla_{\mathbf{y}} f_{t+1}(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t)\|^2 \\ &\quad + \left(18d_2^2 \ell_{f,1}^2 + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})d_2^2\right) \rho_{\mathbf{r}}^2 + 18d_2^2 \ell_{g,1}^2 \frac{\rho_{\mathbf{r}}^2}{\rho_{\mathbf{v}}^2} \\ &\quad + \frac{36}{\rho_{\mathbf{v}}^2} \mathbb{E}\|\nabla_{\mathbf{y}} g_{t+1}(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t + \rho_{\mathbf{v}} \mathbf{v}_t)\|^2 \\ &\quad + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2 \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2})d_2 \mathbb{E}\|\mathbf{y}_{t+1} - \mathbf{y}_t\|^2 \\ &\quad + 27d_2 \ell_{g,1}^2 \mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2 + 3(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b\rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b}) \lambda_{t+1}^2. \end{aligned} \quad (224)$$

Combining the outcomes .

Let

$$\begin{aligned}
\Lambda &:= \Gamma \sum_{t=1}^T \left(\mathbb{E}[\hat{\theta}_{t+1}^{\mathbf{y}}] - \mathbb{E}[\hat{\theta}_t^{\mathbf{y}}] \right) \\
&+ \Upsilon \sum_{t=1}^T \left(\mathbb{E}[\hat{\theta}_{t+1}^{\mathbf{y}}] - \mathbb{E}[\hat{\theta}_t^{\mathbf{y}}] \right) + \frac{1}{\Phi} \sum_{t=1}^T \left(\mathbb{E}[\|e_{t+1}^{g_{\rho}}\|^2] - \mathbb{E}[\|e_t^{g_{\rho}}\|^2] \right) \\
&+ \frac{1}{\Psi} \sum_{t=1}^T \left(\mathbb{E}[\|e_{t+1}^M\|^2] - \mathbb{E}[\|e_t^M\|^2] \right) + \frac{1}{\Omega} \sum_{t=1}^T \left(\mathbb{E}[\|e_{t+1}^L\|^2] - \mathbb{E}[\|e_t^L\|^2] \right).
\end{aligned}$$

Here, we have

$$\begin{aligned}
\Gamma &= \frac{11M_f^2}{L_{\mu_g}c_{\beta}}, \quad \Upsilon = \frac{52M_f^2}{L_{\mu_g}c_{\delta}}, \\
\Phi &= \max \left\{ 240 \frac{d_2 \ell_{g,1}^2}{L_f}, \frac{12d_2 \ell_{g,1}^2 L_{\mu_g}^2 c_{\beta}^2}{L_f M_f^2} \right\}, \\
\Psi &= \max \left\{ 720 \frac{d_2 \ell_{f,1}^2}{L_f}, 27 \frac{L_{\mu_g}}{\Upsilon L_f} \ell_{g,1}^2 d_2 c_{\delta}, \frac{144d_2 \ell_{f,1}^2 (\mu_g + \ell_{g,1}) c_{\beta}}{L_f \Gamma}, \frac{36 \ell_{f,1}^2 d_2 L_{\mu_g}^2 c_{\beta}^2}{L_f M_f^2} \right\}, \\
\Omega &= \max \left\{ 720 \frac{d_1 \ell_{f,1}^2}{L_f}, 27 \frac{L_{\mu_g}}{\Upsilon L_f} \ell_{g,1}^2 d_1 c_{\delta}, \frac{144d_1 \ell_{f,1}^2 (\mu_g + \ell_{g,1}) c_{\beta}}{L_f \Gamma}, \frac{36 \ell_{f,1}^2 d_1 L_{\mu_g}^2 c_{\beta}^2}{L_f M_f^2} \right\},
\end{aligned} \tag{225}$$

with

$$\begin{aligned}
c_{\beta} &\geq \sqrt{1760} \frac{L_{\mathbf{y}} M_f}{L_{\mu_g}}, \\
c_{\delta} &\geq \sqrt{33280(1 + 2L_{\mathbf{y}}^2)} \frac{\nu M_f}{L_{\mu_g} \mu_g}, \\
c &\geq \left(\max \left\{ 4L_f, c_{\beta}(\mu_g + \ell_{g,1}), \frac{48L_{\mu_g}^2 d_2 \ell_{g,1}^2 c_{\beta}^2}{M_f^2 \Phi} \right\} \right)^3 + 1, \\
c_{\mathbf{v}} &= \max \left\{ 1080 \ell_{g,1}^2, \frac{324}{M_f^2} \ell_{g,1}^4 c_{\delta}^2, \frac{54L_{\mu_g}^2}{M_f^2} \ell_{g,1}^2 c_{\beta}^2, \frac{216}{\Gamma} \ell_{g,1}^2 c_{\beta}(\mu_g + \ell_{g,1}) \right\} \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right), \\
c_{\gamma} &= \frac{26M_f^2 \Phi}{L_{\mu_g}^2}, \\
c_{\eta} &= 26\Omega, \quad c_{\lambda} = \frac{10\Upsilon}{L_{\mu_g}} c_{\delta} \Psi.
\end{aligned} \tag{226}$$

By adding (223), (222), (224), (143), and (180), along with (209) and considering the fact that α_t decreases with respect to t , and by applying Lemma D.21, we obtain:

$$\begin{aligned} & \sum_{t=1}^T A(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] + \Lambda \\ & \leq 12M + 6V_T + \sum_{t=1}^T B(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) \mathbb{E}[\hat{\theta}_t^{\mathbf{v}}] + \sum_{t=1}^T C(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) \mathbb{E}[\hat{\theta}_t^{\mathbf{y}}] \end{aligned} \quad (227a)$$

$$+ \frac{4\ell_{f,1}\ell_{g,1}}{\mu_g} \sum_{t=1}^T E(\beta_t, \delta_t, \rho_{\mathbf{v}}) \alpha_t^2 (\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2) + \sum_{t=1}^T L(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) \mathbb{E} \|e_t^L\|^2 \quad (227b)$$

$$+ \frac{8\ell_{g,2}^2 p^4 \Upsilon}{L\mu_g} \sum_{t=1}^T \delta_t \rho_{\mathbf{v}}^2 + 4\ell_{g,2}^2 p^4 \sum_{t=1}^T (6\alpha_t - 3L_f \alpha_t^2 + 2\alpha_t^2 E(\beta_t, \delta_t, \rho_{\mathbf{v}})) \rho_{\mathbf{v}}^2 \quad (227c)$$

