Fully First-Order Algorithms for Online Non-Convex Bilevel Optimization

Tingkai Jia **Ting Wang** **Cheng Chen***
MoE Engineering Research Center of HW/SW Co-design Technology and Application
East China Normal University
{51275902086}@stu.ecnu.edu.cn {twang,chchen}@sei.ecnu.edu.cn

Abstract

In this work, we study non-convex-strongly-convex online bilevel optimization (OBO) using only first-order oracle. Existing OBO algorithms are mainly based on hypergradient descent, which requires access to a Hessian-vector product (HVP) oracle and potentially incurs high computational costs. By reformulating the original OBO problem as a single-level online problem with inequality constraints and constructing a sequence of Lagrangian function, we eliminate the need for HVPs arising from implicit differentiation. Specifically, we propose a fully first-order algorithm for OBO, and provide theoretical guarantees showing that it achieves regret of $O(1 + V_T + H_{2,T})$ with a total of $O(T \log T)$ iterations, where V_T measures the variation in function values and $H_{2,T}$ characterizes the drift variation of the inner-level optimal solution. We also establish a sublinear regret bound under the single-loop structure by introducing additional gradient-variation terms. Furthermore, we develop an improved variant with an adaptive inner-iteration scheme, which removes the dependence on $H_{2,T}$ and achieves regret of $O(\log T + V_T)$. Finally, under the stochastic OBO setting, we establish the regret bound for the fully first-order algorithm, i.e., $O(T^{2/3}(1 + \sigma^2) + V_T + H_{2,T})$. Numerical experiments demonstrate the feasibility of our algorithm and support our theoretical findings.

1 Introduction

Online bilevel optimization (OBO) has attracted growing attention in recent years. It extends bilevel learning to streaming settings, where the underlying data distribution may drift over time. The goal is to generate a sequence of decisions $\{\mathbf{x}_t\}_{t=1}^T$ that tracks the time-varying optimal solutions $\{\mathbf{x}_t^*\}_{t=1}^T$ over the horizon $t \in [T]$ by solving

$$\min_{\mathbf{x}\in\mathcal{X}} F_t(\mathbf{x}) := f_t\left(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})\right) \quad \text{s.t.} \quad \mathbf{y}_t^*(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{y}\in\mathbb{R}^{d_2}} g_t(\mathbf{x}, \mathbf{y}), \tag{P1}$$

where the upper-level function $f_t : \mathcal{X} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is smooth and possibly non-convex, while the lower-level function $g_t : \mathcal{X} \times \mathbb{R}^{d_2} \to \mathbb{R}$ is smooth and μ_g-strongly convex with respect to $\mathbf{y}$, which ensures the existence of a unique minimizer $\mathbf{y}_t^*(\mathbf{x})$ for all given $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{d_1}$.

Existing OBO methods [31, 18, 2, 14] build on hypergradient descent, which requires an inner iterative procedure to approximately solve the lower-level problem, and then update the upper decision $\mathbf{x}$ using the hypergradient $\nabla F_t(\mathbf{x})$. Achieving a sufficiently accurate hypergradient typically requires many Hessian-vector products (HVP) to control the approximation error [8, 26]. However, this reliance on HVPs may limit the practical applicability of existing methods, particularly when second-order computations are prohibitively expensive or unavailable.

*Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of algorithms on bilevel local regret without window averaging as defined in Eq.(2) in terms of gradient queries, required Hessian-vector product (HVP) queries, fully single-loop structure, local regret bounds, and corresponding theorems. V_T and $H_{2,T}$ are defined in Eq.(4). $E_{2,T}$ and P_T are defined in Eq.(9) and Theorem 3.2, respectively.

Algorithm	**Grad. Query**	**HVP Query**	**Fully S-Loop**	**Local Regret Bound**
SOBOW [18]	$O(T)$	$O(T^2)$	✗	$O(1+V_T+H_{2,T})$
OBBO [2]	$O(T\log T)$	$O(T\log T)$	✗	$O(1+V_T+H_{2,T})$
FSOBO [14]	$O(T)$	$O(T)$	✓	$O(1+V_T+H_{2,T}+E_{2,T})$
AOBO [14]	$O(T\log T)$	$O(T\log T)$	✗	$O(1+V_T)$
Algorithm 1	$O(T\log T)$	None	✗	$O(1+V_T+H_{2,T})$
Algorithm 1	$O(T)$	None	✓	$O(T^{1/3}(1+V_T)+T^{2/3}P_T)$
Algorithm 2	$O(T^2+TH_{2,T})$	None	✗	$O(\log T+V_T)$

Table 2: Comparison of stochastic algorithms on bilevel local regret in terms of sample complexities, required HVP term numbers, window-averaged regret form, local regret bounds, and corresponding theorems. In SOGD, $\Delta_T = O(1)+V_T$, $\Psi_T = P_T + E^f_{\mathbf{x},T} + E^g_{\mathbf{xy},T} + E^g_{\mathbf{yy},T}$, where $E^f_{\mathbf{x},T} := \sum_{t=2}^{T}\sup_{\mathbf{z}}\|\nabla_{\mathbf{x}} f_t(\mathbf{z}) - \nabla_{\mathbf{x}} f_{t-1}(\mathbf{z})\|^2$, $E^g_{\mathbf{xy},T} := \sum_{t=2}^{T}\sup_{\mathbf{z}}\|\nabla^2_{\mathbf{xy}} g_t(\mathbf{z}) - \nabla^2_{\mathbf{xy}} g_{t-1}(\mathbf{z})\|^2$.

Algorithm	**Sample Comp.**	**HVP Query**	**WinAvg.**	**Local Regret Bound**
SOBBO [2]	$O(wT)$	$O(T\log w)$	✓	$O(\frac{T}{w}(1+\sigma^2)+V_T+H_{2,T})$
SOGD [24]	$O(T)$	$O(T)$	✗	$O(T^{1/3}(\sigma^2+\Delta_T)+T^{2/3}\Psi_T)$
Algorithm 3	$O(T^{5/3}\log T)$	None	✗	$O(T^{2/3}(1+\sigma^2)+V_T+H_{2,T})$

For example, in large-scale bilevel learning with high-dimensional neural networks, repeated second-order computations and differentiation through long inner optimization trajectories can incur substantial runtime and memory overhead [30]. Similar difficulties arise when the lower-level problem contains millions of variables or when only a limited number of inner updates can be afforded under strict computational budgets [23]. Moreover, in stochastic-policy or reinforcement-learning settings, second-order estimators can also be difficult to implement and may suffer from high variance [28]. Motivated by these challenges, this work studys fully first-order OBO methods that avoid HVPs while maintaining favorable theoretical guarantees.

1.1 Our Contributions

We propose a fully first-order online bilevel optimizer (F^{2}OBO), presented in Algorithm 1. Theoretical analysis shows that it can achieve an $O(1+V_T+H_{2,T})$ upper bound on the standard bilevel local regret, without requiring any second-order information. Moreover, our method supports fully single-loop structure, which performs only one inner-step update per t. By introducing additional regularities of environmental variation, we can still guarantee sublinear regret, i.e., $O(T^{1/3}(1+V_T)+T^{2/3}P_T)$.

Building on the adaptive-iteration strategy in AOBO [14], we further enhance F^{2}OBO in a similar spirit to propose adaptive-iteration fully first-order online bilevel optimizer (AF^{2}OBO), presented in Algorithm 2. Theoretical analysis shows that it eliminates the dependence on the drift variation $H_{2,T}$ of the inner-level optimal solution. Although this comes at the cost of a larger iteration overhead and a weaker regret guarantee, i.e., $O(\log T+V_T)$ with an additional fixed $O(\log T)$ term, our method still ensures sublinear regret even when $\mathbf{y}^*_t(x)$ drifts drastically (e.g., $H_{2,T}=O(T)$ or larger), as shown in Table 1. This robustness is significant compared with static inner-iteration methods.

Finally, we extend F^{2}OBO to the stochastic OBO problem and propose stochastic fully first-order online bilevel optimizer (SF^{2}OBO), presented in Algorithm 3. Compared with SOBBO, we reduce the variance of the biased gradient by appropriately increasing the batch size, thereby eliminating the reliance on window-averaged regret and obtaining $O(T^{2/3}(1+\sigma^2)+V_T+H_{2,T})$. Compared

with SOGD, our result imposes a milder requirement on environmental variation, i.e., when $V_T = H_{2,T} = O(T^{2/3})$, SF^{2}OBO still achieves sublinear regret, whereas SOGD suffers at least $O(T^{4/3})$ superlinear regret. A comparison of the regret bounds is presented in Table 2.

1.2 Related Work

Bilevel Optimization Bilevel optimization (BO) was originally introduced by Bracken and McGill [4]. Existing BO methods mainly include hypergradient-based methods, penalty-based methods, and Hessian- or Jacobian-free methods. Hypergradient-based methods are typically divided into approximate implicit differentiation (AID) [9, 8, 10, 12, 25] and iterative differentiation (ITD) approaches [22, 7, 26, 21, 12]. Penalty-based methods reformulate BO as a constrained single-level problem using inner-level optimality conditions, thereby avoiding implicit-gradient computation and allowing extensions to non-strongly convex inner problems under suitable constraint qualification conditions [27, 19, 11, 1, 20, 5]. Hessian- or Jacobian-free methods further avoid explicit second-order derivatives by estimating hypergradients with zeroth-order approximations [32, 13, 33].

Online Bilevel Optimization. Online bilevel optimization (OBO) has attracted growing attention in recent years. Early studies by Tarzanagh et al. [31] and Lin et al. [18] initiated the investigation of OBO and developed a window-averaged hypergradient framework. Building on this line of work, Bohne et al. [2] employed iterative differentiation (ITD) to compute hypergradients and proposed the OBBO algorithm, together with its stochastic counterpart SOBBO. Later, Nazari et al. [24] introduced a momentum-based search direction and developed SOGD for the stochastic OBO setting. More recently, Jia et al. [14] revisited bilevel local regret without window averaging, proposed the first adaptive-inner-iteration algorithm, and established optimal guarantees. They further introduced a new window-averaging analysis framework that removes the need for sublinear variation assumptions. In addition, Bohne et al. [3] studied non-stationary functional bilevel optimization, which can also be interpreted within the existing OBO framework.

2 Preliminaries

2.1 Notations and Assumptions

We use $\|\cdot\|$ to denote the Euclidean norm. Given the inner-level function g_t in problem (P1), its partial gradients with respect to the upper- and inner-level variables are denoted by $\nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y})$ and $\nabla_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y})$. The mixed derivative $\nabla^2_{\mathbf{xy}} g_t(\mathbf{x}, \mathbf{y})$ is defined as the Jacobian of $\nabla_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y})$ with respect to $\mathbf{x}$, while $\nabla^2_{\mathbf{yy}} g_t(\mathbf{x}, \mathbf{y})$ denotes the Hessian with respect to $\mathbf{y}$. For the composition $f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))$, we use $\nabla f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))$ to denote its total derivative with respect to $\mathbf{x}$. Finally, the condition number of the inner-level objective with respect to $\mathbf{y}$ is denoted by $\kappa_g := L_{g,1}/\mu_g$.

We make the following assumptions on the upper- and inner-level functions. These assumptions are also commonly used in existing OBO studies [18, 24, 2] and are standard in this setting.

Assumption 2.1. For all $t \in [T]$, f_t is $L_{f,0}$-Lipschitz continuous, and $\nabla f_t(\mathbf{x}, \mathbf{y})$ is $L_{f,1}$-Lipschitz continuous. Moreover, there exists a positive constant M such that $|f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))| \le M, \quad \forall \mathbf{x} \in \mathcal{X}$.

Assumption 2.2. For all $t \in [T]$, g_t satisfies the following conditions. For any given $\mathbf{x} \in \mathcal{X}$, $g_t(\mathbf{x}, \mathbf{y})$ is μ_g-strongly convex in $\mathbf{y}$. In addition, $\nabla g_t(\mathbf{x}, \mathbf{y})$ is $L_{g,1}$-Lipschitz continuous, and both $\nabla^2_{\mathbf{xy}} g_t(\mathbf{x}, \mathbf{y})$ and $\nabla^2_{\mathbf{yy}} g_t(\mathbf{x}, \mathbf{y})$ are $L_{g,2}$-Lipschitz continuous.

For stochastic OBO, we additionally give the following assumption to bound the variance of inner- and upper-level first-order gradients, which is common for stochastic regret analysis.

Assumption 2.3. For all $t \in [T]$, the stochastic gradients $\nabla f_t(\mathbf{x}, \mathbf{y}; \xi)$, $\nabla g_t(\mathbf{x}, \mathbf{y}; \zeta)$ are unbiased:

$$\mathbb{E}_\xi \left[\nabla f_t(\mathbf{x}, \mathbf{y}; \xi)\right] = \nabla f_t(\mathbf{x}, \mathbf{y}), \quad \mathbb{E}_\zeta \left[\nabla g_t(\mathbf{x}, \mathbf{y}; \zeta)\right] = \nabla g_t(\mathbf{x}, \mathbf{y}).$$

Moreover, there exists $\sigma > 0$ such that their conditional variances are uniformly bounded:

$$\mathbb{E}\left[\left\|\nabla f_t(\mathbf{x}, \mathbf{y}; \xi) - \nabla f_t(\mathbf{x}, \mathbf{y})\right\|^2\right] \le \sigma^2, \quad \mathbb{E}\left[\left\|\nabla g_t(\mathbf{x}, \mathbf{y}; \zeta) - \nabla g_t(\mathbf{x}, \mathbf{y})\right\|^2\right] \le \sigma^2.$$

2.2 Single-Level Reformulation of OBO Problem

For non-convex-strongly-convex OBO, the uniqueness of the inner-level solution $\mathbf{y}_t^*(\mathbf{x})$ for any $\mathbf{x} \in \mathcal{X}$ yields a unique closed-form expression for the hypergradient of F_t, given by

$$\begin{aligned}\nabla F_t(\mathbf{x}) =& \nabla_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) + \nabla \mathbf{y}_t^*(\mathbf{x}) \nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) \\ =& \nabla_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) - \nabla^2_{\mathbf{xy}} g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) \left[\nabla^2_{\mathbf{yy}} g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))\right]^{-1} \nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})). \end{aligned} \tag{1}$$

We consider the following local regret notion as the evaluation criterion for the algorithms:

$$\mathrm{BLReg}(T) = \sum_{t=1}^{T} \left\| \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \nabla F_t(\mathbf{x}_t), \gamma \right) \right\|^2. \tag{2}$$

Here, $\mathcal{G}_{\mathcal{X}}$ denotes the gradient mapping with respect to the constraint set $\mathcal{X}$, defined as

$$\mathcal{G}_{\mathcal{X}} \left(\mathbf{x}, \mathbf{g}, \gamma \right) := \frac{1}{\gamma} \left(\mathbf{x} - \operatorname*{argmin}_{\mathbf{u} \in \mathcal{X}} \left\{ \langle \mathbf{g}, \mathbf{u} \rangle + \frac{1}{2\gamma} \|\mathbf{u} - \mathbf{x}\|^2 \right\} \right), \tag{3}$$

for given $\mathbf{g} \in \mathbb{R}^{d_1}$ and $\gamma > 0$. We mainly use the following two variation regularities to establish our regret bound, which are also used in Lin et al. [18], Bohne et al. [2] and Nazari et al. [24]:

$$V_T := \sum_{t=2}^{T} \sup_{\mathbf{x} \in \mathcal{X}} |F_{t-1}(\mathbf{x}) - F_t(\mathbf{x})| \quad \text{and} \quad H_{2,T} := \sum_{t=2}^{T} \sup_{\mathbf{x} \in \mathcal{X}} \left\| \mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x}) \right\|^2. \tag{4}$$

Following previous works [15, 29, 19, 6], the original problem (P1) can be equivalently written as an inequality constrained single-level problem, shown as:

$$\min_{\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathbb{R}^{d_2}} f_t(\mathbf{x}, \mathbf{y}) \quad \text{s.t.} \quad g_t(\mathbf{x}, \mathbf{y}) - g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) \leq 0. \tag{P2}$$

By treating the constraint as a penalty term, we construct the following Lagrangian function $\mathcal{L}_t$ for some $\lambda_t > 0$:

$$\mathcal{L}_t(\mathbf{x}, \mathbf{y}, \lambda_t) = f_t(\mathbf{x}, \mathbf{y}) + \lambda_t \left(g_t(\mathbf{x}, \mathbf{y}) - g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) \right). \tag{5}$$

We can finally approximately transform problem (P2) into an min-max optimization problem for $\mathcal{L}_t^*(\mathbf{x})$ as follows:

$$\min_{\mathbf{x} \in \mathcal{X}} \mathcal{L}_t^*(\mathbf{x}) := \mathcal{L}_t(\mathbf{x}, \mathbf{y}_{\lambda,t}^*(\mathbf{x}), \lambda_t) \quad \text{s.t.} \quad \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}) = \operatorname{argmin}_{\mathbf{y} \in \mathbb{R}^{\mathrm{d}_2}} \mathcal{L}_{\mathrm{t}}(\mathbf{x}, \mathbf{y}, \lambda_{\mathrm{t}}). \tag{P3}$$

Here the multiplier λ_t represents the degree to which the solution to (P3) conforms to the constraint, and we have $\lim_{\lambda_t \to \infty} \mathcal{L}_t^*(\mathbf{x}) = f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))$.

Based on the above, we approximate the OBO problem by an online min-max optimization problem. The ingenuity of this approach lies in its elimination of the requirement for a second-order gradient oracle arising from the implicit term $\nabla \mathbf{y}_t^*(\mathbf{x})$ in Eq.(1). Specifically, we have

$$\begin{aligned}\nabla \mathcal{L}_t^*(\mathbf{x}) =& \nabla_{\mathbf{x}} \mathcal{L}_t(\mathbf{x}, \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}), \lambda_t) + \nabla \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}) \nabla_{\mathbf{y}} \mathcal{L}_t(\mathbf{x}, \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}), \lambda_t) \\ =& \nabla_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}_{\lambda_t,t}^*(\mathbf{x})) + \lambda_t \left(\nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}_{\lambda_t,t}^*(\mathbf{x})) - \nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) \right), \end{aligned} \tag{6}$$

note that $\nabla_{\mathbf{y}} \mathcal{L}_t(\mathbf{x}, \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}), \lambda_t) = 0$ holds for all $t \in [T]$ and any $\lambda_t > 0$, thus the hypergradient does not require the calculation of $\nabla \mathbf{y}_{\lambda_t,t}^*(\mathbf{x})$.

3 A Fully First-Order Algorithm for Online Bilevel Optimization

This section develops and analyzes a fully first-order algorithm for determistic OBO.

3.1 Fully First-order Online Bilevel Optimizer (F^{2}OBO)

Similar to existing OBO algorithms [31, 18, 2], we still attempt to apply online gradient descent (OGD) to problem (P3). Consider a sequence $\{\lambda_t\}_{t=1}^T$ that governs our approximation, the hypergradient $\nabla \mathcal{L}_t^*(\mathbf{x}_t)$ is calculated at the current iterate $\mathbf{x}_t$ and used to update $\mathbf{x}_{t+1}$ at each time step t. Therefore,

Algorithm 1 Fully First-order Online Bilevel Optimizer (F^{2}OBO)

Input: $\mathbf{x}_1, \mathbf{y}_1, \mathbf{z}_1, \{\lambda_t, \beta_t, \gamma_t, K_t\}_{t=1}^T, \alpha$

for $t = 1$ to T **do**
  Output $\mathbf{x}_t$ and $\mathbf{y}_t$, receive f_t and g_t.
  warm start $\mathbf{y}_t^1 \leftarrow \mathbf{y}_t$, $\mathbf{z}_t^1 \leftarrow \mathbf{z}_t$
  for $k = 1$ to K_t **do**
    $\mathbf{z}_t^{k+1} \leftarrow \mathbf{z}_t^k - \alpha \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{z}_t^k)$
    $\mathbf{y}_t^{k+1} \leftarrow \mathbf{y}_t^k - \beta_t \left(\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^k) + \lambda_t \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t^k)\right)$
  end for
  update $\mathbf{y}_{t+1} \leftarrow \mathbf{y}_t^{K_t+1}$, $\mathbf{z}_{t+1} \leftarrow \mathbf{z}_t^{K_t+1}$
  calculate $\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) = \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_{t+1}) + \lambda_t \left(\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{z}_{t+1})\right)$
  $\mathbf{x}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{u} \in \mathcal{X}} \left\{ \left\langle \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \mathbf{u} \right\rangle + \frac{1}{2\gamma_t} \|\mathbf{u} - \mathbf{x}_t\|^2 \right\}$
  $\lambda_{t+1} \leftarrow \Lambda(t, \lambda_t)$
end for

according to Eq.(6), we need to solve two subproblems to obtain $\mathbf{y}_t^*(\mathbf{x}_t)$ and $\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)$. We can solve both subproblems simultaneously via a nested inner loop of gradient descent, starting from $\mathbf{z}_t^1 = \mathbf{z}_t$ and $\mathbf{y}_t^1 = \mathbf{y}_t$ for K times, as shown below:

$$\mathbf{z}_t^{k+1} \leftarrow \mathbf{z}_t^k - \alpha \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{z}_t^k) \quad \text{and} \quad \mathbf{y}_t^{k+1} \leftarrow \mathbf{y}_t^k - \beta \nabla_{\mathbf{y}} \mathcal{L}_t(\mathbf{x}_t, \mathbf{y}_t^k, \lambda_t). \tag{7}$$

This simple strategy yields excellent theoretical performance because both g_t and $\mathcal{L}_t$ is strongly convex and smooth with respect to $\mathbf{y}$ under Assumptions 2.1, 2.2 and if set $\lambda_t \geq 2L_{f,1}/\mu_g$ for all $t \in [T]$, which can also be seen in several offline works [15, 6]. Finally with $\mathbf{z}_{t+1} = \mathbf{z}_t^{K+1}$ and $\mathbf{y}_{t+1} = \mathbf{y}_t^{K+1}$, the approximated hypergradient $\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)$ can be calculated by

$$\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) := \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_{t+1}) + \lambda_t \left(\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{z}_{t+1})\right).$$

The multiplier λ_t must be carefully chosen: a small λ_t may cause the solution of (P3) to deviate from that of (P1), leading to accumulated bias, while a large λ_t may induce overly large gradients in Eq.(7) and destabilize optimization [15]. To balance this trade-off, we update $\lambda_{t+1} = \Lambda(t, \lambda_t)$ via a designed non-decreasing schedule, yielding the final fully first-order online bilevel optimizer (F^{2}OBO) in Algorithm 1. The following theorem gives a regret upper bound of F^{2}OBO.

