Online Nonconvex Bilevel Optimization with Bregman Divergences

Jason Bohne[1,2] JASON.BOHNE@STONYBROOK.EDU
David Rosenberg[2] DROSENBERG44@BLOOMBERG.NET
Gary Kazantsev[2] GKAZANTSEV@BLOOMBERG.NET
Paweł Polak[1,3,4] PAWEL.POLAK@STONYBROOK.EDU
[1] *Stony Brook University,* [2] *Bloomberg,* [3] *Institute for Advanced Computational Science,* [4] *Center of Excellence in Wireless and Information Technology*

Abstract

Bilevel optimization methods are increasingly relevant within machine learning, especially for tasks such as hyperparameter optimization and meta-learning. Compared to the offline setting, online bilevel optimization (OBO) offers a more dynamic framework by accommodating time-varying functions and sequentially arriving data. This study addresses the online nonconvex-strongly convex bilevel optimization problem. In deterministic settings, we introduce a novel online Bregman bilevel optimizer (OBBO) that utilizes adaptive Bregman divergences. We demonstrate that OBBO enhances the known sublinear rates for bilevel local regret through a novel hypergradient error decomposition that adapts to the underlying geometry of the problem. In stochastic contexts, we introduce the first stochastic online bilevel optimizer (SOBBO), which employs a window averaging method for updating outer-level variables using a weighted average of recent stochastic approximations of hypergradients. This approach not only achieves sublinear rates of bilevel local regret but also serves as an effective variance reduction strategy, obviating the need for additional stochastic gradient samples at each timestep. Experiments on online hyperparameter optimization and online meta-learning highlight the superior performance, efficiency, and adaptability of our Bregman-based algorithms compared to established online and offline bilevel benchmarks.

1. Introduction

Online bilevel optimization (OBO) is a recently introduced strategy that complements traditional offline bilevel optimization approaches. It is tailored for dynamic environments encountered in machine learning tasks such as online hyperparameter optimization [24], online meta-learning [34], domain adaptation [13], and dataset augmentation [29]. The deterministic OBO problem, for parameters $\boldsymbol{\lambda} \in \mathcal{X} \subseteq \mathbb{R}^{d_1}$, $\boldsymbol{\beta} \in \mathbb{R}^{d_2}$, and the online running index $\forall t \in [1, T]$, is defined as:

$$\widehat{\boldsymbol{\lambda}}_t \in \underset{\boldsymbol{\lambda} \in \mathcal{X} \subseteq \mathbb{R}^{d_1}}{\arg\min} \left\{ f_t\left(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})\right) + h(\boldsymbol{\lambda}) \right\}, \quad \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}) \in \underset{\boldsymbol{\beta} \in \mathbb{R}^{d_2}}{\arg\min}\, g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}), \tag{1}$$

where $F_t(\boldsymbol{\lambda}) \triangleq f_t\left(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})\right)$ is a nonconvex and smooth outer objective function, $h(\boldsymbol{\lambda})$ is a convex and potentially nonsmooth regularization term, and $g_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ is a smooth inner objective function that is μ_g-strongly convex in $\boldsymbol{\beta}$. Single-level online optimization problems are often formulated with time-varying objectives that are only accessible through a stochastic oracle, for example, a stochastic gradient oracle in the online learning of an LSTM in [1] and in the online learning of a GAN in [17]. Similarly, many machine learning tasks in offline bilevel optimization are structured using stochastic loss functions to capture the inherent randomness and perturbations from using sample batches. Examples include the meta-learning task of [21] and the data hyper-cleaning task as in [20]. This motivates us to introduce the first stochastic online bilevel optimization problem:

$$\widehat{\boldsymbol{\lambda}}_t \in \underset{\boldsymbol{\lambda}\in\mathcal{X}\subseteq\mathbb{R}^{d_1}}{\arg\min} \left\{\mathbb{E}_\epsilon \left[f_t(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}), \epsilon)\right] + h(\boldsymbol{\lambda})\right\}, \qquad \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}) \in \underset{\boldsymbol{\beta}\in\mathbb{R}^{d_2}}{\arg\min}\, \mathbb{E}_\zeta\left[g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)\right] \tag{2}$$

where $F_t(\boldsymbol{\lambda}) \triangleq \mathbb{E}_\epsilon \left[f_t(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}), \epsilon)\right]$ is a nonconvex and smooth outer objective function, $h(\boldsymbol{\lambda})$ is a convex and potentially nonsmooth regularization term, and $\mathbb{E}_\zeta\left[g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)\right]$ is a smooth inner objective function that is μ_g-strongly convex in $\boldsymbol{\beta}$. Further, ϵ and ζ are independent and identically distributed random variables from error distributions D and D', respectively.