$$+ \left(\frac{12}{L\mu_g} \frac{\Gamma}{\beta_T} + \frac{48\nu^2}{L\mu_g \mu_g^2} \frac{\Upsilon}{\delta_T} \right) H_{2,T} + \sum_{t=1}^T M(\delta_t) \mathbb{E} \|e_t^M\|^2 \quad (227d)$$

$$+ \sum_{t=1}^T Q(\beta_t, \rho_{\mathbf{v}}) \mathbb{E} \|e_t^{g\rho}\|^2 + \sum_{t=1}^T S(\beta_t, \rho_{\mathbf{v}}) \mathbb{E} [\|\nabla_{\mathbf{y}} g_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2] \quad (227e)$$

$$+ \sum_{t=1}^T Z (3d_2^2 \ell_{f,1}^2 \delta_t^2 \rho_{\mathbf{r}}^2 + (12\ell_{f,0}^2 + 6\ell_{g,1}^2 p^2) \delta_t^2) \quad (227f)$$

$$+ \frac{36}{\Psi} D_{\mathbf{y},T} + \frac{36}{\Omega} D_{\mathbf{x},T} + \frac{12}{\Phi} G_{\mathbf{y},T} + \frac{18}{\Psi \rho_{\mathbf{v}}^2} G_{\mathbf{v},T} + \frac{18}{\Omega \rho_{\mathbf{v}}^2} G_{\mathbf{x},T} \quad (227g)$$

$$+ 2 \sum_{t=1}^T \frac{\gamma_{t+1}^2}{\Phi} \frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b} + 3 \left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b \rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b} \right) \sum_{t=1}^{T+1} \frac{\lambda_{t+1}^2}{\Psi} + 3 \left(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{b \rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b} \right) \sum_{t=1}^T \frac{\eta_{t+1}^2}{\Omega} \quad (227h)$$

$$+ R(\rho_{\mathbf{v}}) T \rho_{\mathbf{r}}^2 + \dot{R}(\rho_{\mathbf{v}}) T \rho_{\mathbf{s}}^2 + 18T \ell_{g,1}^2 \left(\frac{d_1^2 \rho_{\mathbf{s}}^2}{\Omega \rho_{\mathbf{v}}^2} + \frac{d_2^2 \rho_{\mathbf{r}}^2}{\Psi \rho_{\mathbf{v}}^2} \right) + \sum_{t=1}^T D(\alpha_t, \beta_t, \delta_t) (\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2). \quad (227i)$$

Here,

$$\begin{aligned} E(\beta_t, \delta_t, \rho_{\mathbf{v}}) &:= \frac{4L_{\mathbf{y}}^2}{L\mu_g} \frac{\Gamma}{\beta_t} + \frac{16\nu^2}{L\mu_g \mu_g^2} (2L_{\mathbf{y}}^2 + 1) \frac{\Upsilon}{\delta_t} \\ &\quad + 24d_2 \frac{\ell_{g,1}^2}{\Phi} + 6(12\ell_{f,1}^2 + \frac{9\ell_{g,1}^2}{2\rho_{\mathbf{v}}^2}) \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right), \\ A(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) &:= \alpha_t - L_f \alpha_t^2 - 4E(\beta_t, \delta_t, \rho_{\mathbf{v}}) \alpha_t^2, \\ B(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) &:= -\frac{L\mu_g}{4} \Upsilon \delta_t + 2M_f^2 (6\alpha_t - 3L_f \alpha_t^2 + 2\alpha_t^2 E(\beta_t, \delta_t, \rho_{\mathbf{v}})), \\ C(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) &:= -\frac{L\mu_g}{2} \Gamma \beta_t + Z 6\ell_{g,1}^2 \frac{\delta_t^2}{\rho_{\mathbf{v}}^2} \\ &\quad + 2M_f^2 (6\alpha_t - 3L_f \alpha_t^2 + 2\alpha_t^2 E(\beta_t, \delta_t, \rho_{\mathbf{v}})), \\ Z &:= 27\ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right). \end{aligned} \quad (228)$$

Moreover,

$$\begin{aligned}
M(\delta_t) &:= -\frac{\lambda_{t+1}}{\Psi} + Z2\delta_t^2 + \frac{8\Upsilon}{L_{\mu_g}}\delta_t, \\
D(\alpha_t, \beta_t, \delta_t) &:= 6\ell_{f,1}(1 + 2\frac{\ell_{g,1}}{\mu_g}) + \frac{3\ell_{f,1}\ell_{g,1}}{\mu_g}(\alpha_t - L_f\alpha_t^2) \\
&\quad + \frac{24\ell_{g,1}}{L_{\mu_g}\mu_g}\frac{\Gamma}{\beta_t} + \frac{96\ell_{g,1}\nu^2}{L_{\mu_g}\mu_g^3}\frac{\Upsilon}{\delta_t}, \\
F(\rho_{\mathbf{v}}) &:= 24d_2\frac{\ell_{g,1}^2}{\Phi} + (72\ell_{f,1}^2 + \frac{27\ell_{g,1}^2}{\rho_{\mathbf{v}}^2})(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}), \\
S(\beta_t, \rho_{\mathbf{v}}) &:= -\frac{2\beta_t\Gamma}{\mu_g + \ell_{g,1}} + \beta_t^2\Gamma + 2F(\rho_{\mathbf{v}})\beta_t^2, \\
Q(\beta_t, \rho_{\mathbf{v}}) &:= \frac{2}{L_{\mu_g}}\Gamma\beta_t - \frac{\gamma_{t+1}}{\Phi} + 2F(\rho_{\mathbf{v}})\beta_t^2, \\
R(\rho_{\mathbf{v}}) &:= 9d_2^2\frac{\ell_{g,1}^2}{\Phi} + 18d_2^2\frac{\ell_{f,1}^2}{\Psi} + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})\frac{d_2^2}{\Psi}, \\
\dot{R}(\rho_{\mathbf{v}}) &:= 18d_1^2\frac{\ell_{f,1}^2}{\Omega} + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2})\frac{d_1^2}{\Omega}, \\
L(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) &:= -\frac{\eta_{t+1}}{\Omega} + 4(6\alpha_t - 3L_f\alpha_t^2 + 2\alpha_t^2E(\beta_t, \delta_t, \rho_{\mathbf{v}})).
\end{aligned} \tag{229}$$

We then provide bounds for the terms in (227a)-(227i).