Theorem 3.1. *Under Assumptions 2.1, 2.2, let* $\Lambda(t, \lambda_t) = \left(1 + \frac{1}{t}\right)^\tau \lambda_t$ *for some* $\tau > 0$, $\lambda_1 > \frac{2L_{f,1}}{\mu_g}$, $\alpha \leq \frac{1}{L_{g,1}}$, $\beta_t \leq \frac{1}{2\lambda_t L_{g,1}}$, $\gamma_t \equiv \gamma \leq \min\{\frac{1}{2L_F}, \frac{1}{16\kappa_g}\sqrt{\frac{1-\rho}{C_\lambda}}\}$ *and* $K_t \equiv K \geq \max\{1 - \frac{\ln 2}{\ln \rho}, \frac{\ln c - 2\tau \ln T}{\ln \rho}\}$ *for some constants* $\rho \in (0, 1)$, C_λ *and* $c > 0$, *Algorithm 1 can guarantee*

$$\mathrm{BLReg}(T) \leq O\Big(\sum_{t=1}^T \frac{1}{t^{2\tau}} + V_T + H_{2,T}\Big).$$

Theorem 3.1 shows that our regret bound mainly depends on two variation regularities, V_T and $H_{2,T}$. Specifically, if we set $\tau > 1/2$, then we achieve a regret of $O(1 + V_T + H_{2,T})$, which is comparable to SOBOW [18] and OBBO [2]. Our algorithm requires a time complexity of $O(T \log T)$, which is on par with OBBO and better than that of SOBOW.

3.2 Single-Loop F^{2}OBO

Although sufficiently many inner-loop iterations can yield favorable theoretical guarantees, it may be difficult to deploy in practice due to the substantial computational overhead, particularly in time-critical settings where performing multiple inner-loop iterations per round may be impractical. Here we proof that F^{2}OBO still achieves sublinear regret under the single-loop structure, i.e., $K_t = 1$. To control the accumulated error aused by inexactly solving the inner subproblems, we introduce the following two variation regularities to further restrict the first-order gradient changes:

$$E_{\mathbf{y},T}^f = \sum_{t=2}^T \sup_{\mathbf{x},\mathbf{y}} \|\nabla_{\mathbf{y}} f_t(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} f_{t-1}(\mathbf{x}, \mathbf{y})\|^2, \quad E_{\mathbf{y},T}^g = \sum_{t=2}^T \sup_{\mathbf{x},\mathbf{y}} \|\nabla_{\mathbf{y}} g_t(\mathbf{x}, \mathbf{y}) - \nabla_{\mathbf{y}} g_{t-1}(\mathbf{x}, \mathbf{y})\|^2.$$

We believe that these additional terms are necessary in this case. Similar ideas have been adopted in the analyses of existing second-order single-loop methods, such as FSOBO [14] and SOGD [24].

Theorem 3.2. *Under Assumptions 2.1, 2.2, let* $\Lambda(t,\lambda_t)=\left(1+\frac{1}{t}\right)^{\tau}\lambda_t$ *for some* $\tau\in(0,\frac{1}{2})$, $\lambda_1>\frac{2L_{f,1}}{\mu_g}$, $\alpha\le\frac{1}{L_{g,1}}$, $\beta_t=\frac{1}{2\lambda_t L_{g,1}}$, $\gamma_t<\min\{\frac{1}{2L_F},\frac{1}{8\lambda_1 t^{\tau}}\sqrt{\frac{1-\rho}{3C_g}}\}$ *and* $K_t\equiv 1$ *for some constant* C_g, *with* $P_T:=H_{2,T}+E^f_{\mathbf{y},T}+E^g_{\mathbf{y},T}$, *Algorithm 1 can guarantee*

$$\mathrm{BLReg}(T)\le O\left(T^{\tau}(1+V_T)+T^{1-2\tau}+T^{2\tau}P_T\right).$$

Theorem 3.2 shows that SF^{2}OBO enjoys a sublinear regret guarantee under polynomially sublinear environmental variation. In particular, suppose that $V_T=O(T^{\alpha})$ and $P_T=O(T^{\beta})$ for some $\alpha,\beta\in[0,1)$, it suffices to choose τ such that $\tau+\alpha<1$, $2\tau+\beta<1$, and $\tau>0$. Equivalently, any $0<\tau<\min\{1-\alpha,(1-\beta)/2\}$ is valid. Since $\alpha,\beta<1$, this interval is always nonempty. Therefore, for any polynomially sublinear variation rates $V_T=O(T^{\alpha})$ and $P_T=O(T^{\beta})$, there exists a valid choice of τ such that every term in the regret bound is sublinear in T, yielding $\mathrm{BLReg}(T)=o(T)$. In particular, by setting $\tau=1/3$, we obtain

$$\mathrm{BLReg}(T)\le O\left(T^{1/3}(1+V_T)+T^{2/3}P_T\right), \tag{8}$$

a sublinear regret guarantee holds when $V_T=o(T^{2/3})$ and $P_T=o(T^{1/3})$.

Remark 3.3. Compared with the regret bound of FSOBO as shown in Table 1, Eq.(8) incurs an additional T-dependent factor, leading to a weaker guarantee. This degradation stems from the cross-problem error induced by fully first-order oracle feedback. In the single-loop setting, this error is excessively amplified, even with additional variation regularities to control it, we believe it is difficult to eliminate entirely. However, existing second-order single-loop algorithms typically require assuming sublinear variation of the Hessian or Jacobian of g_t. For example, FSOBO assumes

$$E_{2,T}=E^f_{\mathbf{y},T}+E^g_{\mathbf{yy},T},\quad\text{where}\quad E^g_{\mathbf{yy},T}=\sum_{t=2}^{T}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla^2_{\mathbf{yy}}g_t(\mathbf{x},\mathbf{y})-\nabla^2_{\mathbf{yy}}g_{t-1}(\mathbf{x},\mathbf{y})\right\|^2, \tag{9}$$

is sublinear. Such assumptions is harder to justify and guarantee. In contrast, our first-order algorithm only requires a sublinear first-order gradient changes, which is easier to verify and more practical.

4 An Adaptive Variant of F^{2}OBO

F^{2}OBO solves the two inner subproblems using a fixed number of iterations at each time step t. This static strategy is inevitably sensitive to changes, which is described by $H_{2,T}$, and may fail to guarantee sublinear regret when the lower-level optimal solutions exhibit linear or even superlinear drift. In this section, we consider improving F^{2}OBO by applying an adaptive inner iteration approach [14] to improve the regret upper bound.

4.1 Adaptive-iteration Fully First-order Online Bilevel Optimizer (AF^{2}OBO)

Considering (P3), $\mathcal{L}_t$ remains $(\lambda_t\mu_g/2)$-strongly convex and $(2\lambda_t L_{g,1})$-smooth if $\lambda_t\ge 2L_{f,1}/\mu_g$ for all $t\in[T]$. The distance between the current iterate and the optimum can be upper bounded by the norm of a computable gradient. Therefore, the tolerance parameters $\delta_{\mathbf{y}}$ and $\delta_{\mathbf{z}}$ can be used as stopping criteria for obtaining sufficiently accurate approximate solutions $\mathbf{y}_{t+1}$ and $\mathbf{z}_{t+1}$:

$$\frac{\lambda_t\mu_g}{2}\left\|\mathbf{y}_{t+1}-\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|\le\left\|\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x}_t,\mathbf{y}_{t+1},\lambda_t)\right\|\le\delta_{\mathbf{y}}\quad\text{and} \tag{10}$$

$$\mu_g\left\|\mathbf{z}_{t+1}-\mathbf{y}^*_t(\mathbf{x}_t)\right\|\le\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{z}_{t+1})\right\|\le\delta_{\mathbf{z}}. \tag{11}$$

We use two adaptive iterative processes to solve for $\mathbf{y}^*_t(\mathbf{x}_t)$ and $\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)$ respectively to constitute our adaptive-iteration fully first-order online bilevel optimizer (AF^{2}OBO), as shown in Algorithm 2. The following theorem provides the regret and computational cost guarantees of AF^{2}OBO.

Theorem 4.1. *Under Assumptions 2.1, 2.2, let* $\Lambda(t,\lambda_t)=\left(1+\frac{1}{t}\right)^{\tau}\lambda_t$ *for some* $\tau>0$, $\lambda_1>\frac{2L_{f,1}}{\mu_g}$, $\alpha\le\frac{1}{L_{g,1}}$, $\beta_t\le\frac{1}{2\lambda_t L_{g,1}}$, $\gamma\le\frac{1}{2L_F}$ *for all* $t\in[T]$, $\delta_{\mathbf{y}}=\frac{1}{\sqrt{T}}$ *and* $\delta_{\mathbf{z}}=\frac{1}{\sqrt{T^{1+2\tau}}}$, *with* $\mathcal{I}_T$ *representing*

Algorithm 2 Adaptive-iteration Fully First-order Online Bilevel Optimizer (AF^{2}OBO)

Input: $\mathbf{x}_1$, $\mathbf{y}_1$, $\mathbf{z}_1$, $\{\lambda_t, \beta_t\}_{t=1}^T$, α, γ, $\delta_{\mathbf{y}}$, $\delta_{\mathbf{z}}$

for $t = 1$ to T **do**
  Output $\mathbf{x}_t$ and $\mathbf{y}_t$, receive f_t and g_t.
  Set $\mathbf{y}_{t+1} \leftarrow \mathbf{y}_t$, $\mathbf{z}_{t+1} \leftarrow \mathbf{z}_t$
  while $\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{z}_{t+1})\| > \delta_{\mathbf{z}}$ **do**
    $\mathbf{z}_{t+1} \leftarrow \mathbf{z}_{t+1} - \alpha \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{z}_{t+1})$
  end while
  while $\|\nabla_{\mathbf{y}} \mathcal{L}_t(\mathbf{x}_t, \mathbf{y}_{t+1}, \lambda_t)\| > \delta_{\mathbf{y}}$ **do**
    $\mathbf{y}_{t+1} \leftarrow \mathbf{y}_{t+1} - \beta \left(\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_{t+1}) + \lambda_t \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_{t+1})\right)$
  end while
  calculate $\widetilde{\nabla} \mathcal{L}_t^*(\mathbf{x}_t) = \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_{t+1}) + \lambda_t \left(\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{z}_{t+1})\right)$
  $\mathbf{x}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{u} \in \mathcal{X}} \left\{ \left\langle \widetilde{\nabla} \mathcal{L}_t^*(\mathbf{x}_t), \mathbf{u} \right\rangle + \frac{1}{2\gamma} \|\mathbf{u} - \mathbf{x}_t\|^2 \right\}$
  $\lambda_{t+1} \leftarrow \Lambda(t, \lambda_t)$
end for

the total number of inner-loop iterations, Algorithm 2 can guarantee

$$\mathrm{BLReg}(T) \leq O\left(\sum_{t=1}^{T} \frac{1}{t^{2\tau}} + V_T\right) \quad \textit{and} \quad \mathcal{I}_T \leq O\left(T^{1+2\tau} + T^{2\tau} H_{2,T}\right).$$

As shown in Theorem 4.1, the main improvement of our method is that it eliminates the dependency of the regret upper bound on $H_{2,T}$, which describes the drift of the inner optimal solution $\|\mathbf{y}_t^*(\mathbf{x}) - \mathbf{y}_{t-1}^*(\mathbf{x})\|$ for any $\mathbf{x} \in \mathcal{X}$.

4.2 Discussion on the Improved Regret Bound

Note that AOBO [14] achieves the optimal bilevel local regret $\Omega(1 + V_T)$ with $O(T \log T)$ query of gradients and HVPs. Our method achieves the same bound if we set, i.e., $\tau = 1$, which requires $O(T^3)$ first-order gradient evaluations. If we choose $\tau = 1/2$, We can guarantee a sublinear regret of $O(\log T + V_T)$, which does not rely on $H_{2,T}$. In this setting, the total number of gradient quiries is at most $O(T^2 + T H_{2,T})$.

The upper bound in Theorem 4.1 have the advantage when the drift of $\mathbf{y}_t^*(\mathbf{x})$ is more severe. We provide a 1-dimensional example for illustration. Consider a set of $\{f_t, g_t\}_{t=1}^T$ within constraints $\mathcal{X} = \{x \,|\, x \in [-1, 1]\}$ and $y \in \mathbb{R}$ that satisfies $V_T = 0$ and $H_{2,T} = O(T)$:

$$f_t(x, y) \equiv \widetilde{f}(y) = c e^{-y^2} \quad \text{and} \quad g_t(x, y) = \frac{\mu_g}{2} \left(y - (-1)^t x\right)^2.$$

Our construction satisfies Assumptions 2.1, 2.2 when constant c is set appropriately. With $y_t^*(x) = (-1)^t x$, $f_t(x, y_t^*(x)) = \widetilde{f}(y_t^*(x)) \equiv c e^{-x^2}$, it holds that

$$V_T = \sum_{t=2}^{T} \sup_{x \in [-1,1]} \left|\widetilde{f}(y_t^*(x)) - \widetilde{f}(y_{t-1}^*(x))\right| = 0, \quad H_{2,T} = \sum_{t=2}^{T} \sup_{x \in [-1,1]} \left|(-1)^t x - (-1)^{t-1} x\right|^2 = 4T - 4.$$

In the case of oscillatory drift in the inner-level optimal solution, our AF^{2}OBO still guarantees a sublinear regret bound of $O(\log T)$ if $\tau = 1/2$, while the regret guarantees of static inner-iteration methods, such as OBBO and F^{2}OBO, will suffer from linear regret.

5 F^{2}OBO with Stochastic First-Order Gradient

In this section, we study the stochastic OBO problem with our fully first-order algorithm, where the objective functions are formulated as:

$$f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) = \mathbb{E}_{\xi \sim \mathcal{D}_t^f} \left[f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}); \xi)\right], \quad g_t(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\zeta \sim \mathcal{D}_t^g} \left[g_t(\mathbf{x}, \mathbf{y}; \zeta)\right],$$

where $\{\mathcal{D}_t^f, \mathcal{D}_t^g\}_{t=1}^T$ are the time-varying upper- and inner-level distributions, respectively.

Algorithm 3 Stochastic Fully First-order Online Bilevel Optimizer (SF^{2}OBO)

Input: $\mathbf{x}_1, \mathbf{y}_1, \{\lambda_t, \alpha_t, \beta_t, \gamma_t, K_t, B_t, \widehat{B}_t\}_{t=1}^T$

for $t = 1$ to T **do**
  Output $\mathbf{x}_t$ and $\mathbf{y}_t$, receive f_t and g_t.
  warm start $\mathbf{y}_t^1 \leftarrow \mathbf{y}_t$, $\mathbf{z}_t^1 \leftarrow \mathbf{z}_t$
  for $k = 1$ to K_t **do**
    draw samples $\xi_t^k \sim \mathcal{D}_t^f$ and $\zeta_t^k \sim \mathcal{D}_t^g$ with batch size B_t
    $\mathbf{z}_t^{k+1} \leftarrow \mathbf{z}_t^k - \alpha_t \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{z}_t^k; \zeta_t^k)$
    $\mathbf{y}_t^{k+1} \leftarrow \mathbf{y}_t^k - \beta_t \left(\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^k; \xi_t^k) + \lambda_t \nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t^k; \zeta_t^k)\right)$
  end for
  update $\mathbf{y}_{t+1} \leftarrow \mathbf{y}_t^{K_t+1}$, $\mathbf{z}_{t+1} \leftarrow \mathbf{z}_t^{K_t+1}$, draw samples $\xi_t \sim \mathcal{D}_t^f$, $\zeta_t \sim \mathcal{D}_t^g$ with batch size $\widehat{B}_t$
  calculate $\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t) = \nabla_{\mathbf{x}} f_t(\mathbf{x}_t, \mathbf{y}_{t+1}; \xi_t) + \lambda_t \left(\nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{y}_{t+1}; \zeta_t) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t, \mathbf{z}_{t+1}; \zeta_t)\right)$
  $\mathbf{x}_{t+1} \leftarrow \operatorname{argmin}_{\mathbf{u} \in \mathcal{X}} \left\{ \left\langle \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \mathbf{u} \right\rangle + \frac{1}{2\gamma_t} \|\mathbf{u} - \mathbf{x}_t\|^2 \right\}$
  $\lambda_{t+1} \leftarrow \Lambda(t, \lambda_t)$
end for

We replace the exact gradients in F^{2}OBO with stochastic gradients, and control the accumulated variance in the regret through the inner- and upper-level batch sizes B_t and $\widehat{B}_t$, respectively, for all $t \in [T]$. This leads to Algorithm 3. The following theorem establishes a regret bound.

Theorem 5.1. *Under Assumption 2.1, 2.2 and 2.3, Let* $\Lambda(t, \lambda_t) = \left(1 + \frac{1}{t}\right)^\tau \lambda_t$ *for some* $\tau \in (0, \frac{1}{2})$, $\lambda_1 > \frac{2L_{f,1}}{\mu_g}$, $\alpha_t = \frac{1}{L_{g,1} t^a}$, $\beta_t = \frac{1}{2\lambda_1 L_{g,1} t^b}$, $\gamma_t \equiv \gamma \le \min\{\frac{1}{2L_F}, \frac{1}{48\kappa_g L_{g,1}} \sqrt{\frac{1-\rho}{7c_\lambda}}\}$ *for some constant* $c_\lambda > 0$, $K_t = 4\kappa_g \max\{t^{b-\tau}, t^a\} \log t$, $\widehat{B}_t = t^{B_1}$ *and* $B_t = t^{B_2}$, *Algorithm 3 can guarantee*

$$\mathbb{E}\left[\mathrm{BLReg}(T)\right] \le O\left(T^{1-2\tau} + \sigma^2 \max\{T^{1+2\tau-B_1}, T^{1+3\tau-b-B_2}, T^{1+2\tau-a-B_2}\} + V_T + H_{2,T}\right).$$

As shown in Theorem 5.1, the three variance terms arise from the updates of $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$, respectively. The first $T^{1+2\tau-B_1}$ term is the main bottleneck, due to the bias of the gradient mapping and the dependence of $\nabla\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t)$ variance on λ_t. This makes the parameter selection more challenging, shown as $\mathbb{E}[\|\nabla\mathcal{L}_t^*(\mathbf{x}; \mathcal{S}_t) - \nabla\mathcal{L}_t^*(\mathbf{x})\|^2] \le (1 + 2\lambda_t^2)\sigma^2$.

Corollary 5.2. *Under Assumption 2.1, 2.2 and 2.3, with the same parameter setting in Theorem 5.1, let* $\tau = \frac{1}{6}, a = 0, b = \frac{1}{6}, B_1 = B_2 = \frac{2}{3}$, *which means* $K_t = 4\kappa_g \log t$, *Algorithm 3 can guarantee*

$$\mathbb{E}\left[\mathrm{BLReg}(T)\right] \le O\left(T^{2/3}(1 + \sigma^2) + V_T + H_{2,T}\right).$$

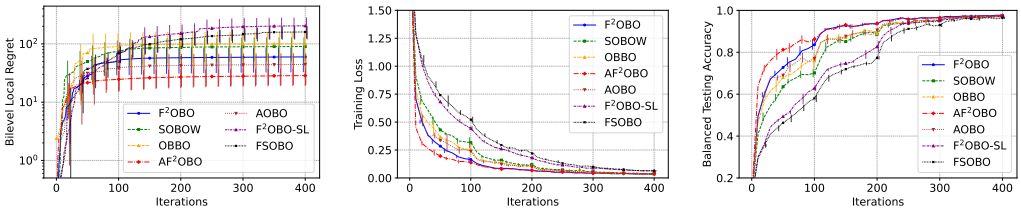
Corollary 5.2 gives a specific result of SF^{2}OBO. SOBBO [2] reduces the upper-level variance by using a window-averaged hypergradient, which yields a variance term of $O(T\sigma^2/w)$. However, this approach inherently relies on the analysis of window-averaged regret, which may not accurately characterize the algorithm's performance in dynamic environments. SOGD [24] establishes results under the standard local regret by employing variance reduction techniques and achieve a $O(T^{1/3}\sigma^2)$. However, its guarantees crucially depend on sublinear environmental variation assumptions as shown in Table 2, and additionally require a bounded variance assumption on stochastic second-order gradients, i.e., $\mathbb{E}\|\nabla_{\mathbf{xy}}^2 g_t(\mathbf{z}; \zeta) - \nabla_{\mathbf{xy}}^2 g_t(\mathbf{z})\|^2 \le \sigma^2$ and $\mathbb{E}\|\nabla_{\mathbf{yy}}^2 g_t(\mathbf{z}; \zeta) - \nabla_{\mathbf{yy}}^2 g_t(\mathbf{z})\|^2 \le \sigma^2$.

6 Experiments

In this section, we provide the empirical studies on the applications of parametric loss tuning for imbalanced data tasks and online hyperparameter optimization.

6.1 Parametric Loss Tuning for Imbalanced Data

Tuning for imbalanced data addresses skewed class distributions, where models, especially high-capacity deep networks, often bias toward the majority class and perform poorly on minority classes. Parametric training loss can be used to address this issue by balancing accuracy and fairness while



(a) regret (b) training loss (c) test accuracy

Figure 1: Performance of different OBO algorithms on the parametric loss tuning for imbalanced data over five runs. F^{2}OBO-SL denotes F^{2}OBO under the single-loop structure, i.e., $K_t \equiv 1$.

Table 3: Performance of different stochastic OBO algorithms on online hyperparameter optimization.