Current methods in OBO of [24, 34] are limited as: (i) they assume smooth outer level objectives, excluding OBO problems where the outer level objective is separable into a smooth and nonsmooth term; (ii) they use Euclidean proximal operators, preventing the application of more sophisticated gradient step techniques of Bregman divergences as explored in the offline bilevel optimization of [20]; and (iii) they restrict the theoretical discussion to a deterministic setting, whereas stochasticity is commonly assumed in both online optimization [16] and offline bilevel optimization [21].

Addressing these limitations, our contributions to OBO include introducing a novel family of online Bregman bilevel optimizers (OBBO) that utilize Bregman divergences to solve a larger class of OBO problems. These problems include cases where the outer-level objective is a nonconvex composite function separable into a smooth and potentially nonsmooth term, and the inner-level objective is strongly convex. Due to the application of Bregman divergences to our novel hypergradient error decomposition, we show that OBBO improves the previously best known sublinear rate of bilevel local regret. Further, our hypergradient error decomposition can be a theoretical interest of its own, as we provide a decomposition not expressed in previous works ([24, 34]) which depend on the variation of outer level variables. Additionally, we formulate the first stochastic OBO problem and provide an algorithm (SOBBO) that achieves a sublinear rate of bilevel local regret. We demonstrate the computational efficiency of SOBBO compared to offline stochastic bilevel algorithms, as it incorporates a variance reduction technique without requiring additional gradient samples. In Appendix E, empirical results on online hyperparameter optimization and online meta-learning validate our theoretical contributions and practicality of our improved methods.

2. Preliminaries

Notations and Assumptions Let $\|\cdot\|$ denote the ℓ_2 norm for vectors and the spectral norm for matrices, with $\langle\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\rangle$ denoting the inner product between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. For a function $g_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ and its stochastic variation $g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)$ we denote the gradient as $\nabla g_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ and $\nabla g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)$ respectively, with partial derivatives denoted, for example with respect to $\boldsymbol{\lambda}$, as $\nabla_{\boldsymbol{\lambda}} g_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ and $\nabla_{\boldsymbol{\lambda}} g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)$ respectively. For the deterministic OBO problem, we make the following assumptions that are standard in OBO [24, 34], whereas, in the stochastic setting, our assumptions are typical in stochastic bilevel optimization of [11, 20, 21].

Assumption A *For all $t \in [1, \dots, T]$ and $\forall\boldsymbol{\lambda} \in \mathcal{X}$, $\forall\boldsymbol{\beta} \in \mathbb{R}^{d_2}$, we have the following:*
A1. $f_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ and $\nabla f_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ are respectively $\ell_{f,0}$-Lipschitz and $\ell_{f,1}$-Lipschitz continuous.
A2. $\nabla g_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ is $\ell_{g,1}$-Lipschitz continuous.
A3. $\nabla^2_{\boldsymbol{\lambda},\boldsymbol{\beta}} g_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ and $\nabla^2_{\boldsymbol{\beta},\boldsymbol{\beta}} g_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ are both $\ell_{g,2}$-Lipschitz continuous.
In the stochastic case A1.-A3. hold for $f_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \epsilon)$ and $g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)$ for any $\epsilon \sim D$ and $\zeta \sim D'$.