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 $\beta_t = c_{\beta}\alpha_t$, $\delta_t = c_{\delta}\alpha_t$ in (31), we have

$$\begin{aligned}
\alpha_t^2 E(\beta_t, \delta_t, \rho_{\mathbf{v}}) &= \frac{4L_{\mathbf{y}}^2}{L_{\mu_g}}\frac{\Gamma\alpha_t^2}{\beta_t} + \frac{16\nu^2}{L_{\mu_g}\mu_g^2}(2L_{\mathbf{y}}^2 + 1)\frac{\Upsilon\alpha_t^2}{\delta_t} \\
&\quad + 24d_2\frac{\ell_{g,1}^2}{\Phi}\alpha_t^2 + (72\ell_{f,1}^2\alpha_t^2 + \frac{27\ell_{g,1}^2}{\rho_{\mathbf{v}}^2}\alpha_t^2)(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \\
&\leq \frac{44L_{\mathbf{y}}^2}{L_{\mu_g}^2}M_f^2\frac{\alpha_t}{c_{\beta}^2} + \frac{832\nu^2}{L_{\mu_g}^2\mu_g^2}(1 + 2L_{\mathbf{y}}^2)M_f^2\frac{\alpha_t}{c_{\delta}^2} \\
&\quad + 6\frac{d_2\ell_{g,1}^2}{L_f\Phi}\alpha_t + (\frac{18\ell_{f,1}^2}{L_f}\alpha_t + \frac{27\ell_{g,1}^2}{c_{\mathbf{v}}}\alpha_t)(\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \\
&\leq \frac{\alpha_t}{8},
\end{aligned} \tag{230}$$

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 (225), $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}}\alpha_t$ and $\alpha_t \leq 1/4L_f$ in (31);

the second inequality follows from $c_{\beta} \geq \sqrt{1760\frac{L_{\mathbf{y}}^2M_f^2}{L_{\mu_g}^2}}$, $c_{\delta} \geq \sqrt{33280\frac{\nu^2M_f^2}{L_{\mu_g}^2\mu_g^2}(1 + 2L_{\mathbf{y}}^2)}$, in (226);

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 $c_{\mathbf{v}} \geq 1080\ell_{g,1}^2(\frac{d_2}{\Psi} + \frac{d_1}{\Omega})$ in (225).

Moreover, we have

$$\begin{aligned}
A(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) &= \alpha_t - L_f\alpha_t^2 - 4E(\beta_t, \delta_t, \rho_{\mathbf{v}})\alpha_t^2 \\
&\geq \alpha_t - L_f\alpha_t^2 - \frac{\alpha_t}{2} \\
&\geq \frac{\alpha_t}{4},
\end{aligned} \tag{231}$$

where the last inequality is by $\alpha_t \leq 1/4L_f$ in (226).

Bounding (227a).

From $\delta_t = c_\delta \alpha_t$ in (31), we have

$$\begin{aligned}
B(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) &= -\frac{L_{\mu_g}}{4} \Upsilon \delta_t + 2M_f^2 (6\alpha_t - 3L_f \alpha_t^2 + 2\alpha_t^2 E(\beta_t, \delta_t, \rho_{\mathbf{v}})) \\
&\leq -\frac{L_{\mu_g}}{4} \Upsilon c_\delta \alpha_t + 12M_f^2 \alpha_t - 6M_f^2 L_f \alpha_t^2 + \frac{M_f^2}{2} \alpha_t \\
&\leq \left(-\frac{L_{\mu_g}}{4} \Upsilon c_\delta + \frac{25}{2} M_f^2 \right) \alpha_t \\
&\leq -\frac{1}{2} M_f^2 \alpha_t,
\end{aligned} \tag{232}$$

where the first inequality follows from (230); the last inequality is by $\Upsilon = \frac{52M_f^2}{L_{\mu_g} c_\delta}$ in (225).

From (228), we obtain

$$Z = 27\ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right).$$

Thus, from $\beta_t = c_\beta \alpha_t$, $\delta_t = c_\delta \alpha_t$ and $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}} \alpha_t$ in (31), we have

$$\begin{aligned}
C(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) &= -\frac{L_{\mu_g}}{2} \Gamma \beta_t + 162 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right) \ell_{g,1}^4 \frac{\delta_t^2}{\rho_{\mathbf{v}}^2} \\
&\quad + 2M_f^2 (6\alpha_t - 3L_f \alpha_t^2 + 2\alpha_t^2 E(\beta_t, \delta_t, \rho_{\mathbf{v}})) \\
&\leq -\frac{L_{\mu_g}}{2} \Gamma c_\beta \alpha_t + 162 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right) \ell_{g,1}^4 \frac{c_\delta^2}{c_{\mathbf{v}}} \alpha_t + \frac{9}{2} M_f^2 \alpha_t \\
&= -\frac{11}{2} M_f^2 \alpha_t + 162 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right) \ell_{g,1}^4 \frac{c_\delta^2}{c_{\mathbf{v}}} \alpha_t + \frac{9}{2} M_f^2 \alpha_t \\
&\leq -\frac{1}{2} M_f^2 \alpha_t,
\end{aligned} \tag{233}$$

where the first inequality follows from (230); the second equality follows from $\Gamma = \frac{11M_f^2}{L_{\mu_g} c_\beta}$ in (225);

the last inequality is by $c_{\mathbf{v}} \geq \frac{324}{M_f^2} \ell_{g,1}^4 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right) c_\delta^2$.

Thus, from (232) and (233), we get

$$(227a) \leq \mathcal{O}(V_T). \tag{234}$$

Bounding (227b).