Algorithm	Training loss	Test accuracy	Running time
SOBBO	1.3884 ± 0.0519	0.7979 ± 0.0005	150.80 ± 1.84
SOGD	2.0910 ± 0.0477	0.7607 ± 0.0005	$\mathbf{41.28 \pm 1.51}$
SF^{2}OBO	$\mathbf{1.2914 \pm 0.0664}$	$\mathbf{0.8003 \pm 0.0016}$	103.02 ± 1.77

preventing overfitting, and we use it to optimize the model's performance on the training set $\mathcal{D}_t^{\mathrm{tr}}$ with $\lambda = (\gamma_j, \Delta_j)_{j=1}^J$, which means that the logits adjustment is treated as an upper variable and optimized by validation set $\mathcal{D}_t^{\mathrm{val}}$.

$$\min_{\lambda \in \mathbb{R}^{d_1}} F_t(\lambda) = \frac{1}{|\mathcal{D}_t^{\mathrm{val}}|} \sum_{(x_i,y_i)\in\mathcal{D}_t^{\mathrm{val}}} u_{y_i} \mathcal{L}(f(\theta_t^*(\lambda); x_i), y_i) \quad \text{s.t.} \quad \theta_t^*(\lambda) = \operatorname*{argmin}_{\theta \in \mathbb{R}^{d_2}} g_t(\lambda, \theta),$$

$$\text{where} \quad g_t(\lambda, \theta) = -\frac{1}{|\mathcal{D}_t^{\mathrm{tr}}|} \sum_{(x_i,y_i)\in\mathcal{D}_t^{\mathrm{tr}}} \log \frac{\exp(\gamma_{y_i}[f(\theta; x_i)]_{y_i} + \Delta_{y_i})}{\sum_{j=1}^J \exp(\gamma_j [f(\theta; x_i)]_j + \Delta_j)}.$$

Here u_j is the inverse class frequency (i.e., the reciprocal of the proportion of class-j samples) [17]. We use a lightweight MNIST CNN with four 3×3 conv blocks (Conv-ReLU-2 $\times$ 2 MaxPool–BN), followed by flattening and a 10-way linear classifier.

We evaluated performance over 400 timesteps in four 100-timestep phases, transitioning from 0.6^i to 0.9^i distribution for each class $i = 0, 1, \ldots, 9$. We tune learning rates $\alpha, \beta, \gamma \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5\}$, set $\lambda_1 = 5.0$ and $\tau = 1$. In Figure 1, we compare the performance of different algorithms, where AF^{2}OBO achieves the best balanced testing accuracy. SF^{2}OBO achieves better empirical performance than FSOBO. However, it exhibits a larger local regret, which also corroborates our theoretical results.

6.2 Online Hyperparameter Optimization

We formulate online hyperparameter optimization as a bilevel problem, where the inner problem updates the model parameters and the upper problem optimizes hyperparameters according to validation performance. Under a dynamic data stream, both parameters and hyperparameters are updated at each time step to adapt to distribution shifts. At t, the optimizer receives $\mathcal{D}_t = \{\mathcal{D}_t^{\mathrm{tr}}, \mathcal{D}_t^{\mathrm{val}}\}$, model parameters θ_t, and hyperparameters λ_t. We optimize the L2 regularization coefficient with the box constraint $\lambda_t \in \Lambda = [0, 0.1]^{d_1}$ for a linear model, and define the objective as follows:

$$\min_{\lambda \in \Lambda} F_t(\lambda) = \frac{1}{|\mathcal{D}_t^{\mathrm{val}}|} \sum_{(x_i,y_i)\in\mathcal{D}_t^{\mathrm{val}}} \mathcal{L}(f(\theta_t^*(\lambda); x_i), y_i) \quad \text{s.t.} \quad \theta_t^*(\lambda) = \operatorname*{argmin}_{\theta \in \mathbb{R}^{d_2}} g_t(\lambda, \theta),$$

$$\text{where} \quad g_t(\lambda, \theta) = \frac{1}{|\mathcal{D}_t^{\mathrm{tr}}|} \sum_{(x_i,y_i)\in\mathcal{D}_t^{\mathrm{tr}}} \mathcal{L}\left(f(\theta; x_i), y_i\right) + \Omega(\lambda, \theta).$$

Here, $\Omega(\lambda, \theta) = \frac{1}{2} \sum_{i=0}^{d-1} \lambda_i \|\theta_i\|^2$. Given the optimal model parameter $\theta_t^*(\lambda_t)$, we update the hyperparameters using the validation loss. We evaluate stochastic OBO methods on the 20 Newsgroups text classification task [16]. We repeat each experiment five times. Each run lasts 10,000 iterations, where

5%, 10%, and 15% label noise is injected at iterations 2,500, 5,000, and 7,500 to simulate distribution shifts. Table 3 compares the training loss, test accuracy, and running time of different methods. The results demonstrate that SF^{2}OBO achieves the best overall performance with competitive efficiency.

7 Conclusion and Discussion

This work studies first-order algorithms for non-convex-strongly-convex OBO problem. We propose F^{2}OBO, a fully first-order algorithm, and prove that, under suitable parameter choices, it achieves regret and computational complexity comparable to those of second-order algorithms. We also establish a regret upper bound under the more efficient single-loop structure, which involves an additional gradient-variation term. We further develop an adaptive-iteration variant, AF^{2}OBO, which removes the dependence on $H_{2,T}$ by dynamically adjusting the number of inner iterations, at the cost of additional computation. Finally, we extend our analysis to the stochastic OBO setting, propose SF^{2}OBO, and establish an upper bound on its expected regret.

Limitation and future work Our theoretical analysis is still restricted to OBO problems with strongly convex inner-level objectives, which limits the applicability of our algorithm. In future work, we will extend the analysis of first-order algorithms to OBO problems whose inner-level objectives are non-convex but satisfy the PL condition. Developing an appropriate regret notion and designing efficient algorithms for this setting is an interesting and meaningful direction.

Acknowledgment

This work was supported by Natural Science Foundation of China (No. 62306116) and the Shanghai Special Fund for Promoting High-Quality Industrial Development (No. 20250307).

References

[1] Michael Arbel and Julien Mairal. Non-convex bilevel games with critical point selection maps. *Advances in Neural Information Processing Systems*, 35:8013–8026, 2022.

[2] Jason Bohne, David Rosenberg, Gary Kazantsev, and Pawel Polak. Online nonconvex bilevel optimization with bregman divergences. In *Proceedings of the 16th Annual Workshop on Optimization for Machine Learning*, 2024.

[3] Jason Bohne, Ieva Petrulionyte, Michael Arbel, Julien Mairal, and Pawel Polak. Non-stationary functional bilevel optimization. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300, pages 5275–5283, 2026.

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

[5] Lesi Chen, Jing Xu, and Jingzhao Zhang. On finding small hyper-gradients in bilevel optimization: Hardness results and improved analysis. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 947–980. PMLR, 2024.

[6] Lesi Chen, Yaohua Ma, and Jingzhao Zhang. Near-optimal nonconvex-strongly-convex bilevel optimization with fully first-order oracles. *Journal of Machine Learning Research*, 26(109): 1–56, 2025.

[7] 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.

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

[9] Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.

[10] Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pages 3748–3758. PMLR, 2020.

[11] Feihu Huang. On momentum-based gradient methods for bilevel optimization with nonconvex lower-level. *arXiv preprint arXiv:2303.03944*, 2023.

[12] 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.

[13] Kaiyi Ji, Mingrui Liu, Yingbin Liang, and Lei Ying. Will bilevel optimizers benefit from loops. *Advances in Neural Information Processing Systems*, 35:3011–3023, 2022.

[14] Tingkai Jia, Haiguang Wang, and Cheng Chen. Achieving better local regret bound for online non-convex bilevel optimization. *arXiv preprint arXiv:2602.06457*, 2026.

[15] Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pages 18083–18113. PMLR, 2023.

[16] Ken Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pages 331–339. Elsevier, 1995.

[17] 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.

[18] 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*, 36:29692–29717, 2023.

[19] Bo Liu, Mao Ye, Stephen Wright, Peter Stone, and Qiang Liu. Bome! bilevel optimization made easy: A simple first-order approach. *Advances in Neural Information Processing Systems*, 35:17248–17262, 2022.

[20] Risheng Liu, Yaohua Liu, Shangzhi Zeng, and Jin Zhang. Towards gradient-based bilevel optimization with non-convex followers and beyond. *Advances in Neural Information Processing Systems*, 34:8662–8675, 2021.

[21] Matthew MacKay, Paul Vicol, Jon Lorraine, David Duvenaud, and Roger Grosse. Self-tuning networks: Bilevel optimization of hyperparameters using structured best-response functions. In *International Conference on Learning Representations*, 2019.

[22] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International Conference on Machine Learning*, pages 2113–2122. PMLR, 2015.

[23] Akshay Mehra and Jihun Hamm. Penalty method for inversion-free deep bilevel optimization. In *Asian Conference on Machine Learning*, pages 347–362. PMLR, 2021.

[24] Parvin Nazari, Bojian Hou, Davoud Ataee Tarzanagh, Li Shen, and George Michailidis. Stochastic regret guarantees for online zeroth- and first-order bilevel optimization. In *Advances in Neural Information Processing Systems*, 2025.

[25] Fabian Pedregosa. Hyperparameter optimization with approximate gradient. In *International Conference on Machine Learning*, pages 737–746. PMLR, 2016.

[26] Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1723–1732. PMLR, 2019.

[27] Han Shen and Tianyi Chen. On penalty-based bilevel gradient descent method. In *International Conference on Machine Learning*, pages 30992–31015. PMLR, 2023.

[28] Xingyou Song, Wenbo Gao, Yuxiang Yang, Krzysztof Choromanski, Aldo Pacchiano, and Yunhao Tang. Es-maml: Simple hessian-free meta learning. In *International Conference on Learning Representations*, 2020.

[29] Daouda Sow, Kaiyi Ji, Ziwei Guan, and Yingbin Liang. A primal-dual approach to bilevel optimization with multiple inner minima. *arXiv preprint arXiv:2203.01123*, 2022.

[30] 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.

[31] 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.

[32] Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim. Multimodal model-agnostic meta-learning via task-aware modulation. *Advances in Neural Information Processing Systems*, 32, 2019.

[33] Yifan Yang, Peiyao Xiao, and Kaiyi Ji. Achieving $O(\epsilon^{-1.5})$ complexity in hessian/jacobian-free stochastic bilevel optimization. *Advances in Neural Information Processing Systems*, 36: 39491–39503, 2023.

A Experimental Details and Additional Results

In this section, we provide the detailed parameter settings for the experiments in Section 6 and hyperparameter sensitivity experiments for F^{2}OBO.

A.1 Detailed Parameter Settings

In the parametric loss tuning experiment for imbalanced data, all algorithms use an inner learning rate $\alpha = 0.1$ and an outer learning rate $\gamma = 0.001$. For OBBO and F^{2}OBO, we set the number of inner iterations to $K = 5$. For AOBO, we set $\delta = 0.1$; for AF^{2}OBO, we set $\delta_y = 0.5$ and $\delta_z = 0.1$. The number of linear-system solving steps for AOBO and SOBOW is set to 5. For all our first-order algorithms, we use $\lambda_1 = 5$ with $\tau = 1$.

In the online hyperparameter optimization experiment, each algorithm receives a dataset with batch size 256 at every time step. We use an inner learning rate $\alpha = 0.1$ and an outer learning rate $\gamma = 0.001$. For SOGD, we set the momentum coefficient to $\eta = 0.9$ and the linear-system learning rate to $\beta = 0.1$. For SOBBO, we use a window size of $w = 50$ with $K = 5$ inner iterations. For our SF^{2}OBO, we also set $K = 5$ and use $\lambda_1 = 5$ with $\tau = 1$.

A.2 Hyperparameter Sensitivity Experiments for F^{2}OBO

We conduct hyperparameter sensitivity experiments for F^{2}OBO under the parametric loss tuning task for imbalanced data. Specifically, we study the sensitivity with respect to the number of inner iterations K and the penalty-schedule parameter τ in λ_t. All other experimental settings and hyperparameters are identical to those reported in the main paper.

Table 4: Sensitivity of F^{2}OBO to the number of inner iterations K.

Iterations	$K = 1$	$K = 3$	$K = 5$	$K = 10$
100	0.6151 ± 0.0006	0.7471 ± 0.0143	0.8302 ± 0.0215	0.8654 ± 0.0116
200	0.8222 ± 0.0027	0.9227 ± 0.0002	0.9309 ± 0.0098	0.9301 ± 0.0092
300	0.9469 ± 0.0005	0.9600 ± 0.0006	0.9636 ± 0.0010	0.9613 ± 0.0021
400	0.9741 ± 0.0026	0.9767 ± 0.0011	0.9771 ± 0.0023	0.9752 ± 0.0022

Table 5: Sensitivity of F^{2}OBO to the penalty-schedule parameter τ.

Iterations	$\tau = 0$	$\tau = 0.5$	$\tau = 1$	$\tau = 5$
100	0.8365 ± 0.0265	0.8324 ± 0.0208	0.8302 ± 0.0215	0.8483 ± 0.0083
200	0.9307 ± 0.0081	0.9388 ± 0.0031	0.9309 ± 0.0098	0.9336 ± 0.0113
300	0.9637 ± 0.0019	0.9647 ± 0.0007	0.9636 ± 0.0010	0.9623 ± 0.0038
400	0.9759 ± 0.0028	0.9764 ± 0.0019	0.9771 ± 0.0023	0.9749 ± 0.0033

B Preliminary Lemmas

We list all the lemmas used in the proof below.

Lemma B.1 (Lemma 16 in Tarzanagh et al. [31])**.** *Under Assumptions 2.1, 2.2, for all $t \in [T]$, any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we have*

$$\|\mathbf{y}_t^*(\mathbf{x}_1) - \mathbf{y}_t^*(\mathbf{x}_2)\| \leq \frac{L_{g,1}}{\mu_g}\|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Lemma B.2 (Lemma 4.1 in Chen et al. [6])**.** *Under Assumptions 2.1 and 2.2, let $\lambda > \frac{2L_{f,1}}{\mu_g}$, for all $t \in [T]$, it holds that*

$$\left|f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) - \mathcal{L}_{\lambda,t}^*(\mathbf{x})\right| \leq \frac{D_0}{\lambda}, \tag{12}$$

$$\left\|\nabla f_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x})) - \nabla \mathcal{L}_{\lambda,t}^*(\mathbf{x})\right\| \leq \frac{D_1}{\lambda}, \tag{13}$$

$$\mathcal{L}_{\lambda,t}^*(\mathbf{x}) \text{ is } L_{\mathcal{L},1}\text{-smooth.} \tag{14}$$

Lemma B.3 (Lemma 16 in Tarzanagh et al. [31]). *Under Assumptions 2.1, 2.2, for all $t \in [T]$, any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$, we have*

$$\|\nabla f_t(\mathbf{x}_1, \mathbf{y}_t^*(\mathbf{x}_1)) - \nabla f_t(\mathbf{x}_2, \mathbf{y}_t^*(\mathbf{x}_2))\| \leq L_F \|\mathbf{x}_1 - \mathbf{x}_2\|,$$

where

$$L_F := L_{f,1} + \frac{L_{f,0}L_{g,2}}{\mu_g} + \frac{2L_{g,1}L_{f,1}}{\mu_g} + \frac{L_{g,1}^2 L_{f,1}}{\mu_g^2} + \frac{2L_{f,0}L_{g,1}L_{g,2}}{\mu_g^2} + \frac{L_{f,0}L_{g,1}^2 L_{g,2}}{\mu_g^3}.$$

Lemma B.4 (Lemma B.7 in Jia et al. [14]). *Under Assumption 2.2, we have*

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

where V_T is defined in Eq.(4).

Lemma B.5 (Based on Lemma 3.2 in Kwon et al. [15]). *Let $\lambda > \frac{2L_{f,1}}{\mu_g}$, for all $t \in [T]$, it holds that $\|\mathbf{y}_t^*(\mathbf{x}) - \mathbf{y}_{\lambda,t}^*(\mathbf{x})\| \leq \frac{2L_{f,0}}{\lambda\mu_g}$.*

Proof. Lemma 3.2 [15] shows that for any $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ and $\lambda_2 \geq \lambda_1 \geq \frac{2L_{f,1}}{\mu_g}$, it holds that

$$\left\|\mathbf{y}_{\lambda_1,t}^*(\mathbf{x}_1) - \mathbf{y}_{\lambda_2,t}^*(\mathbf{x}_2)\right\| \leq \frac{2(\lambda_2 - \lambda_1)}{\lambda_1\lambda_2}\frac{L_{f,0}}{\mu_g} + 3\kappa_g \left\|\mathbf{x}_2 - \mathbf{x}_1\right\|.$$

Our result can be achieved if $\lambda = \lambda_1$ for any $\lambda \geq \frac{2L_{f,1}}{\mu_g}$, $\lambda_2 \to +\infty$ and $\mathbf{x}_1 = \mathbf{x}_2$. □

Lemma B.6. *Under Assumptions 2.1 and 2.2, for any $\lambda > \frac{2L_{f,1}}{\mu_g}$ and $t \in [T]$, $\mathcal{L}_{\lambda,t}^*(\mathbf{x})$ is $L_{\mathcal{L},0}$-Lipschitz continuous, where*

$$L_{\mathcal{L},0} := L_{f,0} + \frac{2L_{f,0}L_{g,1}}{\mu_g}.$$

Proof. For any $\lambda > \frac{2L_{f,1}}{\mu_g}$, it follows that

$$\begin{aligned}
\|\nabla \mathcal{L}_t^*(\mathbf{x})\| =& \left\|\nabla_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}_{\lambda,t}^*(\mathbf{x})) + \lambda\left(\nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}_{\lambda,t}^*(\mathbf{x})) - \nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))\right)\right\| \\
\leq& \left\|\nabla_{\mathbf{x}} f_t(\mathbf{x}, \mathbf{y}_{\lambda,t}^*(\mathbf{x}))\right\| + \lambda\left\|\nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}_{\lambda,t}^*(\mathbf{x})) - \nabla_{\mathbf{x}} g_t(\mathbf{x}, \mathbf{y}_t^*(\mathbf{x}))\right\| \\
\leq& L_{f,0} + \lambda L_{g,1}\left\|\mathbf{y}_{\lambda,t}^*(\mathbf{x})) - \mathbf{y}_t^*(\mathbf{x})\right\| \\
\leq& L_{f,0} + \frac{2L_{f,0}L_{g,1}}{\mu_g} =: L_{\mathcal{L},0},
\end{aligned}$$

where the last inequality comes from the $L_{f,0}$-Lipschitz continuity of f_t under Assumption 2.2 and Lemma B.5. □

C Proof of Section 3

In this section, we provide the regret upper bound for Algorithm 1 and the regret upper bound under the single-loop structure in Sections C.1 and C.2, respectively. The following lemma establishes a preliminary regret upper bound for the fully first-order algorithm applied to the approximation problem (P3) in the deterministic OBO setting.

Lemma C.1. *Under Assumptions 2.1, 2.2, suppose that $\gamma_t \leq \frac{1}{2L_F}$ for all $t \in [T]$. Consider the update*

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \gamma_t \mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t(\mathbf{x}_t), \gamma_t\right).$$

Then, we obtain

$$\mathrm{BLReg}(T) \leq \frac{16M}{\gamma_T} + \frac{8}{\gamma_T}V_T + 16D_1^2\sum_{t=1}^{T}\frac{1}{\lambda_t^2} + 16\sum_{t=1}^{T}\left\|\nabla\mathcal{L}_t^*(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2.$$

Proof. It begins with the smoothness of $F_t(\mathbf{x})$ shown in Lemma B.3, we have

$$\begin{aligned}
&F_t(\mathbf{x}_{t+1}) - F_t(\mathbf{x}_t) \\
\leq& \langle \nabla F_t(\mathbf{x}_t)), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L_F}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 \\
=& -\gamma_t \left\langle \nabla F_t(\mathbf{x}_t)), \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\rangle + \frac{\gamma_t^2 L_F}{2} \left\| \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\|^2 \\
=& -\gamma_t \left\langle \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\rangle \\
&+ \gamma_t \left\langle \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla F_t(\mathbf{x}_t), \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\rangle + \frac{\gamma_t^2 L_F}{2} \left\| \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\|^2,
\end{aligned}$$

then with Young's inequality,

$$\begin{aligned}
F_t(\mathbf{x}_{t+1}) - F_t(\mathbf{x}_t) \leq& -\gamma_t \left\| \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\|^2 + \frac{\gamma_t}{2} \left\| \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\|^2 \\
&+ \frac{\gamma_t}{2} \left\| \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla F_t(\mathbf{x}_t)) \right\|^2 + \frac{\gamma_t^2 L_F}{2} \left\| \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\|^2 \\
\leq& -\frac{\gamma_t}{2}(1 - \gamma_t L_F) \left\| \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\|^2 + \frac{\gamma_t}{2} \left\| \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla F_t(\mathbf{x}_t)) \right\|^2.
\end{aligned} \tag{15}$$

Moreover, by

$$\begin{aligned}
-\left\| \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\|^2 \leq& \left\| \mathcal{G}_{\mathcal{X}}(\mathbf{x}_t, \nabla F_t(\mathbf{x}_t)), \gamma_t) - \mathcal{G}_{\mathcal{X}} \left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t), \gamma_t\right) \right\|^2 \\
&- \frac{1}{2} \|\mathcal{G}_{\mathcal{X}}(\mathbf{x}_t, \nabla F_t(\mathbf{x}_t)), \gamma_t)\|^2,
\end{aligned} \tag{16}$$

substituting Eq.(16) into Eq.(15) and rearranging the terms, it follows that

$$\begin{aligned}
\|\mathcal{G}_{\mathcal{X}}(\mathbf{x}_t, \nabla F_t(\mathbf{x}_t), \gamma_t)\|^2 \leq& \frac{8}{\gamma_t}(F_t(\mathbf{x}_t) - F_t(\mathbf{x}_{t+1})) + 8 \left\| \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla F_t(\mathbf{x}_t) \right\|^2 \\
\leq& \frac{8}{\gamma_t}(F_t(\mathbf{x}_t) - F_t(\mathbf{x}_{t+1})) + 16 \|\nabla F_t(\mathbf{x}_t) - \nabla \mathcal{L}_t^*(\mathbf{x}_t)\|^2 \\
&+ 16 \left\| \nabla \mathcal{L}_t^*(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) \right\|^2,
\end{aligned}$$

where we choose $\gamma_t \leq \frac{1}{2L_F}$ to ensure that $1 - \gamma_t L_F \geq \frac{1}{2}$. Summing the above inequality over $t = 1, \ldots, T$, and substituting Lemma B.4 and Lemma B.2 equation (13) in it, we obtain

$$\mathrm{BLReg}(T) \leq \frac{16M}{\gamma_t} + \frac{8}{\gamma_t} V_T + 16 D_1^2 \sum_{t=1}^{T} \frac{1}{\lambda_t^2} + 16 \sum_{t=1}^{T} \left\| \nabla \mathcal{L}_t^*(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) \right\|^2.$$

Thus we finish the proof. □

C.1 Proof of Theorem 3.1

In this section, we give the proof of Theorem 3.1. Lemma C.4 provides an upper bound on the approximate hypergradient error of Algorithm 1 by using the inner-loop iteration error bounds established in Lemmas C.2 and C.3. Substituting this bound into Lemma C.1 yields the proof of Theorem C.5.