Assumption B *$g_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ and its stochastic variation $g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)$ are μ_g-strongly convex functions in $\boldsymbol{\beta}$ $\forall t \in [1, T]$ and for all $\boldsymbol{\lambda} \in \mathcal{X}$.*

Assumption C *For all $t \in [1, T]$ and any $\boldsymbol{\lambda} \in \mathcal{X}$ and $\boldsymbol{\beta} \in \mathbb{R}^{d_2}$ the stochastic partial derivative of $\nabla_{\boldsymbol{\beta}} g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)$ satisfies $\mathbb{E}[\nabla_{\boldsymbol{\beta}} g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)] = \nabla_{\boldsymbol{\beta}} g_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ with finite variance $\sigma^2_{g_{\boldsymbol{\beta}}}$. Further the estimated stochastic partial derivative of $\widetilde{\nabla} f_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathcal{E})$ has finite variance σ^2_f.*

Assumption D *Set $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ is closed, convex, and bounded s.t. $\forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathcal{X}$, $\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| \leq S$.*

Assumption E *For all $t \in [1, \ldots, T]$ it holds for finite $Q \in \mathbb{R}$ that $\sup_{\boldsymbol{\lambda} \in \mathcal{X}} |F_t(\boldsymbol{\lambda})| \leq Q$.*

Assumption F *For all $t \in [1, T]$ the distance generating function $\phi_t(\boldsymbol{\lambda})$ is continuously differentiable and ρ-strongly convex with respect to $\|\cdot\|$.*

Bregman Proximal Gradient Introduced in [4], a Bregman divergence $\mathcal{D}_\phi(\boldsymbol{\lambda}_2, \boldsymbol{\lambda}_1) := \phi(\boldsymbol{\lambda}_2) - \phi(\boldsymbol{\lambda}_1) - \langle \nabla \phi(\boldsymbol{\lambda}_1), \boldsymbol{\lambda}_2 - \boldsymbol{\lambda}_1 \rangle$ for a continuously differentiable and ρ-strongly convex function $\phi(\boldsymbol{\lambda})$ offers a generalization to the squared Euclidean distance. Given a Bregman divergence $\mathcal{D}_\phi(\cdot, \cdot)$, our proximal gradient step is

$$\boldsymbol{\lambda}^+ = \underset{\boldsymbol{\lambda} \in \mathcal{X}}{\arg\min} \left\{ \langle \boldsymbol{q}, \boldsymbol{\lambda} \rangle + h(\boldsymbol{\lambda}) + \frac{1}{\alpha} \mathcal{D}_\phi(\boldsymbol{\lambda}, \boldsymbol{u}) \right\}, \tag{3}$$

where $\phi(\boldsymbol{\lambda})$ is a continuously differentiable and ρ-strongly convex function, $h(\boldsymbol{\lambda})$ is a convex and potentially nonsmooth regularization term, $\alpha > 0$ is a step size, and $\boldsymbol{q}, \boldsymbol{u} \in \mathbb{R}^{d_1}$ are the estimate of the gradient, and current reference point, respectively. Proximal gradient methods in offline bilevel optimization have been shown to improve convergence rates as in the Bio-BreD algorithm of [20]. Special cases of the gradient update in (3) include projected gradient descent ($\phi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|^2$, $\mathcal{X} \subseteq \mathbb{R}^{d_1}$, and $h(\boldsymbol{\lambda}) = 0$), as well as proximal gradient descent ($\phi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|^2$ and $\mathcal{X} = \mathbb{R}^{d_1}$). The aforementioned gradient step in (3) can be further extended to a time-varying distance generating function, e.g., $\phi_t(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T \mathbf{H}_t \boldsymbol{\lambda}$ with an adaptive matrix $\mathbf{H}_t$, resulting in an adaptive proximal gradient method with similarities to Adagrad from [7] and Super-Adam of [19]. The proximal gradient step of (3) has led to the introduction of a generalized projection from [12] defined for a step size $\alpha > 0$, $\boldsymbol{q} \in \mathbb{R}^{d_1}$, and $\boldsymbol{u} \in \mathcal{X}$ as $\mathcal{G}_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{q}, \alpha) := \frac{1}{\alpha}\left(\boldsymbol{u} - \boldsymbol{\lambda}^+\right)$. Here $\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}, \nabla f_t(\boldsymbol{\lambda}), \alpha)$ acts as a generalized gradient that simplifies to $\nabla f_t(\boldsymbol{\lambda})$ if $\mathcal{X} = \mathbb{R}^{d_1}$ and $h(\boldsymbol{\lambda}) = 0$.