From (230), we also obtain

$$\begin{aligned}
&\frac{4\ell_{f,1}\ell_{g,1}}{\mu_g} \sum_{t=1}^T E(\beta_t, \delta_t, \rho_{\mathbf{v}}) \alpha_t^2 (\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2) \\
&\leq \frac{4\ell_{f,1}\ell_{g,1}}{\mu_g} \sum_{t=1}^T \frac{\alpha_t}{8} (\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2) \\
&= \mathcal{O} \left(\sum_{t=1}^T \alpha_t (\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2) \right).
\end{aligned} \tag{235}$$

From (229) and $\eta_{t+1} = c_\eta \alpha_t$ in (31), we have

$$\begin{aligned}
L(\alpha_t, \beta_t, \delta_t, \rho_{\mathbf{v}}) &= -\frac{\eta_{t+1}}{\Omega} + 4 (6\alpha_t - 3L_f \alpha_t^2 + 2\alpha_t^2 E(\beta_t, \delta_t, \rho_{\mathbf{v}})) \\
&\leq -\frac{c_\eta}{\Omega} \alpha_t + 25\alpha_t \\
&\leq -\alpha_t,
\end{aligned}$$

where the last inequality is by $c_\eta \geq 26\Omega$ and (230).
Thus, we get

$$\sum_{t=1}^T L(\alpha_t, \beta_t, \delta_t, \rho_v) \mathbb{E} \|e_t^L\|^2 \leq 0. \quad (236)$$

From (236) and (235), we have

$$(227b) \leq \mathcal{O} \left(\sum_{t=1}^T \alpha_t (\rho_s^2 + \rho_r^2) \right). \quad (237)$$

Bounding (227c).

From $\delta_t = c_\delta \alpha_t$ in (31) and Eq. (230), we have

$$\begin{aligned} & \frac{8\ell_{g,2}^2 p^4 \Upsilon}{L_{\mu_g}} \sum_{t=1}^T \delta_t \rho_v^2 + 4\ell_{g,2}^2 p^4 \sum_{t=1}^T (6\alpha_t - 3L_f \alpha_t^2 + 2\alpha_t^2 E(\beta_t, \delta_t, \rho_v)) \rho_v^2 \\ & \leq \frac{8\ell_{g,2}^2 p^4 \Upsilon}{L_{\mu_g}} \sum_{t=1}^T c_\delta \alpha_t \rho_v^2 + 4\ell_{g,2}^2 p^4 \sum_{t=1}^T \frac{25}{4} \alpha_t \rho_v^2. \end{aligned}$$

Thus, from $\rho_v^2 = c_v \alpha_t$ in (31), we have

$$(227c) \leq \mathcal{O} \left(\sum_{t=1}^T \alpha_t^2 \right). \quad (238)$$

Bounding (227d).

From (228), we have

$$Z = 27\ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right). \quad (239)$$

From (229), $\lambda_{t+1} = c_\lambda \alpha_t$ and $\delta_t = c_\delta \alpha_t$ in (31), we have

$$\begin{aligned} M(\delta_t) &= -\frac{\lambda_{t+1}}{\Psi} + Z 2\delta_t^2 + \frac{8\Upsilon}{L_{\mu_g}} \delta_t \\ &= -\frac{c_\lambda \alpha_t}{\Psi} + 27\ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right) 2c_\delta^2 \alpha_t^2 + \frac{8\Upsilon}{L_{\mu_g}} c_\delta \alpha_t \\ &\leq -\frac{2\Upsilon}{L_{\mu_g}} c_\delta \alpha_t + \frac{27}{4L_f} \ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right) 2c_\delta^2 \alpha_t \\ &\leq -\frac{\Upsilon}{L_{\mu_g}} c_\delta \alpha_t, \end{aligned}$$

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 $\delta_t = c_\delta \alpha_t$ in (31), we get

$$\begin{aligned} (227d) &= \left(\frac{12}{L_{\mu_g}} \frac{\Gamma}{\beta_T} + \frac{48\nu^2}{L_{\mu_g} \mu_g^2} \frac{\Upsilon}{\delta_T} \right) H_{2,T} + \sum_{t=1}^T M(\delta_t) \mathbb{E} \|e_t^M\|^2 \\ &\leq \mathcal{O} \left(\frac{H_{2,T}}{\alpha_T} \right). \end{aligned} \quad (240)$$

Bounding (227e).

From (229), we have

$$F(\rho_v) = 24d_2 \frac{\ell_{g,1}^2}{\Phi} + (72\ell_{f,1}^2 + \frac{27\ell_{g,1}^2}{\rho_v^2}) \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right). \quad (241)$$

From (229), $\gamma_{t+1} = c_\gamma \alpha_t$, $\beta_t = c_\beta \alpha_t$ in (31), we have

$$\begin{aligned}
Q(\beta_t, \rho_{\mathbf{v}}) &= -\frac{\gamma_{t+1}}{\Phi} + \frac{2}{L_{\mu_g}} \Gamma \beta_t + 2F(\rho_{\mathbf{v}}) \beta_t^2 \\
&= -\frac{c_\gamma \alpha_t}{\Phi} + \frac{22M_f^2}{L_{\mu_g}^2} \alpha_t + \left(24d_2 \frac{\ell_{g,1}^2}{\Phi} + (72\ell_{f,1}^2 + \frac{27\ell_{g,1}^2}{c_{\mathbf{v}} \alpha_t}) (\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \alpha_t^2 \\
&\leq -\frac{4M_f^2}{L_{\mu_g}^2} \alpha_t + \left(24d_2 \frac{\ell_{g,1}^2}{\Phi} \alpha_t^2 + (72\ell_{f,1}^2 \alpha_t^2 + \frac{27\ell_{g,1}^2 \alpha_t}{c_{\mathbf{v}}}) (\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \\
&\leq -\frac{4M_f^2}{L_{\mu_g}^2} \alpha_t + \left(\frac{6d_2}{L_f} \frac{\ell_{g,1}^2}{\Phi} \alpha_t + (\frac{18}{L_f} \ell_{f,1}^2 \alpha_t + \frac{27\ell_{g,1}^2 \alpha_t}{c_{\mathbf{v}}}) (\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \\
&\leq -\frac{M_f^2}{L_{\mu_g}^2} \alpha_t,
\end{aligned} \tag{242}$$