Lemma C.2. *Under Assumptions 2.1, 2.2, let* $\beta_t \leq \frac{1}{2\lambda_t L_{g,1}}$ *and* $K_t \equiv K \geq 1 - \frac{\ln 2}{\ln \rho_{\mathbf{y}}}$*, where* $\rho_{\mathbf{y}} = 1 - \frac{\beta_t \lambda_t \mu_g}{2}$*, Algorithm 1 can obtain*

$$\begin{aligned}
&\left\| \mathbf{y}_{t+1} - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t) \right\|^2 \\
\leq& \rho_{\mathbf{y}}^{t-1} \left\| \mathbf{y}_2 - \mathbf{y}_{\lambda,1}^*(\mathbf{x}_1) \right\|^2 + \frac{12 L_{f,0}^2}{\mu_g^2} \sum_{j=0}^{t-2} \frac{1}{\lambda_{t-j}^2} \rho_{\mathbf{y}}^{j+K} + 12 \kappa_g^2 \sum_{j=0}^{t-2} \rho_{\mathbf{y}}^{j+K} \|\mathbf{x}_{t-1-j} - \mathbf{x}_{t-j}\|^2 \\
&+ 12 \sum_{j=0}^{t-2} \rho_{\mathbf{y}}^{j+K} \left\| \mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j}) \right\|^2.
\end{aligned}$$

Proof. It begins with

$$\begin{aligned}
&\left\|\mathbf{y}_t^{k+1}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\\
=&\left\|\mathbf{y}_t^{k}-\beta_t\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x}_t,\mathbf{y}_t^k,\lambda_t)-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\\
=&\left\|\mathbf{y}_t^{k}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2-2\beta_t\left\langle\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x}_t,\mathbf{y}_t^k,\lambda_t),\mathbf{y}_t^k-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\rangle+\beta_t^2\left\|\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x}_t,\mathbf{y}_t^k,\lambda_t)\right\|^2\\
\leq&\left\|\mathbf{y}_t^{k}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2-\frac{\beta_t\lambda_t\mu_g}{2}\left\|\mathbf{y}_t^k-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2-\left(\frac{\beta_t}{2\lambda_tL_{g,1}}-\beta_t^2\right)\left\|\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x}_t,\mathbf{y}_t^k,\lambda_t)\right\|^2\\
=&\rho_{\mathbf{y}}\left\|\mathbf{y}_t^{k}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2,
\end{aligned}\tag{17}$$

where we set $\beta_t\leq\frac{1}{2\lambda_tL_{g,1}}$ and $\rho_{\mathbf{y}}:=1-\frac{\beta_t\lambda_t\mu_g}{2}$. The first inequality follows from the fact that, if $\lambda_t\geq\frac{2L_{f,1}}{\mu_g}$ holds for all $t\in[T]$, then $\mathcal{L}_t$ is $(\lambda_t\mu_g/2)$-strongly convex and $(2\lambda_tL_{g,1})$-smooth, for any $\mathbf{x}\in\mathcal{X}$, $\mathbf{y}\in\mathbb{R}^{d_2}$,

$$\begin{aligned}
\frac{\lambda_t\mu_g}{2}\left\|\mathbf{y}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x})\right\|^2&\leq\left\langle\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x},\mathbf{y},\lambda_t),\mathbf{y}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x})\right\rangle\\
\frac{1}{2\lambda_tL_{g,1}}\left\|\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x},\mathbf{y},\lambda_t)\right\|^2&\leq\left\langle\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x},\mathbf{y},\lambda_t),\mathbf{y}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x})\right\rangle.
\end{aligned}$$

Thus we natually have

$$\begin{aligned}
\left\|\mathbf{y}_{t+1}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\leq&\rho_{\mathbf{y}}^{K_t}\left\|\mathbf{y}_t-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\\
\leq&2\rho_{\mathbf{y}}^{K_t}\left\|\mathbf{y}_t-\mathbf{y}_{\lambda_{t-1},t-1}^*(\mathbf{x}_{t-1})\right\|^2+2\rho_{\mathbf{y}}^{K_t}\left\|\mathbf{y}_{\lambda_{t-1},t-1}^*(\mathbf{x}_{t-1})-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2.
\end{aligned}$$

Then it follows that

$$\begin{aligned}
&\left\|\mathbf{y}_{\lambda_{t-1},t-1}^*(\mathbf{x}_{t-1})-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\\
\leq&3\left\|\mathbf{y}_{\lambda_{t-1},t-1}^*(\mathbf{x}_{t-1})-\mathbf{y}_{t-1}^*(\mathbf{x}_{t-1})\right\|^2+3\left\|\mathbf{y}_{t-1}^*(\mathbf{x}_{t-1})-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\
&+3\left\|\mathbf{y}_t^*(\mathbf{x}_t)-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\\
\leq&\frac{6L_{f,0}^2}{\lambda_t^2\mu_g^2}+6\kappa_g^2\left\|\mathbf{x}_{t-1}-\mathbf{x}_t\right\|^2+6\left\|\mathbf{y}_{t-1}^*(\mathbf{x}_t)-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2,
\end{aligned}\tag{18}$$

where the last ineqaulity comes from Lemma B.5 and B.1. Let $K_t\equiv K\geq1-\frac{\ln2}{\ln\rho_{\mathbf{y}}}$ such that $2\rho_{\mathbf{y}}^K\leq\rho_{\mathbf{y}}$, we have

$$\begin{aligned}
&\left\|\mathbf{y}_{t+1}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\\
\leq&\rho_{\mathbf{y}}\left\|\mathbf{y}_t-\mathbf{y}_{\lambda_{t-1},t-1}^*(\mathbf{x}_{t-1})\right\|^2+\frac{12L_{f,0}^2}{\lambda_t^2\mu_g^2}\rho_{\mathbf{y}}^K+12\kappa_g^2\rho_{\mathbf{y}}^K\left\|\mathbf{x}_{t-1}-\mathbf{x}_t\right\|^2\\
&+12\rho_{\mathbf{y}}^K\left\|\mathbf{y}_{t-1}^*(\mathbf{x}_t)-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\
\leq&\rho_{\mathbf{y}}^{t-1}\left\|\mathbf{y}_2-\mathbf{y}_{\lambda,1}^*(\mathbf{x}_1)\right\|^2+\frac{12L_{f,0}^2}{\mu_g^2}\sum_{j=0}^{t-2}\frac{1}{\lambda_{t-j}^2}\rho_{\mathbf{y}}^{j+K}+12\kappa_g^2\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^{j+K}\left\|\mathbf{x}_{t-1-j}-\mathbf{x}_{t-j}\right\|^2\\
&+12\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^{j+K}\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j})-\mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2.
\end{aligned}\tag{19}$$

Thus we finish the proof. □

Lemma C.3. *Under Assumptions 2.1, 2.2, let $\alpha\leq\frac{1}{L_{g,1}}$ and $K_t\equiv K\geq1-\frac{\ln2}{\ln\rho_{\mathbf{z}}}$, where $\rho_{\mathbf{z}}=1-\alpha\mu_g$, Algorithm 1 can obtain*

$$\left\|\mathbf{z}_{t+1}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\leq\rho_{\mathbf{z}}^{t-1}\left\|\mathbf{z}_2-\mathbf{y}_1^*(\mathbf{x}_1)\right\|^2+4\kappa_g^2\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^{j+K}\left\|\mathbf{x}_{t-1-j}-\mathbf{x}_{t-j}\right\|^2$$

$$+4\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^{j+K}\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j})-\mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2.$$

Proof. It begins with

$$\begin{aligned}
&\left\|\mathbf{z}_t^{k+1}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\
=&\left\|\mathbf{z}_t^{k}-\alpha\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{z}_t^k)-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\
=&\left\|\mathbf{z}_t^{k}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2-2\alpha\left\langle\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{z}_t^k),\mathbf{z}_t^k-\mathbf{y}_t^*(\mathbf{x}_t)\right\rangle+\alpha^2\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{z}_t^k)\right\|^2\\
\leq&\left\|\mathbf{z}_t^{k}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2-\alpha\mu_g\left\|\mathbf{z}_t^k-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2-\left(\frac{\alpha}{L_{g,1}}-\alpha^2\right)\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t,\mathbf{z}_t^k)\right\|^2\\
=&\rho_{\mathbf{z}}\left\|\mathbf{z}_t^{k}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2,
\end{aligned}$$

where we set $\alpha\leq\frac{1}{L_{g,1}}$ and $\rho_{\mathbf{z}}:=1-\alpha\mu_g$. the first inequality follows from Assumption 2.1 and 2.2. Then we complete the proof by

$$\begin{aligned}
&\left\|\mathbf{z}_{t+1}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\
\leq&\rho_{\mathbf{z}}^{K_t}\left\|\mathbf{z}_t-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\
\leq&2\rho_{\mathbf{z}}^{K_t}\left\|\mathbf{z}_t-\mathbf{y}_{t-1}^*(\mathbf{x}_{t-1})\right\|^2+2\rho_{\mathbf{z}}^{K_t}\left\|\mathbf{y}_{t-1}^*(\mathbf{x}_{t-1})-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\
\overset{(i)}{\leq}&\rho_{\mathbf{z}}\left\|\mathbf{z}_t-\mathbf{y}_{t-1}^*(\mathbf{x}_{t-1})\right\|^2+4\kappa_g^2\rho_{\mathbf{z}}^{K}\left\|\mathbf{x}_{t-1}-\mathbf{x}_t\right\|^2+4\rho_{\mathbf{z}}^{K}\left\|\mathbf{y}_{t-1}^*(\mathbf{x}_t)-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\
\leq&\rho_{\mathbf{z}}^{t-1}\left\|\mathbf{z}_2-\mathbf{y}_1^*(\mathbf{x}_1)\right\|^2+4\kappa_g^2\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^{j+K}\left\|\mathbf{x}_{t-1-j}-\mathbf{x}_{t-j}\right\|^2\\
&+4\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^{j+K}\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j})-\mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2,
\end{aligned}$$

where in (i), we set $K_t\equiv K\geq 1-\frac{\ln 2}{\ln\rho_{\mathbf{z}}}$ such that $2\rho_{\mathbf{z}}^K\leq\rho_{\mathbf{z}}$. □

Lemma C.4. *Under Assumptions 2.1 and 2.2, let $\Lambda(t,\lambda_t)=\left(1+\frac{1}{t}\right)^{\tau}\lambda_t$ for some $\tau>0$, $\lambda_1>\frac{2L_{f,1}}{\mu_g}$, $\alpha\leq\frac{1}{L_{g,1}}$, $\beta_t\leq\frac{1}{2\lambda_t L_{g,1}}$, $\gamma_t\equiv\gamma\leq\frac{1}{4\kappa_g}\sqrt{\frac{1-\rho}{C_\lambda}}$ and $K_t\equiv K\geq\max\{1-\frac{\ln 2}{\ln\rho},\frac{\ln c-2\tau\ln T}{\ln\rho}\}$ for some $c>0$, the approximation error of hypergradient in Algorithm 1 can be bounded as follows:*

$$\begin{aligned}
\sum_{t=2}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\leq&2c_\tau\Delta_1+\frac{576c\kappa_g^3L_{f,0}^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}+\frac{8\gamma^2C_\lambda\kappa_g^2D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}\\
&+\frac{4\gamma^2C_\lambda\kappa_g^2}{1-\rho}\mathrm{BLReg}(T)+\frac{2C_\lambda}{1-\rho}\sum_{t=2}^{T}\sup_{\mathbf{x}\in\mathbb{R}^{d_1}}\left\|\mathbf{y}_{t-1}^*(\mathbf{x})-\mathbf{y}_t^*(\mathbf{x})\right\|^2.
\end{aligned}$$

where ρ, c, c_τ, C_λ and Δ_1 are some constants.

Proof. It begins with

$$\begin{aligned}
\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\leq&3\left\|\nabla_{\mathbf{x}}f_t(\mathbf{x}_t,\mathbf{y}_{t+1})-\nabla_{\mathbf{x}}f_t(\mathbf{x}_t,\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}))\right\|^2\\
&+3\lambda_t^2\left\|\nabla_{\mathbf{x}}g_t(\mathbf{x}_t,\mathbf{y}_{t+1})-\nabla_{\mathbf{x}}g_t(\mathbf{x}_t,\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t))\right\|^2\\
&+3\lambda_t^2\left\|\nabla_{\mathbf{x}}g_t(\mathbf{x}_t,\mathbf{z}_{t+1})-\nabla_{\mathbf{x}}g_t(\mathbf{x}_t,\mathbf{y}_t^*(\mathbf{x}_t))\right\|^2\\
\overset{(i)}{\leq}&3\left(L_{f,1}^2+\lambda_t^2L_{g,1}^2\right)\left\|\mathbf{y}_{t+1}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2+3\lambda_t^2L_{g,1}^2\left\|\mathbf{z}_{t+1}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\
\overset{(ii)}{\leq}&6\lambda_t^2L_{g,1}^2\underbrace{\left\|\mathbf{y}_{t+1}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2}_{\mathcal{L}_t\text{ inner iters bound: }(a)}+3\lambda_t^2L_{g,1}^2\underbrace{\left\|\mathbf{z}_{t+1}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2}_{g_t\text{ inner iters bound: }(b)},
\end{aligned}\tag{20}$$

where (i) comes from the smoothness of f_t and g_t under Assumption 2.2, and (ii) comes from $\lambda_t > \frac{2L_{f,1}}{\mu_g}$, thus $L_{f,1} < \lambda_t L_{g,1}$. Substituting Lemma C.2 and Lemma C.3 into equation (20), we have

$$
\begin{aligned}
&\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 \\
\leq& 6\lambda_t^2 L_{g,1}^2 \rho_{\mathbf{y}}^{t-1} \left\|\mathbf{y}_2 - \mathbf{y}_{\lambda,1}^*(\mathbf{x}_1)\right\|^2 + 3\lambda_t^2 L_{g,1}^2 \rho_{\mathbf{z}}^{t-1} \|\mathbf{z}_2 - \mathbf{y}_1^*(\mathbf{x}_1)\|^2 + 72\kappa_g^2 L_{f,0}^2 \sum_{j=0}^{t-2} \frac{\lambda_t^2}{\lambda_{t-j}^2} \rho_{\mathbf{y}}^{j+K} \\
&+ 72\lambda_t^2 \kappa_g^2 L_{g,1}^2 \sum_{j=0}^{t-2} \rho_{\mathbf{y}}^{j+K} \|\mathbf{x}_{t-1-j} - \mathbf{x}_{t-j}\|^2 + 72\lambda_t^2 L_{g,1}^2 \sum_{j=0}^{t-2} \rho_{\mathbf{y}}^{j+K} \left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2 \\
&+ 12\lambda_t^2 \kappa_g^2 L_{g,1}^2 \sum_{j=0}^{t-2} \rho_{\mathbf{z}}^{j+K} \|\mathbf{x}_{t-1-j} - \mathbf{x}_{t-j}\|^2 + 12\lambda_t^2 L_{g,1}^2 \sum_{j=0}^{t-2} \rho_{\mathbf{z}}^{j+K} \|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\|^2.
\end{aligned}
$$

Suppose that, for a sufficiently large K, i.e., $K \geq \frac{\ln c - 2\tau \ln T}{\ln \rho}$, $\lambda_t^2 \rho^K \leq c$ holds for some constant $c > 0$, where $\rho := \max\{\rho_{\mathbf{y}}, \rho_{\mathbf{z}}\}$. Then we have

$$
\begin{aligned}
&\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 \\
\leq& 6\lambda_t^2 L_{g,1}^2 \rho_{\mathbf{y}}^{t-1} \left\|\mathbf{y}_2 - \mathbf{y}_{\lambda,1}^*(\mathbf{x}_1)\right\|^2 + 3\lambda_t^2 L_{g,1}^2 \rho_{\mathbf{z}}^{t-1} \|\mathbf{z}_2 - \mathbf{y}_1^*(\mathbf{x}_1)\|^2 + 72c\kappa_g^2 L_{f,0}^2 \sum_{j=0}^{t-2} \frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2} \\
&+ 72c\kappa_g^2 L_{g,1}^2 \sum_{j=0}^{t-2} \rho_{\mathbf{y}}^j \|\mathbf{x}_{t-1-j} - \mathbf{x}_{t-j}\|^2 + 72cL_{g,1}^2 \sum_{j=0}^{t-2} \rho_{\mathbf{y}}^j \left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2 \\
&+ 12c\kappa_g^2 L_{g,1}^2 \sum_{j=0}^{t-2} \rho_{\mathbf{z}}^j \|\mathbf{x}_{t-1-j} - \mathbf{x}_{t-j}\|^2 + 12cL_{g,1}^2 \sum_{j=0}^{t-2} \rho_{\mathbf{z}}^j \|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\|^2 \\
\overset{(i)}{\leq}& \lambda_t^2 \rho^{t-1} \Delta_1' + 72c\kappa_g^2 L_{f,0}^2 \sum_{j=0}^{t-2} \frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2} + C_\lambda \sum_{j=0}^{t-2} \rho^j \left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2 \\
&+ C_\lambda \kappa_g^2 \sum_{j=0}^{t-2} \rho^j \|\mathbf{x}_{t-1-j} - \mathbf{x}_{t-j}\|^2,
\end{aligned} \tag{21}
$$

where in (i), we define $\Delta_1' := 6L_{g,1}^2 \|\mathbf{y}_2 - \mathbf{y}_{\lambda_1,1}^*(\mathbf{x}_1)\|^2 + 3L_{g,1}^2 \|\mathbf{z}_2 - \mathbf{y}_1^*(\mathbf{x}_1)\|^2$, $C_\lambda := 84cL_{g,1}^2$ and $\rho := \max\{\rho_{\mathbf{y}}, \rho_{\mathbf{z}}\}$. After telescoping equation (21), we have

$$
\begin{aligned}
&\sum_{t=2}^{T} \left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 \\
\leq& \Delta_1 \text{`} \sum_{t=2}^{T} \lambda_t^2 \rho^{t-1} + 72c\kappa_g^2 L_{f,0}^2 \sum_{t=2}^{T} \sum_{j=0}^{t-2} \frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2} + C_\lambda \sum_{t=2}^{T} \sum_{j=0}^{t-2} \rho^j \left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2 \\
&+ C_\lambda \kappa_g^2 \sum_{t=2}^{T} \sum_{j=0}^{t-2} \rho^j \|\mathbf{x}_{t-1-j} - \mathbf{x}_{t-j}\|^2.
\end{aligned} \tag{22}
$$

Note that, for $\rho = e^{-\theta}$ and $\lambda_t = \prod_{i=1}^{t} (1 + \frac{1}{i})^\tau \lambda_1 = \lambda_1 t^\tau$, we have

$$
\sum_{t=1}^{T} \lambda_t^2 \rho^{t-1} = \lambda_1^2 \sum_{t=1}^{T} t^{2\tau} \rho^{t-1} \leq \frac{\lambda_1^2}{\rho} \sum_{t=1}^{\infty} t^{2\tau} \rho^t = \lambda_1^2 e^\theta \sum_{t=1}^{\infty} t^{2\tau} e^{-\theta t} \leq \lambda_1^2 e^\theta \left(\int_0^\infty x^{2\tau} e^{-\theta x} \, \mathrm{d}x + 1 \right).
$$

Using the standard Gamma integral

$$
\int_0^\infty x^{s-1} e^{-\theta x} \, \mathrm{d}x = \frac{\Gamma(s)}{\theta^s}, \qquad s > 0, \ \theta > 0,
$$

we obtain

$$\int_0^\infty x^{2\tau} e^{-\theta x}\,\mathrm{d}x = \frac{\Gamma(2\tau+1)}{\theta^{2\tau+1}}.$$

Therefore,

$$\sum_{t=1}^{T} \lambda_t^2 \rho^{t-1} \le \lambda_1^2 e^{\theta}\left(\frac{\Gamma(2\tau+1)}{\theta^{2\tau+1}}+1\right) =: c_\tau. \tag{23}$$

Also, if we set $m = t-j$, then

$$\sum_{t=2}^{T}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2} < \sum_{t=1}^{T}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2} = \sum_{t=1}^{T}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j}{(t-j)^{2\tau}\lambda_1^2} = \frac{1}{\lambda_1^2}\sum_{m=2}^{T}\frac{1}{m^{2\tau}}\sum_{j=0}^{T-m}\rho_{\mathbf{y}}^j \le \frac{4L_{g,1}}{\lambda_1^2\mu_g}\sum_{m=2}^{T}\frac{1}{m^{2\tau}}.$$

Substituting the above inequality and equation (23) into equation (22), and adopting a fixed step-size policy where $\gamma_t \equiv \gamma$, it holds that

$$\begin{aligned}
&\sum_{t=2}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 \\
\le& c_\tau\Delta_1' + \frac{288c\kappa_g^3 L_{f,0}^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + \frac{C_\lambda\kappa_g^2}{1-\rho}\sum_{t=2}^{T}\left\|\mathbf{x}_{t-1}-\mathbf{x}_t\right\|^2 + \frac{C_\lambda}{1-\rho}\sum_{t=2}^{T}\left\|\mathbf{y}_{t-1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2 \\
\le& c_\tau\Delta_1' + \frac{288c\kappa_g^3 L_{f,0}^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + \frac{2\gamma^2 C_\lambda\kappa_g^2}{1-\rho}\sum_{t=2}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla F_t(\mathbf{x}_t)\right\|^2 + \frac{2\gamma^2 C_\lambda\kappa_g^2}{1-\rho}\mathrm{BLReg}(T) \\
&+ \frac{C_\lambda}{1-\rho}\sum_{t=2}^{T}\sup_{\mathbf{x}\in\mathbb{R}^{d_1}}\left\|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\right\|^2 \\
\le& c_\tau\Delta_1' + \frac{288c\kappa_g^3 L_{f,0}^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + \frac{4\gamma^2 C_\lambda\kappa_g^2 D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + \frac{4\gamma^2 C_\lambda\kappa_g^2}{1-\rho}\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 \\
&+ \frac{2\gamma^2 C_\lambda\kappa_g^2}{1-\rho}\mathrm{BLReg}(T) + \frac{C_\lambda}{1-\rho}\sum_{t=2}^{T}\sup_{\mathbf{x}\in\mathbb{R}^{d_1}}\left\|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\right\|^2,
\end{aligned}$$

where the first inequality comes from the fact that $\sum_{j=0}^{t-2}\rho^j < \sum_{j=0}^{\infty}\rho^j = \frac{1}{1-\rho}$ for $\forall\rho\in(0,1)$, and the last inequality follows Lemma B.2 Eq.(13). Let $\gamma \le \frac{1}{4\kappa_g}\sqrt{\frac{1-\rho}{C_\lambda}}$ such that $1 - \frac{4\gamma^2 C_\lambda\kappa_g^2}{1-\rho} \ge \frac{1}{2}$, thus we complete the proof by

$$\begin{aligned}
\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 \le& 2c_\tau\Delta_1 + \frac{576c\kappa_g^3 L_{f,0}^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + \frac{8\gamma^2 C_\lambda\kappa_g^2 D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} \\
&+ \frac{4\gamma^2 C_\lambda\kappa_g^2}{1-\rho}\mathrm{BLReg}(T) + \frac{2C_\lambda}{1-\rho}\sum_{t=2}^{T}\sup_{\mathbf{x}\in\mathbb{R}^{d_1}}\left\|\mathbf{y}_{t-1}^*(\mathbf{x}) - \mathbf{y}_t^*(\mathbf{x})\right\|^2,
\end{aligned}$$

where $\Delta_1 := \Delta_1' + \|\widetilde{\nabla}\mathcal{L}_1^*(\mathbf{x}_1) - \nabla\mathcal{L}_1^*(\mathbf{x}_1)\|^2$. □

Theorem C.5 (Restatement of Theorem 3.1)**.** *Under Assumptions 2.1, 2.2, let* $\Lambda(t,\lambda_t) = \left(1+\frac{1}{t}\right)^\tau\lambda_t$ *for some* $\tau > 0$, $\lambda_1 > \frac{2L_{f,1}}{\mu_g}$, $\alpha \le \frac{1}{L_{g,1}}$, $\beta_t \le \frac{1}{2\lambda_t L_{g,1}}$, $\gamma_t \equiv \gamma \le \min\{\frac{1}{2L_F}, \frac{1}{16\kappa_g}\sqrt{\frac{1-\rho}{C_\lambda}}\}$ *and* $K_t \equiv K \ge \max\{1 - \frac{\ln 2}{\ln\rho}, \frac{\ln c - 2\tau\ln T}{\ln\rho}\}$ *for some* $c > 0$*, Algorithm 1 can guarantee*

$$\begin{aligned}
\mathrm{BLReg}(T) \le& \frac{32M}{\gamma} + \frac{16}{\gamma}V_T + \frac{32D_1^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + 64c_\tau\Delta_1 + \frac{18432c\kappa_g^3 L_{f,0}^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} \\
&+ \frac{256\gamma^2 C_\lambda\kappa_g^2 D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + \frac{64C_\lambda}{1-\rho}H_{2,T}
\end{aligned}$$

$$=O\left(\sum_{t=1}^{T}\frac{1}{t^{2\tau}}+V_T+H_{2,T}\right),$$

where ρ, c, c_τ, C_λ and Δ_1 are some constants.

Proof. We substitute Lemma C.4 into Lemma C.1 to get

$$\begin{aligned}\mathrm{BLReg}(T)\leq&\frac{16M}{\gamma}+\frac{8}{\gamma}V_T+\frac{16D_1^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}+32c_\tau\Delta_1+\frac{9216c\kappa_g^3L_{f,0}^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}+\frac{32C_\lambda}{1-\rho}H_{2,T}\\&+\frac{64\gamma^2C_\lambda\kappa_g^2}{1-\rho}\mathrm{BLReg}(T)+\frac{128\gamma^2C_\lambda\kappa_g^2D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}.\end{aligned}$$

Let $\gamma\leq\frac{1}{16\kappa_g}\sqrt{\frac{1-\rho}{C_\lambda}}$ such that $1-\frac{64\gamma^2C_\lambda\kappa_g^2}{1-\rho}\geq\frac{1}{2}$, thus finally have

$$\begin{aligned}\mathrm{BLReg}(T)\leq&\frac{32M}{\gamma}+\frac{16}{\gamma}V_T+\frac{32D_1^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}+64c_\tau\Delta_1+\frac{18432c\kappa_g^3L_{f,0}^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}\\&+\frac{256\gamma^2C_\lambda\kappa_g^2D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}+\frac{64C_\lambda}{1-\rho}H_{2,T}\\=&O\left(\sum_{t=1}^{T}\frac{1}{t^{2\tau}}+V_T+H_{2,T}\right).\end{aligned}$$

Our final bound depends on the choice of τ:

$$\begin{cases}\mathrm{BLReg}(T)\leq O\left(1+V_T+H_{2,T}\right) & \text{if set } \tau\in(\frac{1}{2},\infty)\\ \mathrm{BLReg}(T)\leq O\left(\log T+V_T+H_{2,T}\right) & \text{if set } \tau=\frac{1}{2}\\ \mathrm{BLReg}(T)\leq O\left(T^{1-\tau}+V_T+H_{2,T}\right) & \text{if set } \tau\in(0,\frac{1}{2})\end{cases}.$$

□

C.2 Proof of Theorem 3.2

In this section, we prove the regret upper bound of F^{2}OBO under $K_t\equiv 1$. We present single-loop counterparts of Lemmas C.2 and C.3 in Lemmas C.6 and C.7, respectively. Then, in Lemma C.8, we establish the final upper bound on the hypergradient approximation error, which completes the proof of Theorem C.9.

Lemma C.6. *Under Assumptions 2.1, 2.2, let $\beta_t=\frac{1}{2\lambda_tL_{g,1}}$ and $K_t\equiv 1$, Algorithm 1 can obtain*

$$\begin{aligned}&\left\|\mathbf{y}_{t+1}-\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2\\\leq&\rho_{\mathbf{y}}^{t-1}\left\|\mathbf{y}_2-\mathbf{y}^*_{\lambda_1,1}(\mathbf{x}_1)\right\|^2+288\kappa_g^3\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\left\|\mathbf{x}_{t-1-j}-\mathbf{x}_{t-j}\right\|^2+\frac{1536L_{f,0}^2\kappa_g^3}{\mu_g^2}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j(\lambda_{t-j}-\lambda_{t-1-j})^2}{\lambda_{t-j}^2\lambda_{t-1-j}^2}\\&+\frac{384\kappa_g}{\mu_g^2}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_{t-j}(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1-j}(\mathbf{x},\mathbf{y})\right\|^2\\&+\frac{384\kappa_g}{\mu_g^2}\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_{t-j}(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1-j}(\mathbf{x},\mathbf{y})\right\|^2,\end{aligned}$$

where $\rho_{\mathbf{y}}=1-\frac{1}{8\kappa_g}$.

Proof. We begin with equation (17) in Lemma C.2, it follows that

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

$$
\begin{aligned}
&\leq \left(1-\frac{\beta_t\lambda_t\mu_g}{2}\right)(1+\theta)\left\|\mathbf{y}_t-\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})\right\|^2 \\
&\quad+\left(1-\frac{\beta_t\lambda_t\mu_g}{2}\right)\left(1+\frac{1}{\theta}\right)\left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})-\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 \\
&\overset{(i)}{\leq}\left(1-\frac{1}{8\kappa_g}\right)\left\|\mathbf{y}_t-\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})\right\|^2 \\
&\quad+16\kappa_g\left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})-\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2, \qquad (24)
\end{aligned}
$$

where in (i), we set $\theta=\frac{\beta_t\lambda_t\mu_g}{4}$ such that $(1-\frac{\beta_t\lambda_t\mu_g}{2})(1+\frac{\beta_t\lambda_t\mu_g}{4})=1-\frac{\beta_t\lambda_t\mu_g}{4}-\frac{\beta_t^2\lambda_t^2\mu_g^2}{8}\leq 1-\frac{\beta_t\lambda_t\mu_g}{4}$, and $\beta_t=\frac{1}{2\lambda_t L_{g,1}}$. Then it begins with

$$
\begin{aligned}
\left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})-\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 \leq& 2\left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})-\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_t)\right\|^2 \\
&+2\left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_t)-\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 \\
\overset{(i)}{\leq}&18\kappa_g^2\left\|\mathbf{x}_{t-1}-\mathbf{x}_t\right\|^2+2\left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_t)-\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2, \qquad (25)
\end{aligned}
$$

where (i) comes from Lemma 3.2 [15]. Remind that $\mathcal{L}_t(\mathbf{x},\mathbf{y},\lambda)$ is $(\lambda_t\mu_g/2)$-strongly convex in $\mathbf{y}$, the following formula can be obtained immediately

$$
\left\|\mathbf{y}^*_{\lambda_t,t}(\mathbf{x})-\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x})\right\|\leq\frac{2}{\lambda_t\mu_g}\left\|\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}),\lambda_t)\right\| \qquad (26)
$$

We also noticed that

$$
\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}),\lambda_t)=\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))+\lambda_t\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x})) \qquad (27)
$$

$$
\nabla_{\mathbf{y}}\mathcal{L}_{t-1}(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}),\lambda_{t-1})=\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))+\lambda_{t-1}\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))=0, \qquad (28)
$$

then substitute Eq.(28) into Eq.(27), we obtain

$$
\begin{aligned}
\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}),\lambda_t)=&\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))-\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x})) \\
&+\lambda_t\left(\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))\right) \\
&+(\lambda_t-\lambda_{t-1})\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x})).
\end{aligned}
$$

With

$$
\begin{aligned}
\left\|\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))\right\|=&\left\|\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y}^*_{t-1}(\mathbf{x}))\right\| \\
\leq&L_{g,1}\left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x})-\mathbf{y}^*_{t-1}(\mathbf{x})\right\|\leq\frac{2L_{f,0}L_{g,1}}{\lambda_{t-1}\mu_g},
\end{aligned}
$$

it naturally holds that

$$
\begin{aligned}
\left\|\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}),\lambda_t)\right\|\leq&\left\|\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))-\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))\right\| \\
&+\lambda_t\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))\right\| \\
&+\frac{2L_{f,0}L_{g,1}}{\mu_g}\frac{\lambda_t-\lambda_{t-1}}{\lambda_{t-1}}. \qquad (29)
\end{aligned}
$$

Finally we substitute Eq.(29) into Eq.(26) to get

$$
\begin{aligned}
\left\|\mathbf{y}^*_{\lambda_t,t}(\mathbf{x})-\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x})\right\|\leq&\frac{2}{\lambda_t\mu_g}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y})\right\| \\
&+\frac{2}{\mu_g}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y})\right\|
\end{aligned}
$$

$$+\frac{4L_{f,0}L_{g,1}}{\mu_g^2}\frac{\lambda_t-\lambda_{t-1}}{\lambda_t\lambda_{t-1}}$$

and natually have

$$\begin{aligned}\left\|\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}))-\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}))\right\|^2 \leq&\frac{12}{\lambda_t^2\mu_g^2}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y})\right\|^2\\&+\frac{12}{\mu_g^2}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y})\right\|^2\\&+\frac{48L_{f,0}^2L_{g,1}^2}{\mu_g^4}\frac{(\lambda_t-\lambda_{t-1})^2}{\lambda_t^2\lambda_{t-1}^2}.\end{aligned}\tag{30}$$

Substitute Eq.(30) into Eq.(25) to obtain

$$\begin{aligned}&\left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})-\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2\\\leq&18\kappa_g^2\left\|\mathbf{x}_{t-1}-\mathbf{x}_t\right\|^2+\frac{24}{\lambda_t^2\mu_g^2}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y})\right\|^2\\&+\frac{24}{\mu_g^2}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y})\right\|^2+\frac{96L_{f,0}^2L_{g,1}^2}{\mu_g^4}\frac{(\lambda_t-\lambda_{t-1})^2}{\lambda_t^2\lambda_{t-1}^2}.\end{aligned}\tag{31}$$

Finally, substitute Eq.(31) into Eq.(24), we finish the proof by

$$\begin{aligned}&\left\|\mathbf{y}_{t+1}-\mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2\\\leq&\rho_{\mathbf{y}}\left\|\mathbf{y}_t-\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})\right\|^2+288\kappa_g^3\left\|\mathbf{x}_{t-1}-\mathbf{x}_t\right\|^2+\frac{1536L_{f,0}^2\kappa_g^3}{\mu_g^2}\frac{(\lambda_t-\lambda_{t-1})^2}{\lambda_t^2\lambda_{t-1}^2}\\&+\frac{384\kappa_g}{\lambda_t^2\mu_g^2}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y})\right\|^2+\frac{384\kappa_g}{\mu_g^2}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y})\right\|^2\\\leq&\rho_{\mathbf{y}}^{t-1}\left\|\mathbf{y}_2-\mathbf{y}^*_{\lambda_1,1}(\mathbf{x}_1)\right\|^2+288\kappa_g^3\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\left\|\mathbf{x}_{t-1-j}-\mathbf{x}_{t-j}\right\|^2+\frac{1536L_{f,0}^2\kappa_g^3}{\mu_g^2}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j(\lambda_{t-j}-\lambda_{t-1-j})^2}{\lambda_{t-j}^2\lambda_{t-1-j}^2}\\&+\frac{384\kappa_g}{\mu_g^2}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_{t-j}(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1-j}(\mathbf{x},\mathbf{y})\right\|^2\\&+\frac{384\kappa_g}{\mu_g^2}\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_{t-j}(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1-j}(\mathbf{x},\mathbf{y})\right\|^2,\end{aligned}$$

where $\rho_{\mathbf{y}}:=1-\frac{1}{8\kappa_g}$. □

Lemma C.7. *Under Assumptions 2.1, 2.2, let $\alpha\leq\frac{1}{L_{g,1}}$ and $K_t\equiv 1$, Algorithm 1 can obtain*

$$\begin{aligned}\left\|\mathbf{z}_{t+1}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\leq&\rho_{\mathbf{z}}^{t-1}\left\|\mathbf{z}_2-\mathbf{y}_1^*(\mathbf{x}_1)\right\|^2+\frac{8\kappa_g^2}{\alpha\mu_g}\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\left\|\mathbf{x}_{t-1-j}-\mathbf{x}_{t-j}\right\|^2\\&+\frac{8}{\alpha\mu_g}\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\sup_{\mathbf{x}}\left\|\mathbf{y}^*_{t-1-j}(\mathbf{x})-\mathbf{y}^*_{t-j}(\mathbf{x})\right\|^2,\end{aligned}$$

where $\rho_{\mathbf{z}}=1-\frac{\alpha\mu_g}{2}$.

Proof. It follows that

$$\begin{aligned}&\left\|\mathbf{z}_{t+1}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\leq(1-\alpha\mu_g)\left\|\mathbf{z}_t-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\\\leq&(1-\alpha\mu_g)(1+\theta)\left\|\mathbf{z}_t-\mathbf{y}^*_{t-1}(\mathbf{x}_{t-1})\right\|^2+(1-\alpha\mu_g)\left(1+\frac{1}{\theta}\right)\left\|\mathbf{y}^*_{t-1}(\mathbf{x}_{t-1})-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\end{aligned}$$

$$
\begin{aligned}
&\leq \left(1-\frac{\alpha\mu_g}{2}\right)\left\|\mathbf{z}_t-\mathbf{y}_{t-1}^*(\mathbf{x}_{t-1})\right\|^2+\frac{8\kappa_g^2}{\alpha\mu_g}\left\|\mathbf{x}_{t-1}-\mathbf{x}_t\right\|^2+\frac{8}{\alpha\mu_g}\sup_{\mathbf{x}}\left\|\mathbf{y}_{t-1}^*(\mathbf{x})-\mathbf{y}_t^*(\mathbf{x})\right\|^2\\
&\leq \rho_{\mathbf{z}}^{t-1}\left\|\mathbf{z}_2-\mathbf{y}_1^*(\mathbf{x}_1)\right\|^2+\frac{8\kappa_g^2}{\alpha\mu_g}\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\left\|\mathbf{x}_{t-1-j}-\mathbf{x}_{t-j}\right\|^2\\
&\quad+\frac{8}{\alpha\mu_g}\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\sup_{\mathbf{x}}\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x})-\mathbf{y}_{t-j}^*(\mathbf{x})\right\|^2,
\end{aligned}
$$

where in the second inequality, we set $\theta=\frac{\alpha\mu_g}{2}$ such that $(1-\alpha\mu_g)(1+\frac{\alpha\mu_g}{2})\leq 1-\frac{\alpha\mu_g}{2}$, and $\rho_{\mathbf{z}}:=1-\frac{\alpha\mu_g}{2}$. □

Lemma C.8. *Under Assumptions 2.1 and 2.2, let $\Lambda(t,\lambda_t)=\left(1+\frac{1}{t}\right)^{\tau}\lambda_t$ for some $\tau>0$, $\lambda_1>\frac{2L_{f,1}}{\mu_g}$, $\alpha\leq\frac{1}{L_{g,1}}$, $\beta_t=\frac{1}{2\lambda_t L_{g,1}}$, $\gamma_t\leq\frac{1}{\lambda_1 t^{\tau}}\sqrt{\frac{1-\rho}{6C_g}}$ and $K_t\equiv 1$, Algorithm 1 can obtain*

$$
\begin{aligned}
\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\leq&2c_\tau\Delta_1+2c_\lambda C_S+\frac{6c_\gamma C_g}{1-\rho}\mathrm{BLReg}(T)+\frac{6c_\gamma C_g D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}\\
&+\frac{2c_g'\lambda_T^2}{1-\rho_{\mathbf{z}}}H_{2,T}+\frac{2c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}E_{\mathbf{y},T}^f+\frac{2c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}E_{\mathbf{y},T}^g,
\end{aligned}
$$

where $\rho=\max\{\rho_{\mathbf{y}},\rho_{\mathbf{z}}\}$, $\rho_{\mathbf{y}}=1-\frac{1}{8\kappa_g}$, $\rho_{\mathbf{z}}=1-\frac{\alpha\mu_g}{2}$, Δ_1, c_τ, c_λ C_S, c_g, c_g', C_g are some constants.

Proof. We begin with equation (20) in Lemma C.4, which gives

$$
\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\leq 6\lambda_t^2L_{g,1}^2\left\|\mathbf{y}_{t+1}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2+3\lambda_t^2L_{g,1}^2\left\|\mathbf{z}_{t+1}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2.
$$

Substituting the bounds from Lemma C.6 and Lemma C.7 into the above inequality yields

$$
\begin{aligned}
\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\leq&6\lambda_t^2\rho_{\mathbf{y}}^{t-1}L_{g,1}^2\left\|\mathbf{y}_2-\mathbf{y}_{\lambda,1}^*(\mathbf{x}_1)\right\|^2+3\lambda_t^2\rho_{\mathbf{z}}^{t-1}L_{g,1}^2\left\|\mathbf{z}_2-\mathbf{y}_1^*(\mathbf{x}_1)\right\|^2\\
&+c_\lambda\lambda_t^2\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j(\lambda_{t-j}-\lambda_{t-1-j})^2}{\lambda_{t-j}^2\lambda_{t-1-j}^2}+C_g\lambda_t^2\sum_{j=0}^{t-2}\rho^j\left\|\mathbf{x}_{t-1-j}-\mathbf{x}_{t-j}\right\|^2\\
&+c_g\lambda_t^2\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_{t-j}(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1-j}(\mathbf{x},\mathbf{y})\right\|^2\\
&+c_g\lambda_t^2\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_{t-j}(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1-j}(\mathbf{x},\mathbf{y})\right\|^2\\
&+c_g'\lambda_t^2\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\sup_{\mathbf{x}}\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x})-\mathbf{y}_{t-j}^*(\mathbf{x})\right\|^2,
\end{aligned}
$$

where, for notational simplicity, we define

$$
\begin{aligned}
c_\lambda&:=6L_{g,1}^2\cdot\frac{1536L_{f,0}^2\kappa_g^3}{\mu_g^2},\\
C_g&:=6L_{g,1}^2\cdot 288\kappa_g^3+3L_{g,1}^2\cdot\frac{8\kappa_g^2}{\alpha\mu_g},\\
c_g&:=6L_{g,1}^2\cdot\frac{384\kappa_g}{\mu_g},\\
c_g'&:=3L_{g,1}^2\cdot\frac{8}{\alpha\mu_g},\\
\rho&:=\max\{\rho_{\mathbf{y}},\rho_{\mathbf{z}}\},
\end{aligned}
$$

and have

$$
\begin{aligned}
&\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\\
\leq& c_\tau\Delta_1+c_\lambda\sum_{t=1}^{T}\lambda_t^2\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j(\lambda_{t-j}-\lambda_{t-1-j})^2}{\lambda_{t-j}^2\lambda_{t-1-j}^2}+C_g\sum_{t=1}^{T}\lambda_t^2\sum_{j=0}^{t-2}\rho^j\left\|\mathbf{x}_{t-1-j}-\mathbf{x}_{t-j}\right\|^2\\
&+c_g\sum_{t=1}^{T}\lambda_t^2\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_{t-j}(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1-j}(\mathbf{x},\mathbf{y})\right\|^2\\
&+c_g\sum_{t=1}^{T}\lambda_t^2\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_{t-j}(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1-j}(\mathbf{x},\mathbf{y})\right\|^2\\
&+c_g'\sum_{t=1}^{T}\lambda_t^2\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\sup_{\mathbf{x}}\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x})-\mathbf{y}_{t-j}^*(\mathbf{x})\right\|^2,
\end{aligned}
\tag{32}
$$

where $\Delta_1:=6L_{g,1}^2\|\mathbf{y}_2-\mathbf{y}_{\lambda_1,1}^*(\mathbf{x}_1)\|^2+3L_{g,1}^2\|\mathbf{z}_2-\mathbf{y}_1^*(\mathbf{x}_1)\|^2+\|\widetilde{\nabla}\mathcal{L}_1^*(\mathbf{x}_1)-\nabla\mathcal{L}_1^*(\mathbf{x}_1)\|^2$. For the second term in Eq.(32), we define $S(T)$ and let $m=t-j$, it follows that

$$
S(T):=\sum_{t=2}^{T}\lambda_t^2\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\frac{(\lambda_{t-j}-\lambda_{t-1-j})^2}{\lambda_{t-j}^2\lambda_{t-1-j}^2}=\sum_{t=2}^{T}\lambda_t^2\sum_{m=2}^{t}\rho_{\mathbf{y}}^{t-m}\frac{(\lambda_m-\lambda_{m-1})^2}{\lambda_m^2\lambda_{m-1}^2},
$$

with $\lambda_m-\lambda_{m-1}=\lambda_1(m^\tau-(m-1)^\tau)=\lambda_1\tau\xi^{\tau-1}\leq\lambda_1\tau(m-1)^{\tau-1}$ for some $\xi\in[m-1,m]$, it holds that

$$
\frac{(\lambda_m-\lambda_{m-1})^2}{\lambda_m^2\lambda_{m-1}^2}\leq\frac{\lambda_1^2\tau^2(m-1)^{2\tau-2}}{\lambda_1^4m^{2\tau}(m-1)^{2\tau}}\leq\frac{\tau^2}{\lambda_1^2}\frac{1}{(m-1)^2m^{2\tau}}\approx\frac{\tau^2}{\lambda_1^2}\frac{1}{m^{2\tau+2}}.
$$