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 $c_{\mathbf{v}} \geq \frac{54L_{\mu_g}^2 \ell_{g,1}^2}{M_f^2} (\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) c_\beta^2$, $\Phi \geq \frac{12d_2 \ell_{g,1}^2 L_{\mu_g}^2 c_\beta^2}{L_f M_f^2}$, 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 (229), $\beta_t = c_\beta \alpha_t$, $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}} \alpha_t$ in (31) and (241), we have

$$\begin{aligned}
S(\beta_t, \rho_{\mathbf{v}}) &= -\frac{2\beta_t \Gamma}{\mu_g + \ell_{g,1}} + \beta_t^2 \Gamma + 2F(\rho_{\mathbf{v}}) \beta_t^2 \\
&= -\frac{2c_\beta \alpha_t \Gamma}{\mu_g + \ell_{g,1}} + c_\beta^2 \alpha_t^2 \Gamma + \left(24d_2 \frac{\ell_{g,1}^2}{\Phi} + (72\ell_{f,1}^2 + \frac{27\ell_{g,1}^2}{c_{\mathbf{v}} \alpha_t}) (\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \alpha_t^2 \\
&\leq -\frac{c_\beta \alpha_t \Gamma}{\mu_g + \ell_{g,1}} + \left(24d_2 \frac{\ell_{g,1}^2}{\Phi} \alpha_t^2 + (72\ell_{f,1}^2 \alpha_t^2 + \frac{27\ell_{g,1}^2 \alpha_t}{c_{\mathbf{v}}}) (\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \\
&\leq -\frac{c_\beta \alpha_t \Gamma}{\mu_g + \ell_{g,1}} + \left(\frac{6d_2}{L_f} \frac{\ell_{g,1}^2}{\Phi} \alpha_t + (\frac{18}{L_f} \ell_{f,1}^2 \alpha_t + \frac{27\ell_{g,1}^2 \alpha_t}{c_{\mathbf{v}}}) (\frac{d_2}{\Psi} + \frac{d_1}{\Omega}) \right) 2c_\beta^2 \\
&\leq -\frac{c_\beta \alpha_t \Gamma}{4(\mu_g + \ell_{g,1})},
\end{aligned} \tag{243}$$

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 $\Phi \geq \frac{24d_2 \ell_{g,1}^2 (\mu_g + \ell_{g,1})}{L_f c_\beta \Gamma}$, and $\Psi \geq \frac{144d_2 \ell_{f,1}^2 (\mu_g + \ell_{g,1}) c_\beta}{L_f \Gamma}$, and $\Omega \geq \frac{144d_1 \ell_{f,1}^2 (\mu_g + \ell_{g,1}) c_\beta}{L_f \Gamma}$.

Thus, we get

$$(227e) = \sum_{t=1}^T Q(\beta_t, \rho_{\mathbf{v}}) \mathbb{E} \|e_t^{g_\rho}\|^2 + \sum_{t=1}^T S(\beta_t, \rho_{\mathbf{v}}) \mathbb{E} [\|\nabla_{\mathbf{y}} g_{t,\rho}(\mathbf{x}_t, \mathbf{y}_t)\|^2] \leq 0. \tag{244}$$

Bounding (227f) .

From (228), we have

$$Z = 27\ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right).$$

Thus, from $\delta_t = c_\delta \alpha_t$ in (31), we have

$$\begin{aligned}
(227f) &= \sum_{t=1}^T Z \left(3d_2^2 \ell_{f,1}^2 \delta_t^2 \rho_{\mathbf{r}}^2 + (12\ell_{f,0}^2 + 6\ell_{g,1}^2 p^2) \delta_t^2 \right) \\
&= \sum_{t=1}^T 27\ell_{g,1}^2 \left(\frac{d_2}{\Psi} + \frac{d_1}{\Omega} \right) (3d_2^2 \ell_{f,1}^2 \rho_{\mathbf{r}}^2 + (12\ell_{f,0}^2 + 6\ell_{g,1}^2 p^2)) c_\delta^2 \alpha_t^2 \\
&= \mathcal{O} \left(\sum_{t=1}^T (d_1 + d_2) (\alpha_t^2 \rho_{\mathbf{r}}^2 + \alpha_t^2) \right). \tag{245}
\end{aligned}$$

Bounding (227g). From $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}} \alpha_t$ in (31), we have

$$\begin{aligned}
(227g) &= \frac{36}{\Psi} D_{\mathbf{y},T} + \frac{36}{\Omega} D_{\mathbf{x},T} + \frac{12}{\Phi} G_{\mathbf{y},T} + \frac{36}{\Psi \rho_{\mathbf{v}}^2} G_{\mathbf{v},T} + \frac{36}{\Omega \rho_{\mathbf{v}}^2} G_{\mathbf{x},T} \\
&= \mathcal{O} \left(D_{\mathbf{y},T} + D_{\mathbf{x},T} + G_{\mathbf{y},T} + \frac{1}{\alpha_T} (G_{\mathbf{v},T} + G_{\mathbf{x},T}) \right). \tag{246}
\end{aligned}$$

Bounding (227h). From $\gamma_{t+1} = c_\gamma \alpha_t$, $\eta_{t+1} = c_\eta \alpha_t$, $\lambda_{t+1} = c_\lambda \alpha_t$ and $\rho_{\mathbf{v}}^2 = c_{\mathbf{v}} \alpha_t$ in (31), we have

$$\begin{aligned}
(227h) &= 2 \sum_{t=1}^T \frac{\gamma_{t+1}^2}{\Phi} \frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b} + 3 \left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b \rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b} \right) \sum_{t=1}^{T+1} \frac{\lambda_{t+1}^2}{\Psi} + 3 \left(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{b \rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b} \right) \sum_{t=1}^T \frac{\eta_{t+1}^2}{\Omega} \\
&= 2 \sum_{t=1}^T \frac{c_\gamma^2 \alpha_t^2}{\Phi} \frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b} + 3 \left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b \rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b} \right) \sum_{t=1}^{T+1} \frac{c_\lambda^2 \alpha_t^2}{\Psi} + 3 \left(\frac{\hat{\sigma}_{g_{\mathbf{x}}}^2}{b \rho_{\mathbf{v}}^2} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b} \right) \sum_{t=1}^T \frac{c_\eta^2 \alpha_t^2}{\Omega} \\
&= \mathcal{O} \left(\left(\frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b} + \frac{\hat{\sigma}_{g_{\mathbf{y}}}^2}{b \alpha_t} + \frac{\hat{\sigma}_{f_{\mathbf{y}}}^2}{b} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b \alpha_t} + \frac{\hat{\sigma}_{f_{\mathbf{x}}}^2}{b} \right) \sum_{t=1}^T \alpha_t^2 \right). \tag{247}
\end{aligned}$$