Thus we have

$$
S(T)=\frac{\tau^2}{\lambda_1^2}\sum_{t=2}^{T}\lambda_t^2\sum_{m=2}^{t}\frac{\rho_{\mathbf{y}}^{t-m}}{m^{2\tau+2}}=\tau^2\sum_{t=2}^{T}t^{2\tau}\sum_{m=2}^{t}\frac{\rho_{\mathbf{y}}^{t-m}}{m^{2\tau+2}}=\tau^2\sum_{t=2}^{T}t^{2\tau}\sum_{j=0}^{t-2}\frac{\rho^j}{(t-j)^{2\tau+2}},
$$

for $0\leq j\leq\lfloor\frac{t}{2}\rfloor$,

$$
\sum_{j=0}^{\lfloor\frac{t}{2}\rfloor}\frac{\rho_{\mathbf{y}}^j}{(t-j)^{2\tau+2}}\leq\left(\frac{2}{t}\right)^{2\tau+2}\sum_{j=0}^{\infty}\rho_{\mathbf{y}}^j\leq\left(\frac{2}{t}\right)^{2\tau+2}\frac{1}{1-\rho_{\mathbf{y}}},
$$

for $\lfloor\frac{t}{2}\rfloor+1\leq j\leq t-2$,

$$
\sum_{j=\lfloor\frac{t}{2}\rfloor+1}^{t-2}\frac{\rho_{\mathbf{y}}^j}{(t-j)^{2\tau+2}}\leq\frac{1}{2^{2\tau+2}}\sum_{j=\lfloor\frac{t}{2}\rfloor+1}^{t-2}\rho_{\mathbf{y}}^j\leq\frac{1}{2^{2\tau+2}}\cdot\frac{\rho_{\mathbf{y}}^{\lfloor\frac{t}{2}\rfloor+1}}{1-\rho_{\mathbf{y}}}.
$$

Finally, we have

$$
S(T)\leq\tau^2\sum_{t=2}^{T}\left(\frac{2^{2\tau+2}}{1-\rho_{\mathbf{y}}}\cdot\frac{1}{t^2}+\frac{\rho_{\mathbf{y}}^{\lfloor\frac{t}{2}\rfloor+1}t^{2\tau}}{2^{2\tau+2}(1-\rho_{\mathbf{y}})}\right)=:C_S,
$$

for any $\tau>0$, $C_S=O(1)$. Thus it holds that

$$
\begin{aligned}
&\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\\
\leq& c_\tau\Delta_1+c_\lambda C_S+\frac{C_g}{1-\rho}\sum_{t=1}^{T}\lambda_t^2\left\|\mathbf{x}_{t-1}-\mathbf{x}_t\right\|^2+\frac{c_g'\lambda_T^2}{1-\rho_{\mathbf{z}}}\sum_{t=1}^{T}\sup_{\mathbf{x}\in\mathcal{X}}\left\|\mathbf{y}_{t-1}^*(\mathbf{x})-\mathbf{y}_t^*(\mathbf{x})\right\|^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}\sum_{t=1}^{T}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y})\right\|^2\\
&+\frac{c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}\sum_{t=1}^{T}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y})\right\|^2.\\
\leq& c_\tau\Delta_1+c_\lambda C_S+\frac{C_g}{1-\rho}\sum_{t=1}^{T}\lambda_t^2\gamma_t^2\left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t,\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t),\gamma_t\right)\right\|^2\\
&+\frac{c_g'\lambda_T^2}{1-\rho_{\mathbf{z}}}\sum_{t=1}^{T}\sup_{\mathbf{x}\in\mathcal{X}}\left\|\mathbf{y}_{t-1}^*(\mathbf{x})-\mathbf{y}_t^*(\mathbf{x})\right\|^2\\
&+\frac{c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}\sum_{t=1}^{T}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y})\right\|^2\\
&+\frac{c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}\sum_{t=1}^{T}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y})\right\|^2\\
\leq& c_\tau\Delta_1+c_\lambda C_S+\frac{C_g}{1-\rho}\sum_{t=1}^{T}\lambda_t^2\gamma_t^2\Big(3\left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t,\nabla F_t(\mathbf{x}_t),\gamma_t\right)\right\|^2+3\left\|\nabla F_t(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\\
&+\left\|\nabla\mathcal{L}_t^*(\mathbf{x}_t)-\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\Big)+\frac{c_g'\lambda_T^2}{1-\rho_{\mathbf{z}}}\sum_{t=1}^{T}\sup_{\mathbf{x}\in\mathcal{X}}\left\|\mathbf{y}_{t-1}^*(\mathbf{x})-\mathbf{y}_t^*(\mathbf{x})\right\|^2\\
&+\frac{c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}\sum_{t=1}^{T}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y})\right\|^2\\
&+\frac{c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}\sum_{t=1}^{T}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y})\right\|^2.
\end{aligned}
$$

Now suppose that for sufficiently small γ_t such that $\lambda_t^2\gamma_t^2\leq c_\gamma$ for some constant $c_\gamma>0$, we obtain

$$
\begin{aligned}
&\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\\
\leq& c_\tau\Delta_1+c_\lambda C_S+\frac{3c_\gamma C_g}{1-\rho}\mathrm{BLReg}(T)+\frac{3c_\gamma C_g D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}+\frac{3c_\gamma C_g}{1-\rho}\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\\
&+\frac{c_g'\lambda_T^2}{1-\rho_{\mathbf{z}}}H_{2,T}+\frac{c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}E_{\mathbf{y},T}^f+\frac{c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}E_{\mathbf{y},T}^g.
\end{aligned}
$$

Let $1-\frac{3c_\gamma C_g}{1-\rho}\geq\frac{1}{2}$, which means

$$
\begin{aligned}
c_\gamma\leq&\frac{1-\rho}{6C_g}\\
\gamma\leq&\sqrt{\frac{1-\rho}{6\lambda_t^2C_g}}=\frac{1}{\lambda_1 t^\tau}\sqrt{\frac{1-\rho}{6C_g}},
\end{aligned}
$$

finally we finish the proof that

$$
\begin{aligned}
\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\leq&2c_\tau\Delta_1+2c_\lambda C_S+\frac{6c_\gamma C_g}{1-\rho}\mathrm{BLReg}(T)+\frac{6c_\gamma C_g D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}\\
&+\frac{2c_g'\lambda_T^2}{1-\rho_{\mathbf{z}}}H_{2,T}+\frac{2c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}E_{\mathbf{y},T}^f+\frac{2c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}E_{\mathbf{y},T}^g.
\end{aligned}
$$

□

Theorem C.9 (Restatement of Theorem 3.2)**.** *Under Assumptions 2.1, 2.2, let* $\Lambda(t,\lambda_t)=\left(1+\frac{1}{t}\right)^{\tau}\lambda_t$ *for some* $\tau\in(0,\frac{1}{2})$, $\lambda_1>\frac{2L_{f,1}}{\mu_g}$, $\alpha\le\frac{1}{L_{g,1}}$, $\beta_t=\frac{1}{2\lambda_t L_{g,1}}$, $\gamma_t<\min\{\frac{1}{2L_F},\frac{1}{8\lambda_1 t^{\tau}}\sqrt{\frac{1-\rho}{3C_g}}\}$ *and* $K_t\equiv 1$*, Algorithm 1 can guarantee*

$$\mathrm{BLReg}(T)\le O\left(T^{\tau}(1+V_T)+T^{1-2\tau}+T^{2\tau}(H_{2,T}+E^f_{\mathbf{y},T}+E^g_{\mathbf{y},T})\right),$$

where ρ*,* C_g *are some constants and*

$$E^f_{\mathbf{y},T}=\sum_{t=2}^{T}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}f_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}f_{t-1}(\mathbf{x},\mathbf{y})\right\|^2,$$
$$E^g_{\mathbf{y},T}=\sum_{t=2}^{T}\sup_{\mathbf{x},\mathbf{y}}\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x},\mathbf{y})-\nabla_{\mathbf{y}}g_{t-1}(\mathbf{x},\mathbf{y})\right\|^2.$$

Proof. After substituting Lemma C.8 into Lemma C.1, we obtain

$$\begin{aligned}\mathrm{BLReg}(T)\le&\frac{16M}{\gamma_T}+\frac{8}{\gamma_T}V_T+16D_1^2\sum_{t=1}^{T}\frac{1}{\lambda_t^2}+32c_\tau\Delta_1+32c_\lambda C_S+\frac{96c_\gamma C_g}{1-\rho}\mathrm{BLReg}(T)\\&+\frac{96c_\gamma C_g D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}+\frac{32c'_g\lambda_T^2}{1-\rho_{\mathbf{z}}}H_{2,T}+\frac{32c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}E^f_{\mathbf{y},T}+\frac{32c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}E^g_{\mathbf{y},T}.\end{aligned}$$

Let $1-\frac{96c_\gamma C_g}{1-\rho}\ge\frac{1}{2}$, $c_\gamma\le\frac{1-\rho}{192C_g}$, which means

$$\begin{aligned}\lambda_t^2\gamma_t^2\le&\frac{1-\rho}{192C_g}\\ \gamma_t\le&\frac{1}{8\lambda_1 t^{\tau}}\sqrt{\frac{1-\rho}{3C_g}}.\end{aligned}$$

Thus we finally have

$$\begin{aligned}\mathrm{BLReg}(T)\le&\frac{16M}{\gamma_T}+\frac{8}{\gamma_T}V_T+\frac{16D_1^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}+32c_\tau\Delta_1+32c_\lambda C_S+\frac{96c_\gamma C_g D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}\frac{1}{t^{2\tau}}\\&+\frac{32c'_g\lambda_T^2}{1-\rho_{\mathbf{z}}}H_{2,T}+\frac{32c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}E^f_{\mathbf{y},T}+\frac{32c_g\lambda_T^2}{1-\rho_{\mathbf{y}}}E^g_{\mathbf{y},T}\\=&O\left(T^{\tau}(1+V_T)+T^{1-2\tau}+T^{2\tau}(H_{2,T}+E^f_{\mathbf{y},T}+E^g_{\mathbf{y},T})\right).\end{aligned}$$

□

D Proof of Section 4

In this section, we present a regret upper bound proof for AF^{2}OBO. AF^{2}OBO improves the upper bound of the hypergradient error for the approximation problem (P3) by using a dynamic inner iteration technique, which eliminates the dependence on $H_{2,T}$, as shown in Lemma D.1 below. Finally, following the same proof process as F^{2}OBO, we give the proof in Theorem D.2.

Lemma D.1. *Under Assumptions 2.1 and 2.2, let* $\Lambda(t,\lambda_t)=\left(1+\frac{1}{t}\right)^{\tau}\lambda_t$ *for some* $\tau>0$, $\lambda_1>\frac{2L_{f,1}}{\mu_g}$, $\delta_{\mathbf{y}}=\frac{1}{\sqrt{T}}$ *and* $\delta_{\mathbf{z}}=\frac{1}{\sqrt{T^{1+2\tau}}}$*, the approximation error of hypergradient in Algorithm 2 can be bounded as*

$$\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\le 24\kappa_g^2+3\lambda_1^2\kappa_g^2.$$

Proof. We first bound the hypergradient approximation error as follows:

$$\begin{aligned}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 &\leq 6\lambda_t^2 L_{g,1}^2 \left\|\mathbf{y}_{t+1} - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2 + 3\lambda_t^2 L_{g,1}^2 \left\|\mathbf{z}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2 \\ &\leq 6\lambda_t^2 L_{g,1}^2 \frac{4\delta_{\mathbf{y}}^2}{\lambda_t^2\mu_g^2} + 3\lambda_t^2 L_{g,1}^2 \frac{\delta_{\mathbf{z}}^2}{\mu_g^2} \\ &= 24\kappa_g^2\delta_{\mathbf{y}}^2 + 3\lambda_t^2\kappa_g^2\delta_{\mathbf{z}}^2. \end{aligned} \tag{33}$$

The second inequality follows from the error criteria

$$\frac{\lambda_t\mu_g}{2}\left\|\mathbf{y}_{t+1} - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\| \leq \left\|\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x}_t, \mathbf{y}_{t+1}, \lambda_t)\right\| \leq \delta_{\mathbf{y}},$$

and

$$\mu_g\left\|\mathbf{z}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\right\| \leq \left\|\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{z}_{t+1})\right\| \leq \delta_{\mathbf{z}},$$

which are implied by the μ_g-strong convexity of g_t and the $(\lambda_t\mu_g/2)$-strong convexity of $\mathcal{L}_t$, respectively. Therefore, summing over $t = 1, \ldots, T$ gives

$$\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 \leq 24\kappa_g^2\delta_{\mathbf{y}}^2 T + 3\lambda_T^2\kappa_g^2\delta_{\mathbf{z}}^2 T.$$

By choosing $\delta_{\mathbf{y}} = \frac{1}{\sqrt{T}}$ and $\delta_{\mathbf{z}} = \frac{1}{\sqrt{T^{1+2\tau}}}$, we obtain

$$\sum_{t=1}^{T}\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 \leq 24\kappa_g^2 + 3\lambda_1^2\kappa_g^2.$$

□

Theorem D.2 (Restatement of Theorem 4.1)**.** *Under Assumptions 2.1, 2.2, let* $\Lambda(t, \lambda_t) = \left(1 + \frac{1}{t}\right)^{\tau}\lambda_t$ *for some* $\tau > 0$*,* $\lambda_1 > \frac{2L_{f,1}}{\mu_g}$*,* $\alpha \leq \frac{1}{L_{g,1}}$*,* $\beta_t \leq \frac{1}{2\lambda_t L_{g,1}}$*,* $\gamma \leq \frac{1}{2L_F}$ *for all* $t \in [T]$*,* $\delta_{\mathbf{y}} = \frac{1}{\sqrt{T}}$ *and* $\delta_{\mathbf{z}} = \frac{1}{\sqrt{T^{1+2\tau}}}$*, Algorithm 2 can guarantee*

$$\mathrm{BLReg}(T) \leq \frac{16M}{\gamma} + \frac{8}{\gamma}V_T + \frac{16D_1^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + 384\kappa_g^2 + 48\lambda_1^2\kappa_g^2 = O\left(\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + V_T\right).$$

The total number of inner iterations $\mathcal{I}_T$ *satisfies*

$$\mathcal{I}_T \leq O\left(T^{1+2\tau} + T^{2\tau}H_{2,T}\right).$$

Proof. We begins with Lemma C.1 and substitute Lemma D.1 into it to obtain

$$\mathrm{BLReg}(T) \leq \frac{16M}{\gamma} + \frac{8}{\gamma}V_T + \frac{16D_1^2}{\lambda_1^2}\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + 384\kappa_g^2 + 48\lambda_1^2\kappa_g^2 = O\left(\sum_{t=1}^{T}\frac{1}{t^{2\tau}} + V_T\right),$$

where $\gamma_t \equiv \gamma \leq \frac{1}{2L_F}$, $\delta_{\mathbf{y}} = \frac{1}{\sqrt{T}}$ and $\delta_{\mathbf{z}} = \frac{1}{\sqrt{T^{1+2\tau}}}$. Now consider the two inner-loop iteration process. The sequences $\{\mathbf{y}_t^k\}_{k=1}^{K_t}$ and $\{\mathbf{z}_t^m\}_{m=1}^{M_t}$ are generated by

$$\begin{aligned}\mathbf{y}_t^{k+1} &\leftarrow \mathbf{y}_t^k - \beta_t\left(\nabla_{\mathbf{y}} f_t(\mathbf{x}_t, \mathbf{y}_t^k) + \lambda_t\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{y}_t^k)\right), \quad \text{and} \\ \mathbf{z}_t^{m+1} &\leftarrow \mathbf{z}_t^m - \alpha\nabla_{\mathbf{y}} g_t(\mathbf{x}_t, \mathbf{z}_t^m),\end{aligned}$$

respectively. For $\{\mathbf{y}_t^k\}_{k=1}^{K_t}$, it holds that

$$\frac{\delta_{\mathbf{y}}^2}{4\lambda_t^2 L_{g,1}^2} \leq \left\|\mathbf{y}_t^{K_t} - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2 \leq \rho_{\mathbf{y}}^{K_t - 1}\left\|\mathbf{y}_t^1 - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2, \tag{34}$$

where $\rho_{\mathbf{y}} \in (0, 1)$. The second inequality follows from equation (17) in Lemma C.2 under the stepsize condition $\beta_t \leq 1/(2\lambda_t L_{g,1})$ for all $t \in [T]$. The first inequality follows from the fact that $\mathcal{L}_t(\mathbf{x}_t, \cdot, \lambda_t)$ is $(2\lambda_t L_{g,1})$-smooth with respect to $\mathbf{y}$. Hence,

$$2\lambda_t L_{g,1}\left\|\mathbf{y}_t^{K_t} - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\| \geq \left\|\nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x}_t, \mathbf{y}_t^{K_t}, \lambda_t)\right\| > \delta_{\mathbf{y}}.$$

After rearranging Eq.(34), we have

$$\begin{aligned}\rho_{\mathbf{y}}^{K_t-1} \geq & \frac{\delta_{\mathbf{y}}^2}{4\lambda_t^2 L_{g,1}^2 \left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2} \\ K_t \leq & 1 + \frac{4\lambda_t^2 L_{g,1}^2}{\ln \rho_{\mathbf{y}}^{-1}} \left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 + \frac{2\ln \delta_{\mathbf{y}}}{\ln \rho_{\mathbf{y}}}.\end{aligned}$$

Let $\mathcal{I}_T^{\mathbf{y}}$ denotes the total iteration number of $\mathbf{y}$ and $\mathcal{I}_T^{\mathbf{z}}$ denote the total iteration number of $\mathbf{z}$, it immediately follows that

$$\mathcal{I}_T^{\mathbf{y}} := \sum_{t=1}^{T} K_t \leq \left(1 + \frac{2\ln \delta_{\mathbf{y}}}{\ln \rho_{\mathbf{y}}}\right) T + \frac{4L_{g,1}^2}{\ln \rho_{\mathbf{y}}^{-1}} \sum_{t=1}^{T} \lambda_t^2 \left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2. \tag{35}$$

Then it begins with

$$\begin{aligned}& \sum_{t=1}^{T} \lambda_t^2 \left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 \\ \leq & 2\sum_{t=1}^{T} \lambda_t^2 \left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})\right\|^2 + 2\sum_{t=1}^{T} \lambda_t^2 \left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1}) - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 \\ \leq & 2\sum_{t=1}^{T} \lambda_t^2 \frac{4\delta_{\mathbf{y}}^2}{\lambda_t^2 \mu_g^2} + 2\sum_{t=1}^{T} \lambda_t^2 \left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1}) - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2.\end{aligned}$$

Remind of equation (18), we have

$$\begin{aligned}\left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1}) - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 \leq & \frac{6L_{f,0}^2}{\lambda_t^2 \mu_g^2} + 6\kappa_g^2 \left\|\mathbf{x}_{t-1} - \mathbf{x}_t\right\|^2 + 6\left\|\mathbf{y}^*_{t-1}(\mathbf{x}_t) - \mathbf{y}^*_t(\mathbf{x}_t)\right\|^2 \\ \leq & \frac{6L_{f,0}^2}{\lambda_t^2 \mu_g^2} + 6\gamma^2 \kappa_g^2 L_{\mathcal{L},0}^2 + 6\left\|\mathbf{y}^*_{t-1}(\mathbf{x}_t) - \mathbf{y}^*_t(\mathbf{x}_t)\right\|^2,\end{aligned}$$

where the last inequality comes from Lemma B.6. Thus we obtain

$$\sum_{t=1}^{T} \lambda_t^2 \left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 \leq \frac{8\delta_{\mathbf{y}}^2}{\mu_g^2} T + \frac{12L_{f,0}^2}{\mu_g^2} T + 12\gamma^2 \lambda_1^2 \kappa_g^2 L_{\mathcal{L},0}^2 T^{1+2\tau} + 6\lambda_1^2 T^{2\tau} H_{2,T}.$$

After substituting it into Eq.(35), we finally obtain

$$\begin{aligned}\mathcal{I}_T^{\mathbf{z}} \leq & \left(1 + \frac{2\ln \delta_{\mathbf{y}}}{\ln \rho_{\mathbf{y}}}\right) T + \frac{4L_{g,1}^2}{\ln \rho_{\mathbf{y}}^{-1}} \left(\frac{8\delta_{\mathbf{y}}^2}{\mu_g^2} T + \frac{12L_{f,0}^2}{\mu_g^2} T + 12\gamma^2 \lambda_1^2 \kappa_g^2 L_{\mathcal{L},0}^2 T^{1+2\tau} + 6\lambda_1^2 T^{2\tau} H_{2,T}\right) \\ = & O\left(T^{1+2\tau} + T^{2\tau} H_{2,T}\right).\end{aligned}$$

Based on the similar analysis process for $\mathbf{y}$, we have the following inequality

$$\begin{aligned}\rho_{\mathbf{z}}^{M_t-1} \geq & \frac{\delta_{\mathbf{z}}^2}{L_{g,1}^2 \left\|\mathbf{z}_t - \mathbf{y}^*_t(\mathbf{x}_t)\right\|^2} \\ M_t \leq & 1 + \frac{L_{g,1}^2}{\ln \rho_{\mathbf{z}}^{-1}} \left\|\mathbf{z}_t - \mathbf{y}^*_t(\mathbf{x}_t)\right\|^2 + \frac{2\ln \delta_{\mathbf{z}}}{\ln \rho_{\mathbf{z}}}.\end{aligned}$$

Thus we have

$$\begin{aligned}\mathcal{I}_T^{\mathbf{z}} \leq & \left(1 + \frac{2\ln \delta_{\mathbf{z}}}{\ln \rho_{\mathbf{z}}}\right) T + \frac{L_{g,1}^2}{\ln \rho_{\mathbf{z}}^{-1}} \sum_{t=1}^{T} \left\|\mathbf{z}_t - \mathbf{y}^*_t(\mathbf{x}_t)\right\|^2 \\ \leq & \left(1 + \frac{2\ln \delta_{\mathbf{z}}}{\ln \rho_{\mathbf{z}}}\right) T + \frac{2L_{g,1}^2}{\ln \rho_{\mathbf{z}}^{-1}} \left(\sum_{t=1}^{T} \left\|\mathbf{z}_t - \mathbf{y}^*_{t-1}(\mathbf{x}_{t-1})\right\|^2 + \sum_{t=1}^{T} \left\|\mathbf{y}^*_{t-1}(\mathbf{x}_{t-1}) - \mathbf{y}^*_t(\mathbf{x}_t)\right\|^2\right)\end{aligned}$$

$$\leq \left(1 + \frac{2\ln \delta_{\mathbf{z}}}{\ln \rho_{\mathbf{z}}}\right) T + \frac{2\kappa_g^2 \delta_{\mathbf{z}}^2}{\ln \rho_{\mathbf{z}}^{-1}} T + \frac{4\gamma^2 \kappa_g^2 L_{\mathcal{L},0}^2 L_{g,1}^2}{\ln \rho_{\mathbf{z}}^{-1}} T + \frac{4L_{g,1}^2}{\ln \rho_{\mathbf{z}}^{-1}} H_{2,T} = O\left(T \ln T + H_{2,T}\right).$$

It is true that

$$\mathcal{I}_T = \mathcal{I}_T^{\mathbf{y}} + \mathcal{I}_T^{\mathbf{z}} \leq O\left(T^{1+2\tau} + T^{2\tau} H_{2,T}\right).$$

□

E Proof of Section 5

We first present Lemma E.1, which establishes a preliminary regret upper bound for the fully first-order algorithm applied to the approximation problem equation (P3) in the stochastic OBO setting. Lemmas E.2 and E.3 provide the inner-loop iteration error bounds under stochastic noise, respectively. These bounds further lead to the approximate stochastic hypergradient error bound in Lemma E.4, which ultimately yields the proof of Theorem C.9.

Lemma E.1. *Under Assumptions 2.1, 2.2, suppose that $\gamma_t \leq \frac{1}{2L_F}$ for all $t \in [T]$. Consider the update*

$$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \gamma_t \mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right).$$

Then, we obtain

$$\begin{aligned}\mathbb{E}\left[\mathrm{BLReg}(T)\right] \leq & \sum_{t=1}^{T} \frac{8}{\gamma_t} \mathbb{E}\left[F_t(\mathbf{x}_t) - F_t(\mathbf{x}_{t+1})\right] + 24 D_1^2 \sum_{t=1}^{T} \frac{1}{\lambda_t^2} + 72\sigma^2 \sum_{t=1}^{T} \frac{\lambda_t^2}{\widehat{B}_t} \\ & + 24 \sum_{t=1}^{T} \mathbb{E}\left[\left\|\nabla \mathcal{L}_t^*(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\right].\end{aligned}$$

Proof. We begin with the L_F-smoothness of F_t. By Lemma B.3, we have

$$\begin{aligned}& F_t(\mathbf{x}_{t+1}) - F_t(\mathbf{x}_t) \\ \leq & \left\langle \nabla F_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \right\rangle + \frac{L_F}{2} \left\|\mathbf{x}_{t+1} - \mathbf{x}_t\right\|^2 \\ = & -\gamma_t \left\langle \nabla F_t(\mathbf{x}_t), \mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right)\right\rangle + \frac{\gamma_t^2 L_F}{2} \left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right)\right\|^2 \\ = & -\gamma_t \left\langle \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right)\right\rangle + \frac{\gamma_t^2 L_F}{2} \left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right)\right\|^2 \\ & - \gamma_t \left\langle \nabla F_t(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right)\right\rangle \\ \leq & -\gamma_t \left(1 - \frac{\gamma_t L_F}{2}\right) \left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right)\right\|^2 + \frac{\gamma_t}{2} \left\|\nabla F_t(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t)\right\|^2 \\ & + \frac{\gamma_t}{2} \left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right)\right\|^2 \\ = & -\frac{\gamma_t}{2}\left(1 - \gamma_t L_F\right) \left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right)\right\|^2 + \frac{\gamma_t}{2} \left\|\nabla F_t(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t)\right\|^2.\end{aligned}$$

Let $\gamma_t \leq \frac{1}{2L_F}$, rearranging the above inequality gives

$$\left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right)\right\|^2 \leq \frac{4}{\gamma_t}\left(F_t(\mathbf{x}_t) - F_t(\mathbf{x}_{t+1})\right) + 2\left\|\nabla F_t(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t)\right\|^2.$$

Moreover, by

$$\frac{1}{2}\left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \nabla F_t(\mathbf{x}_t), \gamma_t\right)\right\|^2 \leq \left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t), \gamma_t\right)\right\|^2 + \left\|\nabla F_t(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t)\right\|^2,$$

it follows that

$$\left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \nabla F_t(\mathbf{x}_t), \gamma_t\right)\right\|^2 \leq \frac{8}{\gamma_t}\left(F_t(\mathbf{x}_t) - F_t(\mathbf{x}_{t+1})\right) + 8\left\|\nabla F_t(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t)\right\|^2$$

$$\leq \frac{8}{\gamma_t}\left(F_t(\mathbf{x}_t) - F_t(\mathbf{x}_{t+1})\right) + 24\left\|\nabla F_t(\mathbf{x}_t) - \nabla \mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 + 24\left\|\nabla \mathcal{L}_t^*(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 + 24\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t;\mathcal{S}_t)\right\|^2. \tag{36}$$

Note that, under Assumption 2.3, we naturally have

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}_t}\left[\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t;\mathcal{S}_t)\right\|^2\right] =& \mathbb{E}_{\mathcal{S}_t}\left[\left\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t,\mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t,\mathbf{y}_{t+1};\xi_t)\right.\right. \\
&+\lambda_t\left(\nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{y}_{t+1};\zeta_t)\right) \\
&\left.\left.+\lambda_t\left(\nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{z}_{t+1}) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{z}_{t+1})\right)\right\|^2\right] \\
=& \mathbb{E}_{\xi_t}\left[\left\|\nabla_{\mathbf{x}} f_t(\mathbf{x}_t,\mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} f_t(\mathbf{x}_t,\mathbf{y}_{t+1};\xi_t)\right\|^2\right] \\
&+\lambda_t^2\mathbb{E}_{\zeta_t}\left[\left\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{y}_{t+1}) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{y}_{t+1};\zeta_t)\right\|^2\right] \\
&+\lambda_t^2\mathbb{E}_{\zeta_t}\left[\left\|\nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{z}_{t+1}) - \nabla_{\mathbf{x}} g_t(\mathbf{x}_t,\mathbf{z}_{t+1};\zeta_t)\right\|^2\right] \\
\leq& (1+2\lambda_t^2)\frac{\sigma^2}{\widehat{B}_t} \leq \frac{3\lambda_t^2\sigma^2}{\widehat{B}_t},
\end{aligned}$$

and substituting the above inequality into Eq.(36), we complete the proof by

$$\begin{aligned}
\mathbb{E}\left[\mathrm{BLReg}(T)\right] \leq& \sum_{t=1}^{T}\frac{8}{\gamma_t}\mathbb{E}\left[F_t(\mathbf{x}_t) - F_t(\mathbf{x}_{t+1})\right] + 24D_1^2\sum_{t=1}^{T}\frac{1}{\lambda_t^2} + 72\sigma^2\sum_{t=1}^{T}\frac{\lambda_t^2}{\widehat{B}_t} \\
&+ 24\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\mathcal{L}_t^*(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\right].
\end{aligned}$$

□

Lemma E.2. *Under Assumptions 2.1, 2.2 and 2.3, let $\beta_t \leq \frac{1}{2\lambda_t L_{g,1}}$, Algorithm 3 can obtain*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{y}_{t+1} - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\right] \\
\leq& \rho_{\mathbf{y}}^{t-1}\mathbb{E}\left[\left\|\mathbf{y}_2 - \mathbf{y}_{\lambda_1,1}^*(\mathbf{x}_1)\right\|^2\right] + \frac{4\sigma^2}{\mu_g}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j\beta_{t-j}\lambda_{t-j}}{B_{t-j}} + \frac{12L_{f,0}^2 q_t^{K_t}}{\mu_g^2}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}} \\
&+ 12\kappa_g^2 q_t^{K_t}\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\mathbb{E}\left[\left\|\mathbf{x}_{t-j} - \mathbf{x}_{t-1-j}\right\|^2\right] + 12q_t^{K_t}\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\mathbb{E}\left[\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2\right],
\end{aligned}$$

where $q_t = 1 - \frac{\beta_t\lambda_t\mu_g}{2}$. For each $t \in [T]$, we choose $K_t > 0$ such that $2q_t^{K_t} \leq \rho_{\mathbf{y}}$ for some constant $\rho_{\mathbf{y}} > 0$.

Proof. Let $\widetilde{\mathbf{g}}_{\mathbf{y}} := \nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x}_t,\mathbf{y}_t^k,\lambda_t;\mathcal{S}_t)$, $\mathbf{g}_{\mathbf{y}} := \nabla_{\mathbf{y}}\mathcal{L}_t(\mathbf{x}_t,\mathbf{y}_t^k,\lambda_t)$. With the update rule

$$\mathbf{y}_t^{k+1} \leftarrow \mathbf{y}_t^k - \beta_t\widetilde{\mathbf{g}}_{\mathbf{y}},$$

we have

$$\begin{aligned}
&\mathbb{E}_{\mathcal{S}_t^k}\left[\left\|\mathbf{y}_t^{k+1} - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\right] \\
=& \mathbb{E}_{\mathcal{S}_t^k}\left[\left\|\mathbf{y}_t^k - \beta_t\widetilde{\mathbf{g}}_{\mathbf{y}} - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\right] \\
\leq& \left\|\mathbf{y}_t^k - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2 - 2\beta_t\mathbb{E}_{\mathcal{S}_t^k}\left[\left\langle\widetilde{\mathbf{g}}_{\mathbf{y}}, \mathbf{y}_t^k - \mathbf{y}_{\lambda_t,t}^*((\mathbf{x}_t)\right\rangle\right] + \beta_t^2\mathbb{E}_{\mathcal{S}_t^k}\left[\left\|\widetilde{\mathbf{g}}_{\mathbf{y}}\right\|^2\right] \\
\overset{(i)}{=}& \left\|\mathbf{y}_t^k - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2 - 2\beta_t\left\langle\mathbf{g}_{\mathbf{y}}, \mathbf{y}_t^k - \mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\rangle + \beta_t^2\left\|\mathbf{g}_{\mathbf{y}}\right\|^2 + \beta_t^2(1+\lambda_t^2)\frac{\sigma^2}{B_t}
\end{aligned}$$

$$\overset{(ii)}{\leq} \left\|\mathbf{y}_t^k - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 - \beta_t\frac{\lambda_t\mu_g}{2}\left\|\mathbf{y}_t^k - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 - \left(\frac{\beta_t}{2\lambda_t L_{g,1}} - \beta_t^2\right)\left\|\mathbf{g}_{\mathbf{y}}\right\|^2 + \beta_t^2(1+\lambda_t^2)\frac{\sigma^2}{B_t}$$
$$\leq q_t\left\|\mathbf{y}_t^k - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2 + 2\beta_t^2\lambda_t^2\frac{\sigma^2}{B_t}$$

where (i) comes from the fact that

$$\mathbb{E}_{\mathcal{S}_t^k}\left[\left\|\mathbf{g}_{\mathbf{y}} - \widetilde{\mathbf{g}}_{\mathbf{y}}\right\|^2\right] = \mathbb{E}_{\mathcal{S}_t^k}\left[\left\|\nabla_{\mathbf{y}} f_t(\mathbf{x}_t,\mathbf{y}_t^k) - \nabla_{\mathbf{y}} f_t(\mathbf{x}_t,\mathbf{y}_t^k;\xi_t^k)\right.\right.$$
$$\left.\left. + \lambda_t\left(\nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t^k) - \nabla_{\mathbf{y}} g_t(\mathbf{x}_t,\mathbf{y}_t^k;\zeta_t^k)\right)\right\|^2\right] \leq (1+\lambda_t^2)\frac{\sigma^2}{B_t},$$

and (ii) comes from the $(\lambda_t\mu_g/2)$-strongly convexity and $(2\lambda_t L_{g,1})$-smoothness of $\mathcal{L}_t$ w.r.t. $\mathbf{y}$, and the last inequality comes from $\beta_t \leq \frac{1}{2\lambda_t L_{g,1}}$, let $q_t := 1 - \frac{\beta_t\lambda_t\mu_g}{2}$, we then obtain

$$\mathbb{E}\left[\left\|\mathbf{y}_{t+1} - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2\right] \leq q_t^{K_t}\mathbb{E}\left[\left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2\right] + 2\beta_t^2\lambda_t^2\frac{\sigma^2}{B_t}\sum_{j=0}^{K_t-1} q_t^j$$
$$\leq q_t^{K_t}\mathbb{E}\left[\left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2\right] + \frac{4\beta_t\lambda_t\sigma^2}{\mu_g B_t}. \tag{37}$$

Recall from equation (18) that we also have

$$\left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2$$
$$\leq 2\left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})\right\|^2 + 2\left\|\mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1}) - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2$$
$$\leq 2\left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})\right\|^2 + \frac{12L_{f,0}^2}{\lambda_t^2\mu_g^2} + 12\kappa_g^2\left\|\mathbf{x}_{t-1} - \mathbf{x}_t\right\|^2 + 12\left\|\mathbf{y}^*_{t-1}(\mathbf{x}_t) - \mathbf{y}^*_t(\mathbf{x}_t)\right\|^2,$$

and substituting the above inequality into equation (37), we obtain

$$\mathbb{E}\left[\left\|\mathbf{y}_{t+1} - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2\right] \leq 2q_t^{K_t}\mathbb{E}\left[\left\|\mathbf{y}_t - \mathbf{y}^*_{\lambda_{t-1},t-1}(\mathbf{x}_{t-1})\right\|^2\right] + \frac{4\beta_t\lambda_t\sigma^2}{\mu_g B_t} + \frac{12q_t^{K_t}L_{f,0}^2}{\lambda_t^2\mu_g^2}$$
$$+ 12q_t^{K_t}\kappa_g^2\mathbb{E}\left[\left\|\mathbf{x}_{t-1} - \mathbf{x}_t\right\|^2\right] + 12q_t^{K_t}\mathbb{E}\left[\left\|\mathbf{y}^*_{t-1}(\mathbf{x}_t) - \mathbf{y}^*_t(\mathbf{x}_t)\right\|^2\right].$$

Telescoping the above inequality, we complete the proof by

$$\mathbb{E}\left[\left\|\mathbf{y}_{t+1} - \mathbf{y}^*_{\lambda_t,t}(\mathbf{x}_t)\right\|^2\right]$$
$$\leq \mathbb{E}\left[\left\|\mathbf{y}_2 - \mathbf{y}^*_{\lambda_1,1}(\mathbf{x}_1)\right\|^2\right]\prod_{j=0}^{t-2} 2q_j^{K_j} + \frac{4\sigma^2}{\mu_g}\sum_{j=0}^{t-2}\frac{\beta_{t-j}\lambda_{t-j}}{B_{t-j}}\prod_{m=1}^{j} 2q_{t+1-m}^{K_{t+1-m}}$$
$$+ \frac{6L_{f,0}^2}{\mu_g^2}\sum_{j=0}^{t-2}\frac{1}{\lambda_{t-j}^2}\prod_{m=0}^{j} 2q_{t-m}^{K_{t-m}} + 6\kappa_g^2\sum_{j=0}^{t-2}\mathbb{E}\left[\left\|\mathbf{x}_{t-j} - \mathbf{x}_{t-1-j}\right\|^2\right]\prod_{m=0}^{j} 2q_{t-m}^{K_{t-m}}$$
$$+ 6\sum_{j=0}^{t-2}\mathbb{E}\left[\left\|\mathbf{y}^*_{t-1-j}(\mathbf{x}_{t-j}) - \mathbf{y}^*_{t-j}(\mathbf{x}_{t-j})\right\|^2\right]\prod_{m=0}^{j} 2q_{t-m}^{K_{t-m}}$$
$$\leq \rho_{\mathbf{y}}^{t-1}\mathbb{E}\left[\left\|\mathbf{y}_2 - \mathbf{y}^*_{\lambda_1,1}(\mathbf{x}_1)\right\|^2\right] + \frac{4\sigma^2}{\mu_g}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j\beta_{t-j}\lambda_{t-j}}{B_{t-j}} + \frac{12L_{f,0}^2 q_t^{K_t}}{\mu_g^2}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2}$$
$$+ 12\kappa_g^2 q_t^{K_t}\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\mathbb{E}\left[\left\|\mathbf{x}_{t-j} - \mathbf{x}_{t-1-j}\right\|^2\right] + 12q_t^{K_t}\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\mathbb{E}\left[\left\|\mathbf{y}^*_{t-1-j}(\mathbf{x}_{t-j}) - \mathbf{y}^*_{t-j}(\mathbf{x}_{t-j})\right\|^2\right],$$

where the last inequality follows by assuming that, for all $t \in [T]$, K_t is sufficiently large such that $2q_t^{K_t} \leq \rho_{\mathbf{y}}$ for some constant $\rho_{\mathbf{y}}$. □

Lemma E.3. *Under Assumptions 2.1, 2.2 and 2.3, let $\alpha_t \leq \frac{1}{L_{g,1}}$, Algorithm 3 can obtain*

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{z}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right] \\
\leq&\rho_{\mathbf{z}}^{t-1}\mathbb{E}\left[\left\|\mathbf{z}_2 - \mathbf{y}_1^*(\mathbf{x}_1)\right\|^2\right] + \frac{\sigma^2}{\mu_g}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{z}}^j\alpha_{t-j}}{B_{t-j}} + 4p_t^{K_t}\kappa_g^2\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\mathbb{E}\left[\left\|\mathbf{x}_{t-1-j} - \mathbf{x}_{t-j}\right\|^2\right] \\
&+ 4p_t^{K_t}\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\mathbb{E}\left[\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2\right],
\end{aligned}$$

where $p_t = 1 - \alpha_t\mu_g$. For each $t \in [T]$, we choose $K_t > 0$ such that $2p_t^{K_t} \leq \rho_{\mathbf{z}}$ for some constant $\rho_{\mathbf{z}} > 0$.