Bounding (227i). From $\beta_t = c_\beta \alpha_t$, $\delta_t = c_\delta \alpha_t$ in (31), we have

$$\begin{aligned}
D(\alpha_t, \beta_t, \delta_t) &= 6\ell_{f,1} (1 + 2 \frac{\ell_{g,1}}{\mu_g}) + \frac{3\ell_{f,1} \ell_{g,1}}{\mu_g} (\alpha_t - L_f \alpha_t^2) + \frac{24\ell_{g,1}}{L_{\mu_g} \mu_g} \frac{\Gamma}{\beta_t} + \frac{96\ell_{g,1} \nu^2}{L_{\mu_g} \mu_g^3} \frac{\Upsilon}{\delta_t} \\
&= 6\ell_{f,1} (1 + 2 \frac{\ell_{g,1}}{\mu_g}) + \frac{3\ell_{f,1} \ell_{g,1}}{\mu_g} (\alpha_t - L_f \alpha_t^2) + \frac{24\ell_{g,1}}{L_{\mu_g} \mu_g} \frac{\Gamma}{c_\beta \alpha_t} + \frac{96\ell_{g,1} \nu^2}{L_{\mu_g} \mu_g^3} \frac{\Upsilon}{c_\delta \alpha_t} \\
&= \mathcal{O} \left(\alpha_t + \frac{1}{\alpha_t} \right),
\end{aligned}$$

and

$$\sum_{t=1}^T D(\alpha_t, \beta_t, \delta_t) (\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2) = \mathcal{O} \left(\sum_{t=1}^T \left(\alpha_t + \frac{1}{\alpha_t} \right) (\rho_{\mathbf{s}}^2 + \rho_{\mathbf{r}}^2) \right). \tag{248}$$

Moreover, we have

$$\begin{aligned}
R(\rho_{\mathbf{v}}) &= 9d_2^2 \frac{\ell_{g,1}^2}{\Phi} + 18d_2^2 \frac{\ell_{f,1}^2}{\Psi} + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) \frac{d_2^2}{\Psi} = \mathcal{O} \left(\left(1 + \frac{1}{\rho_{\mathbf{v}}^2} \right) d_2^2 \right), \\
\dot{R}(\rho_{\mathbf{v}}) &= 18d_1^2 \frac{\ell_{f,1}^2}{\Omega} + 6(3\ell_{f,1}^2 + \frac{3\ell_{g,1}^2}{4\rho_{\mathbf{v}}^2}) \frac{d_1^2}{\Omega} = \mathcal{O} \left(\left(1 + \frac{1}{\rho_{\mathbf{v}}^2} \right) d_1^2 \right),
\end{aligned}$$

which, implies that

$$\begin{aligned}
&R(\rho_{\mathbf{v}}) T \rho_{\mathbf{r}}^2 + \dot{R}(\rho_{\mathbf{v}}) T \rho_{\mathbf{s}}^2 + 18d_1^2 \ell_{g,1}^2 \frac{T \rho_{\mathbf{s}}^2}{\Omega \rho_{\mathbf{v}}^2} + 18d_2^2 \ell_{g,1}^2 \frac{T \rho_{\mathbf{r}}^2}{\Psi \rho_{\mathbf{v}}^2} \\
&= \mathcal{O} \left(\left(1 + \frac{1}{\rho_{\mathbf{v}}^2} \right) T (d_1^2 \rho_{\mathbf{s}}^2 + d_2^2 \rho_{\mathbf{r}}^2) + \frac{T}{\rho_{\mathbf{v}}^2} (d_2^2 \rho_{\mathbf{r}}^2 + d_1^2 \rho_{\mathbf{s}}^2) \right). \tag{249}
\end{aligned}$$

From (248), (249) and $\rho_v^2 = c_v \alpha_t$ in (31), we get

$$(227i) \leq \mathcal{O} \left(\sum_{t=1}^T \left(\alpha_t + \frac{1}{\alpha_t} \right) (\rho_s^2 + \rho_r^2) + \left(1 + \frac{1}{\alpha_T} \right) T(d_2^2 \rho_r^2 + d_1^2 \rho_s^2) \right). \quad (250)$$

Combining the outcomes (227i). Combining inequalities (234), (237), (238), (240), (244), (245), (246), (247), and (250) leads to

$$\begin{aligned} & \frac{\alpha_T}{2} \sum_{t=1}^T \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] + \Lambda \\ & \leq \mathcal{O} \left(V_T + \sum_{t=1}^T \alpha_t (\rho_s^2 + \rho_r^2) + \sum_{t=1}^T \alpha_t^2 + \frac{H_{2,T}}{\alpha_T} + \sum_{t=1}^T (d_1 + d_2) (\alpha_t^2 \rho_r^2 + \alpha_t^2) \right) \\ & + \mathcal{O} \left(D_{\mathbf{y}, T} + D_{\mathbf{x}, T} + G_{\mathbf{y}, T} + \frac{1}{\alpha_T} (G_{\mathbf{v}, T} + G_{\mathbf{x}, T}) \right) \\ & + \mathcal{O} \left(\sum_{t=1}^T \left(\frac{\hat{\sigma}_{g_y}^2 \alpha_t^2}{b} + \frac{\hat{\sigma}_{g_y}^2 \alpha_t}{b} + \frac{\hat{\sigma}_{f_y}^2 \alpha_t^2}{b} + \frac{\hat{\sigma}_{g_x}^2 \alpha_t}{b} + \frac{\hat{\sigma}_{f_x}^2 \alpha_t^2}{b} \right) \right) \\ & + \mathcal{O} \left(\sum_{t=1}^T \left(\alpha_t + \frac{1}{\alpha_t} \right) (\rho_s^2 + \rho_r^2) + \left(1 + \frac{1}{\alpha_T} \right) T(d_2^2 \rho_r^2 + d_1^2 \rho_s^2) \right). \end{aligned}$$

From the definition of Λ in (105), we have

$$\begin{aligned} -\Lambda &= \Gamma \sum_{t=1}^T (\mathbb{E}[\theta_t^y] - \mathbb{E}[\theta_{t+1}^y]) \\ &+ \Upsilon \sum_{t=1}^T (\mathbb{E}[\theta_t^v] - \mathbb{E}[\theta_{t+1}^v]) + \frac{1}{\Phi} \sum_{t=1}^T (\mathbb{E}\|e_t^g\|^2 - \mathbb{E}\|e_{t+1}^g\|^2) \\ &+ \frac{1}{\Psi} \sum_{t=1}^T (\mathbb{E}\|e_t^v\|^2 - \mathbb{E}\|e_{t+1}^v\|^2) + \frac{1}{\Omega} \sum_{t=1}^T (\mathbb{E}\|e_t^f\|^2 - \mathbb{E}\|e_{t+1}^f\|^2) \\ &\leq \Gamma \theta_1^y + \Upsilon \theta_1^v + \frac{\hat{\sigma}_{g_y}^2}{\Phi} + \frac{\hat{\sigma}_{g_{yy}}^2 + \hat{\sigma}_{f_y}^2}{\Psi} + \frac{\hat{\sigma}_{g_{xy}}^2 + \hat{\sigma}_{f_x}^2}{\Omega}. \end{aligned} \quad (251)$$

From (28), we have $\hat{\sigma}^2 = \hat{\sigma}_{g_y}^2 + \hat{\sigma}_{g_{yy}}^2 + \hat{\sigma}_{f_y}^2 + \hat{\sigma}_{g_{xy}}^2 + \hat{\sigma}_{f_x}^2$.

Thus, using (251), (31), and rearranging the terms, we get

$$\begin{aligned} & \sum_{t=1}^T \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\ & \leq \frac{2}{\alpha_T} \mathcal{O} \left(V_T + \sum_{t=1}^T \alpha_t (\rho_s^2 + \rho_r^2) + \sum_{t=1}^T \alpha_t^2 + \frac{H_{2,T}}{\alpha_T} + \sum_{t=1}^T (d_1 + d_2) (\alpha_t^2 \rho_r^2 + \alpha_t^2) \right) \\ & + \frac{2}{\alpha_T} \mathcal{O} \left(D_{\mathbf{y}, T} + D_{\mathbf{x}, T} + G_{\mathbf{y}, T} + \frac{1}{\alpha_T} (G_{\mathbf{v}, T} + G_{\mathbf{x}, T}) \right) \\ & + \frac{2}{\alpha_T} \mathcal{O} \left(\sum_{t=1}^T \left(\frac{\hat{\sigma}_{g_y}^2 \alpha_t^2}{b} + \frac{\hat{\sigma}_{g_y}^2 \alpha_t}{b} + \frac{\hat{\sigma}_{f_y}^2 \alpha_t^2}{b} + \frac{\hat{\sigma}_{g_x}^2 \alpha_t}{b} + \frac{\hat{\sigma}_{f_x}^2 \alpha_t^2}{b} \right) \right) \\ & + \frac{2}{\alpha_T} \mathcal{O} \left(\sum_{t=1}^T \left(\alpha_t + \frac{1}{\alpha_t} \right) (\rho_s^2 + \rho_r^2) + \left(1 + \frac{1}{\alpha_T} \right) T(d_2^2 \rho_r^2 + d_1^2 \rho_s^2) \right) \\ & + \frac{2}{\alpha_T} \mathcal{O} (\theta_1^y + \theta_1^v + \hat{\sigma}^2) \\ & \leq \mathcal{O} \left((d_1 + d_2)^{3/4} T^{1/3} (V_T + D_{\mathbf{y}, T} + D_{\mathbf{x}, T} + G_{\mathbf{y}, T} + \Delta_1 + \hat{\sigma}^2) \right. \\ & \quad \left. + (d_1 + d_2)^{3/2} T^{2/3} (H_{2,T} + G_{\mathbf{v}, T} + G_{\mathbf{x}, T}) \right), \end{aligned} \quad (252)$$

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

$$\begin{aligned}
&\frac{1}{2} \sum_{t=1}^T \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\
&\leq \sum_{t=1}^T \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\
&+ \sum_{t=1}^T \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) - \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right].
\end{aligned}$$

From non-expansiveness of the projection operator and Lemma D.4, we have

$$\begin{aligned}
&\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) - \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \\
&\leq \left\| \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) - \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t)) \right\|^2 \\
&\leq \frac{(\rho_s d_1 + \rho_r d_2)^2 \ell_{f,1}^2}{4} \\
&\leq \frac{(\rho_s^2 d_1^2 + \rho_r^2 d_2^2) \ell_{f,1}^2}{2}.
\end{aligned}$$

This implies

$$\begin{aligned}
&\frac{1}{2} \sum_{t=1}^T \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\
&\leq \sum_{t=1}^T \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_{t, \rho}(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] + \frac{T(\rho_s^2 d_1^2 + \rho_r^2 d_2^2) \ell_{f,1}^2}{2}.
\end{aligned}$$

Applying the upper bound in (252) yields

$$\begin{aligned}
&\frac{1}{2} \sum_{t=1}^T \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\
&\leq \mathcal{O} \left((d_1 + d_2)^{3/4} T^{1/3} (V_T + D_{\mathbf{y}, T} + D_{\mathbf{x}, T} + G_{\mathbf{y}, T} + \Delta_1 + \hat{\sigma}^2) \right. \\
&\quad \left. + (d_1 + d_2)^{3/2} T^{2/3} (H_{2, T} + G_{\mathbf{v}, T} + G_{\mathbf{x}, T}) \right) \\
&+ \frac{T(\rho_s^2 d_1^2 + \rho_r^2 d_2^2) \ell_{f,1}^2}{2}.
\end{aligned}$$

Thus, from $\rho_r^2 = \frac{1}{d_2^2 T}$ and $\rho_s^2 = \frac{1}{d_1^2 T}$ in (31), we get

$$\begin{aligned} & \sum_{t=1}^T \mathbb{E} \left[\left\| \mathcal{P}_{\mathcal{X}, \alpha_t}(\mathbf{x}_t; \nabla f_t(\mathbf{x}_t, \mathbf{y}_t^*(\mathbf{x}_t))) \right\|^2 \right] \\ & \leq \mathcal{O} \left((d_1 + d_2)^{3/4} T^{1/3} (V_T + D_{\mathbf{y}, T} + D_{\mathbf{x}, T} + G_{\mathbf{y}, T} + \Delta_1 + \hat{\sigma}^2) \right. \\ & \quad \left. + (d_1 + d_2)^{3/2} T^{2/3} (H_{2, T} + G_{\mathbf{v}, T} + G_{\mathbf{x}, T}) \right). \end{aligned}$$

This completes the proof. $\square$

E Hyperparameter Tuning Results

As detailed in Section 4, we carefully tuned all hyperparameters to ensure stable and fair comparisons. Our analysis indicates that while ZO-SOGD exhibits sensitivity to hyperparameter choices, it remains robust within reasonable ranges. Below, we provide extensive tuning results for ZO-SOGD.