Proof. By applying a derivation similar to that in Lemma E.2, we have

$$\begin{aligned}
&\mathbb{E}_{\zeta_t^k}\left[\left\|\mathbf{z}_t^{k+1} - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right] \\
=&\mathbb{E}_{\zeta_t^k}\left[\left\|\mathbf{z}_t^k - \alpha_t\nabla_{\mathbf{y}}g_t(\mathbf{x}_t, \mathbf{z}_t^k; \zeta_t^k) - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right] \\
\leq&\left\|\mathbf{z}_t^k - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2 - 2\alpha_t\left\langle\nabla_{\mathbf{y}}g_t(\mathbf{x}_t, \mathbf{z}_t^k), \mathbf{z}_t^k - \mathbf{y}_t^*(\mathbf{x}_t)\right\rangle + \alpha_t^2\mathbb{E}_{\zeta_t^k}\left[\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t, \mathbf{z}_t^k; \zeta_t^k)\right\|^2\right] \\
=&\left\|\mathbf{z}_t^k - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2 - 2\alpha_t\left\langle\nabla_{\mathbf{y}}g_t(\mathbf{x}_t, \mathbf{z}_t^k), \mathbf{z}_t^k - \mathbf{y}_t^*(\mathbf{x}_t)\right\rangle + \alpha_t^2\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t, \mathbf{z}_t^k)\right\|^2 + \frac{\alpha_t^2\sigma^2}{B_t} \\
\leq&\left\|\mathbf{z}_t^k - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2 - \alpha_t\mu_g\left\|\mathbf{z}_t^k - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2 - \left(\frac{\alpha_t}{L_{g,1}} - \alpha_t^2\right)\left\|\nabla_{\mathbf{y}}g_t(\mathbf{x}_t, \mathbf{z}_t^k)\right\|^2 + \frac{\alpha_t^2\sigma^2}{B_t} \\
\leq&p_t\left\|\mathbf{z}_t^k - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2 + \frac{\alpha_t^2\sigma^2}{B_t},
\end{aligned}$$

where $p_t := 1 - \alpha_t\mu_g$. Then, after telescoping over K_t iterations, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{z}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right] \\
\leq&p_t^{K_t}\mathbb{E}\left[\left\|\mathbf{z}_t - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right] + \frac{\alpha_t^2\sigma^2}{B_t}\sum_{j=0}^{K_t-1}p_t^j \\
\leq&2p_t^{K_t}\mathbb{E}\left[\left\|\mathbf{z}_t - \mathbf{y}_{t-1}^*(\mathbf{x}_{t-1})\right\|^2\right] + 2p_t^{K_t}\mathbb{E}\left[\left\|\mathbf{y}_{t-1}^*(\mathbf{x}_{t-1}) - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right] + \frac{\alpha_t\sigma^2}{\mu_g B_t} \\
\leq&2p_t^{K_t}\mathbb{E}\left[\left\|\mathbf{z}_t - \mathbf{y}_{t-1}^*(\mathbf{x}_{t-1})\right\|^2\right] + 4p_t^{K_t}\kappa_g^2\mathbb{E}\left[\left\|\mathbf{x}_{t-1} - \mathbf{x}_t\right\|^2\right] \\
&+ 4p_t^{K_t}\mathbb{E}\left[\left\|\mathbf{y}_{t-1}^*(\mathbf{x}_t) - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right] + \frac{\alpha_t\sigma^2}{\mu_g B_t}.
\end{aligned}$$

Finally, telescoping the above inequality yields

$$\begin{aligned}
&\mathbb{E}\left[\left\|\mathbf{z}_{t+1} - \mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right] \\
\leq&\rho_{\mathbf{z}}^{t-1}\mathbb{E}\left[\left\|\mathbf{z}_2 - \mathbf{y}_1^*(\mathbf{x}_1)\right\|^2\right] + \frac{\sigma^2}{\mu_g}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{z}}^j\alpha_{t-j}}{B_{t-j}} + 4p_t^{K_t}\kappa_g^2\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\mathbb{E}\left[\left\|\mathbf{x}_{t-1-j} - \mathbf{x}_{t-j}\right\|^2\right] \\
&+ 4p_t^{K_t}\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\mathbb{E}\left[\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2\right],
\end{aligned}$$

where the last inequality follows by assuming that, for all $t \in [T]$, K_t is sufficiently large such that $2p_t^{K_t} \leq \rho_{\mathbf{z}}$ for some constant $\rho_{\mathbf{z}}$. □

Lemma E.4. *Under Assumptions 2.1, 2.2 and 2.3, let* $\alpha_t \leq \frac{1}{L_{g,1}}$, $\beta_t \leq \frac{1}{2\lambda_t L_{g,1}}$ *and* $\gamma_t \equiv \gamma \leq \frac{1}{8\kappa_g L_{g,1}}\sqrt{\frac{1-\rho}{21c_\lambda}}$, *Algorithm 3 can obtain*

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\right]\\
\leq& 2c_\tau\Delta_1+\frac{54\lambda_T^2\kappa_g L_{g,1}\sigma^2}{1-\rho}\sum_{t=1}^{T}\frac{\beta_t\lambda_t+\alpha_t}{B_t}+\frac{144\kappa_g^2L_{f,0}^2c_\lambda}{1-\rho}\sum_{t=1}^{T}\frac{1}{\lambda_t^2}+\frac{168L_{g,1}^2c_\lambda}{1-\rho}H_{2,T}\\
&+\frac{672\gamma^2\kappa_g^2L_{g,1}^2c_\lambda}{1-\rho}\mathbb{E}\left[\mathrm{BLReg}(T)\right]+\frac{672\gamma^2\kappa_g^2L_{g,1}^2c_\lambda D_1^2}{1-\rho}\sum_{t=1}^{T}\frac{1}{\lambda_t^2}+\frac{2016\gamma^2\kappa_g^2L_{g,1}^2c_\lambda\sigma^2}{1-\rho}\sum_{t=1}^{T}\frac{\lambda_t^2}{\widehat{B}_t},
\end{aligned}$$

where c_τ, Δ_1 *are some constants. For each* $t\in[T]$, *we choose* $K_t>0$ *such that* $\lambda_t^2q_t^{K_t}\leq c_\lambda$, $\lambda_t^2p_t^{K_t}\leq c_\lambda$ *for some constant* c_λ.

Proof. Recall from equation (20) in Lemma C.4 that

$$\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\leq 6\lambda_t^2L_{g,1}^2\left\|\mathbf{y}_{t+1}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2+3\lambda_t^2L_{g,1}^2\left\|\mathbf{z}_{t+1}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2.$$

Then, substituting the results of Lemma E.2 and Lemma E.3 into the above inequality, we obtain

$$\begin{aligned}
&\sum_{t=2}^{T}\mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\right]\\
\leq& 6L_{g,1}^2\sum_{t=2}^{T}\lambda_t^2\mathbb{E}\left[\left\|\mathbf{y}_{t+1}-\mathbf{y}_{\lambda_t,t}^*(\mathbf{x}_t)\right\|^2\right]+3L_{g,1}^2\sum_{t=2}^{T}\lambda_t^2\mathbb{E}\left[\left\|\mathbf{z}_{t+1}-\mathbf{y}_t^*(\mathbf{x}_t)\right\|^2\right]\\
\leq& 6L_{g,1}^2\mathbb{E}\left[\left\|\mathbf{y}_2-\mathbf{y}_{\lambda_1,1}^*(\mathbf{x}_1)\right\|^2\right]\sum_{t=2}^{T}\lambda_t^2\rho_{\mathbf{y}}^{t-1}+24\kappa_gL_{g,1}\sigma^2\sum_{t=2}^{T}\lambda_t^2\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j\beta_{t-j}\lambda_{t-j}}{B_{t-j}}\\
&+72\kappa_g^2L_{f,0}^2\sum_{t=2}^{T}\lambda_t^2q_t^{K_t}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2}+72\kappa_g^2L_{g,1}^2\sum_{t=2}^{T}\lambda_t^2q_t^{K_t}\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\mathbb{E}\left[\left\|\mathbf{x}_{t-j}-\mathbf{x}_{t-1-j}\right\|^2\right]\\
&+72L_{g,1}^2\sum_{t=2}^{T}\lambda_t^2q_t^{K_t}\sum_{j=0}^{t-2}\rho_{\mathbf{y}}^j\mathbb{E}\left[\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j})-\mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2\right]\\
&+3L_{g,1}^2\mathbb{E}\left[\left\|\mathbf{z}_2-\mathbf{y}_1^*(\mathbf{x}_1)\right\|^2\right]\sum_{t=2}^{T}\lambda_t^2\rho_{\mathbf{z}}^{t-1}+3\kappa_gL_{g,1}\sigma^2\sum_{t=2}^{T}\lambda_t^2\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{z}}^j\alpha_{t-j}}{B_{t-j}}\\
&+12\kappa_g^2L_{g,1}^2\sum_{t=2}^{T}\lambda_t^2p_t^{K_t}\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\mathbb{E}\left[\left\|\mathbf{x}_{t-j}-\mathbf{x}_{t-1-j}\right\|^2\right]\\
&+12L_{g,1}^2\sum_{t=2}^{T}\lambda_t^2p_t^{K_t}\sum_{j=0}^{t-2}\rho_{\mathbf{z}}^j\mathbb{E}\left[\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j})-\mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2\right].
\end{aligned}$$

Now, assume that, for a sufficiently large K_t, $\lambda_t^2q_t^{K_t}\leq c_\lambda$ and $\lambda_t^2p_t^{K_t}\leq c_\lambda$. For simplicity, define $\rho:=\max\{\rho_{\mathbf{y}},\rho_{\mathbf{z}}\}$. Then, remind of equation (23), it holds that

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)-\nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\right]\\
\leq& c_\tau\Delta_1+27\kappa_gL_{g,1}\sigma^2\sum_{t=2}^{T}\lambda_t^2\sum_{j=0}^{t-2}\frac{\rho^j}{B_{t-j}}(\beta_{t-j}\lambda_{t-j}+\alpha_{t-j})+72\kappa_g^2L_{f,0}^2c_\lambda\sum_{t=2}^{T}\sum_{j=0}^{t-2}\frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2}
\end{aligned}$$

$$
\begin{aligned}
&+ 84L_{g,1}^2 c_\lambda \sum_{t=2}^{T}\sum_{j=0}^{t-2} \rho^j \mathbb{E}\left[\left\|\mathbf{y}_{t-1-j}^*(\mathbf{x}_{t-j}) - \mathbf{y}_{t-j}^*(\mathbf{x}_{t-j})\right\|^2\right] \\
&+ 84\kappa_g^2 L_{g,1}^2 c_\lambda \sum_{t=2}^{T}\sum_{j=0}^{t-2} \rho^j \mathbb{E}\left[\left\|\mathbf{x}_{t-j} - \mathbf{x}_{t-1-j}\right\|^2\right] \\
\leq & c_\tau \Delta_1 + 27\kappa_g L_{g,1}\sigma^2 \sum_{t=2}^{T} \lambda_t^2 \sum_{j=0}^{t-2} \frac{\rho^j}{B_{t-j}}(\beta_{t-j}\lambda_{t-j} + \alpha_{t-j}) + 72\kappa_g^2 L_{f,0}^2 c_\lambda \sum_{t=2}^{T}\sum_{j=0}^{t-2} \frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2} \\
&+ \frac{84\kappa_g^2 L_{g,1}^2 c_\lambda}{1-\rho} \sum_{t=2}^{T} \mathbb{E}\left[\left\|\mathbf{x}_t - \mathbf{x}_{t-1}\right\|^2\right] + \frac{84L_{g,1}^2 c_\lambda}{1-\rho} H_{2,T},
\end{aligned}
$$

where $\Delta_1 := 6L_{g,1}^2 \mathbb{E}[\|\mathbf{y}_2 - \mathbf{y}_{\lambda_1,1}^*(\mathbf{x}_1)\|^2] + 3L_{g,1}^2 \mathbb{E}[\|\mathbf{z}_2 - \mathbf{y}_1^*(\mathbf{x}_1)\|^2]$. Then we have

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\right] \\
\leq & c_\tau \Delta_1 + 27\kappa_g L_{g,1}\sigma^2 \sum_{t=2}^{T} \lambda_t^2 \sum_{j=0}^{t-2} \frac{\rho^j}{B_{t-j}}(\beta_{t-j}\lambda_{t-j} + \alpha_{t-j}) + 72\kappa_g^2 L_{f,0}^2 c_\lambda \sum_{t=2}^{T}\sum_{j=0}^{t-2} \frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2} \\
&+ \frac{336\kappa_g^2 L_{g,1}^2 c_\lambda}{1-\rho} \sum_{t=1}^{T} \gamma_t^2 \mathbb{E}\Big[\left\|\mathcal{G}_{\mathcal{X}}\left(\mathbf{x}_t, \nabla F_t(\mathbf{x}_t), \gamma_t\right)\right\|^2 + \left\|\nabla F_t(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 \\
&+ \left\|\nabla\mathcal{L}_t^*(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2 + \left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t; \mathcal{S}_t)\right\|^2 \Big] + \frac{84L_{g,1}^2 c_\lambda}{1-\rho} H_{2,T}.
\end{aligned}
$$

Let $\gamma_t \equiv \gamma \leq \frac{1}{8\kappa_g L_{g,1}} \sqrt{\frac{1-\rho}{21 c_\lambda}}$, we can gaurantee that $1 - \frac{336\kappa_g^2 L_{g,1}^2 c_\lambda}{1-\rho} \geq \frac{1}{2}$, thus it holds that

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\right] \\
\leq & 2c_\tau \Delta_1 + 54\kappa_g L_{g,1}\sigma^2 \sum_{t=2}^{T} \lambda_t^2 \sum_{j=0}^{t-2} \frac{\rho^j}{B_{t-j}}(\beta_{t-j}\lambda_{t-j} + \alpha_{t-j}) + 144\kappa_g^2 L_{f,0}^2 c_\lambda \sum_{t=2}^{T}\sum_{j=0}^{t-2} \frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2} \\
&+ \frac{672\gamma^2\kappa_g^2 L_{g,1}^2 c_\lambda}{1-\rho} \mathbb{E}\left[\mathrm{BLReg}(T)\right] + \frac{672\gamma^2\kappa_g^2 L_{g,1}^2 c_\lambda D_1^2}{C_1(1-\rho)} \sum_{t=1}^{T} \frac{1}{\lambda_t^2} + \frac{2016\gamma^2\kappa_g^2 L_{g,1}^2 c_\lambda \sigma^2}{1-\rho} \sum_{t=1}^{T} \frac{\lambda_t^2}{\widehat{B}_t} \\
&+ \frac{168L_{g,1}^2 c_\lambda}{1-\rho} H_{2,T}. \qquad (38)
\end{aligned}
$$

Consider the second term. Letting $m = t - j$, we have

$$
\sum_{t=2}^{T} \lambda_t^2 \sum_{j=0}^{t-2} \frac{\rho^j}{B_{t-j}}(\beta_{t-j}\lambda_{t-j} + \alpha_{t-j}) = \sum_{m=2}^{T} \frac{\beta_m\lambda_m + \alpha_m}{B_m} \sum_{t=m}^{T} \lambda_t^2 \rho^{t-m} \leq \frac{\lambda_T^2}{1-\rho} \sum_{t=1}^{T} \frac{\beta_t\lambda_t + \alpha_t}{B_t}.
$$

For the third term, we similarly have

$$
\sum_{t=2}^{T}\sum_{j=0}^{t-2} \frac{\rho_{\mathbf{y}}^j}{\lambda_{t-j}^2} = \sum_{m=2}^{T} \frac{1}{\lambda_m^2} \sum_{t=m}^{T} \rho_{\mathbf{y}}^{t-m} \leq \frac{1}{1-\rho} \sum_{t=1}^{T} \frac{1}{\lambda_t^2}.
$$

Substituting the above two inequalities into equation (38), we complete the proof by

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\left\|\widetilde{\nabla}\mathcal{L}_t^*(\mathbf{x}_t) - \nabla\mathcal{L}_t^*(\mathbf{x}_t)\right\|^2\right] \\
\leq & 2c_\tau \Delta_1 + \frac{54\lambda_T^2\kappa_g L_{g,1}\sigma^2}{1-\rho} \sum_{t=1}^{T} \frac{\beta_t\lambda_t + \alpha_t}{B_t} + \frac{144\kappa_g^2 L_{f,0}^2 c_\lambda}{1-\rho} \sum_{t=1}^{T} \frac{1}{\lambda_t^2} + \frac{168L_{g,1}^2 c_\lambda}{1-\rho} H_{2,T}
\end{aligned}
$$

$$+\frac{672\gamma^2\kappa_g^2L_{g,1}^2c_\lambda}{1-\rho}\mathbb{E}\left[\mathrm{BLReg}(T)\right]+\frac{672\gamma^2\kappa_g^2L_{g,1}^2c_\lambda D_1^2}{1-\rho}\sum_{t=1}^{T}\frac{1}{\lambda_t^2}+\frac{2016\gamma^2\kappa_g^2L_{g,1}^2c_\lambda\sigma^2}{1-\rho}\sum_{t=1}^{T}\frac{\lambda_t^2}{\widehat{B}_t}.$$

□

Theorem E.5 (Restatement of Theorem 5.1)**.** *Under Assumption 2.1, 2.2 and 2.3, Let* $\Lambda(t,\lambda_t)=\left(1+\frac{1}{t}\right)^\tau\lambda_t$ *for some* $\tau\in(0,\frac{1}{2})$, $\lambda_1>\frac{2L_{f,1}}{\mu_g}$, $\alpha_t=\frac{1}{L_{g,1}t^a}$, $\beta_t=\frac{1}{2\lambda_1L_{g,1}t^b}$, $\gamma_t\equiv\gamma\le\min\{\frac{1}{2L_F},\frac{1}{48\kappa_gL_{g,1}}\sqrt{\frac{1-\rho}{7c_\lambda}}\}$, $K_t=4\kappa_g\max\{t^{b-\tau},t^a\}\log t$, $\widehat{B}_t=t^{B_1}$ *and* $B_t=t^{B_2}$, *Algorithm 3 can guarantee*

$$\mathbb{E}\left[\mathrm{BLReg}(T)\right]\le O\left(T^{1-2\tau}+\sigma^2\max\{T^{1+2\tau-B_1},T^{1+3\tau-b-B_2},T^{1+2\tau-a-B_2}\}+V_T+H_{2,T}\right),$$

where ρ, c_λ *are some constants.*

Proof. Substituting Lemma E.4 into Lemma E.1, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{BLReg}(T)\right]\\
\le&\sum_{t=1}^{T}\frac{8}{\gamma_t}\left(F_t(\mathbf{x}_t)-F_t(\mathbf{x}_{t+1})\right)+24D_1^2\sum_{t=1}^{T}\frac{1}{\lambda_t^2}+72\sigma^2\sum_{t=1}^{T}\frac{\lambda_t^2}{\widehat{B}_t}+\frac{3456\kappa_g^2L_{f,0}^2c_\lambda}{1-\rho}\sum_{t=1}^{T}\frac{1}{\lambda_t^2}\\
&+48c_\tau\Delta_1+\frac{1296\lambda_T^2\kappa_gL_{g,1}\sigma^2}{1-\rho}\sum_{t=1}^{T}\frac{\beta_t\lambda_t+\alpha_t}{B_t}+\frac{16128\gamma^2\kappa_g^2L_{g,1}^2c_\lambda}{1-\rho}\mathbb{E}\left[\mathrm{BLReg}(T)\right]\\
&+\frac{16128\gamma^2\kappa_g^2L_{g,1}^2c_\lambda D_1^2}{1-\rho}\sum_{t=1}^{T}\frac{1}{\lambda_t^2}+\frac{48384\gamma^2\kappa_g^2L_{g,1}^2c_\lambda\sigma^2}{1-\rho}\sum_{t=1}^{T}\frac{\lambda_t^2}{\widehat{B}_t}+\frac{4032L_{g,1}^2c_\lambda}{1-\rho}H_{2,T}.
\end{aligned}$$

Choosing $\gamma\le\frac{1}{48\kappa_gL_{g,1}}\sqrt{\frac{1-\rho}{7c_\lambda}}$ ensures that $1-\frac{4032\gamma^2\kappa_g^2L_{g,1}^2c_\lambda}{1-\rho}\ge\frac{1}{2}$. Moreover, with

$$\lambda_t=\lambda_1t^\tau,\quad\tau\in\left(0,\frac{1}{2}\right),\quad\beta_t=\frac{1}{2\lambda_1L_{g,1}t^b},\quad\alpha_t=\frac{1}{L_{g,1}t^a},\quad\widehat{B}_t=t^{B_1},\quad B_t=t^{B_2},$$

we finally have

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{BLReg}(T)\right]\\
\le&\frac{32M}{\gamma}+\frac{16}{\gamma}V_T+\frac{48D_1^2}{\lambda_1^2}\sum_{t=1}^{T}t^{-2\tau}+144\lambda_1^2\sigma^2\sum_{t=1}^{T}t^{2\tau-B_1}+\frac{6912\kappa_g^2L_{f,0}^2c_\lambda}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}t^{-2\tau}\\
&+\frac{2592\lambda_1^2\kappa_g\sigma^2T^{2\tau}}{1-\rho}\sum_{t=1}^{T}(t^{\tau-b-B_2}+t^{-a-B_2})+\frac{32256\gamma^2\kappa_g^2L_{g,1}^2c_\lambda D_1^2}{\lambda_1^2(1-\rho)}\sum_{t=1}^{T}t^{-2\tau}\\
&+96c_\tau\Delta_1+\frac{96768\gamma^2\lambda_1^2\kappa_g^2L_{g,1}^2c_\lambda\sigma^2}{1-\rho}\sum_{t=1}^{T}t^{2\tau-B_1}+\frac{8064L_{g,1}^2c_\lambda}{1-\rho}H_{2,T}\\
=&O\left(T^{1-2\tau}+\sigma^2\max\{T^{1+2\tau-B_1},T^{1+3\tau-b-B_2},T^{1+2\tau-a-B_2}\}+V_T+H_{2,T}\right),
\end{aligned}$$

which completes the proof. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract, we state that the focus of this paper is the development and theoretical analysis of fully first-order algorithms for OBO problems. In the Introduction, we further clarify our main theoretical contributions in the "Our Contributions" section and present them clearly in tables.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss our limitations in Conclusion and Discussion.

   Guidelines:

   - The answer [N/A] 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.

3. **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: All main theoretical statements are accompanied by formal proofs in the appendix. The assumptions used in the analysis are stated in the preliminaries and are also referenced in the corresponding theorem and lemma statements.

Guidelines:

- The answer [N/A] 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.

4. **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: The paper describes the experimental setup needed to reproduce the reported results, including the datasets, evaluation protocols, baselines, and key hyperparameters. Additional implementation details and scripts are provided in the supplementary material.

Guidelines:

- The answer [N/A] 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.

5. **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 supplementary material contains the code used for the experiments, together with the configurations needed to run the proposed methods and baselines. The implementation specifies the parameter choices and experimental procedures used to generate the main results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.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://neurips.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.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental section specify the datasets, evaluation metrics and optimization settings in our experiments. Further details are included in the submitted code.

   Guidelines:

   - The answer [N/A] 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.

7. **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: The main empirical results are reported over multiple independent runs. We include error bars based on standard deviations to reflect the variability across different random seeds.

   Guidelines:

   - The answer [N/A] 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.

8. **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: [No]

   Justification: The paper does not provide a complete description of the hardware configuration and memory requirements for every experiment.

   Guidelines:

   - The answer [N/A] 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).

9. **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 work is conducted in accordance with the NeurIPS Code of Ethics. It does not involve human-subject studies, private or sensitive data, or deployment scenarios that introduce special ethical concerns.

   Guidelines:

   - The answer [N/A] 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).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: The paper is mainly concerned with theoretical and algorithmic developments for online bilevel optimization.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] 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).

11. **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., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The submission does not release models, datasets, or systems that pose substantial misuse or dual-use risks.

Guidelines:

- The answer [N/A] 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.

12. **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: Existing datasets and software tools used in the experiments are cited where appropriate. We follow their intended usage and respect the corresponding licenses and terms.

Guidelines:

- The answer [N/A] 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.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset associated with this submission is the experimental code included in the supplementary material.

    Guidelines:

    - The answer [N/A] 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.

14. **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: [N/A]

    Justification: The paper does not include crowdsourcing experiments, user studies, or any other form of research involving human participants.

    Guidelines:

    - The answer [N/A] 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.

15. **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: [N/A]

Justification: No human-subject research is conducted in this work, so IRB approval or an equivalent ethics review is not required.

Guidelines:

- The answer [N/A] 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.

16. **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 rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the proposed methods, experiments, or theoretical analysis. Any use of language tools for editing or formatting does not affect the methodology or scientific claims.

Guidelines:

- The answer [N/A] 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 in the NeurIPS handbook for what should or should not be described.