The hyperparameter sensitivity analysis for the adversarial attack scenario reveals critical insights about the algorithm’s attack effectiveness across different parameter configurations. For the inner and outer stepsizes, we observe that the algorithm achieves optimal attack performance with specific combinations that balance perturbation strength and imperceptibility.

Table 2: Hyperparameter tuning results for inner (β) and outer (α) stepsizes in adversarial attack scenario. Values represent test accuracy (mean $\pm$ std) over 5 runs. Lower values indicate better attack performance.

$\beta \backslash \alpha$	$\alpha = 0.001$	$\alpha = 0.005$	$\alpha = 0.01$	$\alpha = 0.1$
$\beta = 0.001$	0.68 ± 0.05	0.59 ± 0.07	0.47 ± 0.06	0.53 ± 0.08
$\beta = 0.005$	0.54 ± 0.06	0.41 ± 0.05	0.35 ± 0.04	0.42 ± 0.05
$\beta = 0.01$	0.48 ± 0.04	0.34 ± 0.05	0.57 ± 0.07	0.39 ± 0.06
$\beta = 0.1$	0.26 ± 0.03	0.43 ± 0.06	0.33 ± 0.04	0.45 ± 0.07

The stepsize analysis reveals that larger inner stepsizes combined with smaller outer stepsizes tend to produce more effective attacks. Specifically, the configuration with $\beta = 0.1$ and $\alpha = 0.001$ achieves the lowest test accuracy of 0.26 ± 0.03 , indicating the most successful adversarial perturbations. This pattern suggests that aggressive updates to the perturbation parameters (β) while maintaining conservative hyperparameter updates (α) creates an effective balance for generating strong yet imperceptible adversarial examples.

Table 3: Performance comparison across different smoothing parameters ($\rho_r = \rho_s$) in adversarial attack scenario.

$\rho_v \backslash \rho_r = \rho_s$	0.001	0.005	0.01	0.05
$\rho_v = 0.001$	0.61 ± 0.06	0.52 ± 0.05	0.48 ± 0.04	0.57 ± 0.06
$\rho_v = 0.005$	0.47 ± 0.05	0.39 ± 0.04	0.35 ± 0.04	0.45 ± 0.05
$\rho_v = 0.01$	0.41 ± 0.04	0.28 ± 0.03	0.31 ± 0.03	0.43 ± 0.05
$\rho_v = 0.05$	0.53 ± 0.06	0.44 ± 0.05	0.40 ± 0.04	0.52 ± 0.06

The smoothing parameter analysis provides additional insights into the algorithm’s convergence behavior in the adversarial setting. The optimal configuration occurs with $\rho_v = 0.01$ and $\rho_r = \rho_s = 0.005$, achieving a test accuracy of 0.28 ± 0.03 . These moderate smoothing values appear to provide the right balance between exploration and exploitation in the adversarial perturbation space, allowing the algorithm to find effective attack directions without excessive oscillation or premature convergence.

Table 4: Performance comparison across different momentum parameters in adversarial attack scenario.

$\gamma_t \backslash \lambda_t = \eta_t$	0.9	0.99	0.999
$\gamma_t = 0.9$	0.35 ± 0.04	0.29 ± 0.03	0.38 ± 0.05
$\gamma_t = 0.99$	0.31 ± 0.03	0.24 ± 0.02	0.33 ± 0.04
$\gamma_t = 0.999$	0.37 ± 0.04	0.32 ± 0.08	0.40 ± 0.05

The momentum parameter investigation reveals that moderate momentum values consistently produce the most effective adversarial attacks. The optimal configuration with $\gamma_t = 0.99$ and $\lambda_t = \eta_t = 0.99$ achieves the lowest test accuracy of 0.24 ± 0.02 , representing the most successful attack performance. This configuration suggests that maintaining momentum across both inner and outer optimization loops helps the algorithm navigate the complex adversarial landscape more effectively than either no momentum or excessive momentum settings.

The comprehensive analysis demonstrates that ZO-SOGD maintains robust attack performance across a broad range of hyperparameter configurations. The algorithm consistently achieves test accuracies below 0.5 across most reasonable parameter combinations, indicating reliable adversarial attack capability. The standard deviations remain low throughout the parameter space, suggesting stable and reproducible attack performance across multiple experimental runs.

The optimal hyperparameter configuration for adversarial attacks consists of inner stepsize $\beta = 0.1$, outer stepsize $\alpha = 0.001$, smoothing parameters $\rho_v = 0.01$ and $\rho_r = \rho_s = 0.005$, and momentum parameters $\gamma_t = \lambda_t = \eta_t = 0.99$. This configuration enables ZO-SOGD to achieve superior attack performance while maintaining the imperceptibility constraints essential for practical adversarial examples.

NeurIPS Paper Checklist

A. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: We provide detailed proofs and implementations of the algorithms.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

B. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Limitations are provided in the Conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

C. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Please refer to Theorems 2.6 and 3.2.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

D. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: It is provided in Section 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

E. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data is publicly available. For experimental results, please refer to Section 4 and the supplement.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

F. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: For experimental results, please refer to Section 4 and the supplement.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

G. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Please refer to Section 4 and the supplement.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

H. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA] .

Justification: All experiments were conducted on the same system and are easily reproducible on a standard personal computer.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

I. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: The authors reviewed the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

J. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA] .

Justification: Theory Paper.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

K. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA] .

Justification: Theory Paper.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

L. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes] .

Justification: The authors cite the original paper that produced the code package or dataset.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

M. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA] .

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

N. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA] .

Justification: Theory paper.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

O. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA] .

Justification: Paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

P. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA] .

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.