Bilevel Local Regret Bilevel local regret is a stationary metric for online bilevel optimization [24, 34] that extends the single-level local regret measure from [17]. The work of [24] in particular defines the bilevel local regret for a window length $w \geq 1$ and a sequence $\{\boldsymbol{\lambda}_t\}_{t=1}^T$ as $BLR_w(T) := \sum_{t=1}^T \|\nabla F_{t,w}(\boldsymbol{\lambda}_t)\|^2$ where for simplicity we have defined $F_{t,w}(\boldsymbol{\lambda}_t) := \frac{1}{w}\sum_{i=0}^{w-1} F_{t-i}(\boldsymbol{\lambda}_{t-i})$ as a time-smoothed outer level objective with $F_t = 0$ $\forall t \leq 0$. To incorporate the generalized projection, we introduce a new bilevel local regret for a window length $w \geq 1$ and sequence $\{\boldsymbol{\lambda}_t\}_{t=1}^T$ as

$$BLR_w(T) := \sum_{t=1}^T \|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\|^2 \tag{4}$$

In the setting where $\mathcal{X} = \mathbb{R}^{d_1}$, $h(\boldsymbol{\lambda}) = 0$, and $\phi_t(\boldsymbol{\lambda}) = \phi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|^2$, our variation of local regret simplifies to the regret measure of [24]. However, our definition offers an important generalization of bilevel local regret when an adaptive distance generating function $\phi_t(\boldsymbol{\lambda})$ or a non-zero regularization term $h(\boldsymbol{\lambda})$ is present. Proposition 1 in [2] shows that in nonstationary environments, there always exists a sequence of well-behaved loss functions for which sublinear regret cannot be achieved. Hence, to derive useful regret bounds of online algorithms, further regularity constraints must be imposed on the sequence, such as a sublinear path variation [35], function variation [2], or gradient variation [6]. In nonstationary OBO one proposed regularity constraint is the p-th order inner level path variation of optimal decisions from [34], and is $H_{p,T} := \sum_{t=2}^{T} \sup_{\boldsymbol{\lambda}\in\mathcal{X}} \left\| \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}) \right\|^p$. A regularity metric on the p-th order variation of the evaluations of the outer level objective function across time is suggested by [24] and is $V_{p,T} := \sum_{t=1}^{T} \sup_{\boldsymbol{\lambda}\in\mathcal{X}} \left| F_{t+1}(\boldsymbol{\lambda}) - F_t(\boldsymbol{\lambda}) \right|^p$.

3. Online Bregman Bilevel Optimizers

Deterministic Algorithm (OBBO) We begin our section on the deterministic online Bregman bilevel optimizer OBBO with a useful lemma on the gradient of the outer level objective $\nabla F_t(\boldsymbol{\lambda})$, often referred as the hypergradient [11]. Here we expand the hypergradient with the chain rule followed by the implicit function theorem.

Lemma 1 *(Lemma 2.1 in [11]) Under Assumption A and B, we have* $\forall \boldsymbol{\lambda} \in \mathcal{X}$, $\forall t \in [1, T]$

$$\nabla F_t(\boldsymbol{\lambda}) = \nabla_{\boldsymbol{\lambda}} f_t(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})) - \nabla^2_{\boldsymbol{\lambda},\boldsymbol{\beta}} g_t(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})) \left(\nabla^2_{\boldsymbol{\beta},\boldsymbol{\beta}} g_t(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}))\right)^{-1} \nabla_{\boldsymbol{\beta}} f(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})). \quad (5)$$

Note the computation of $\nabla F_t(\boldsymbol{\lambda})$ requires knowledge of the optimal inner level variables $\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})$, which are typically inaccessible. This has motivated the construction of computationally efficient gradient estimators for $\nabla F_t(\boldsymbol{\lambda})$ that can provide an approximation to $\nabla F_t(\boldsymbol{\lambda})$ without access to $\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})$. One popular approach is to utilize iterative differentiation techniques for gradient estimation, that is define the estimate $\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K) := \frac{\partial f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}$, see [21, 26]. Hence, we examine our OBBO algorithm in this case, although we remark other hypergradient estimates can be used.

Algorithm 1 OBBO Deterministic Online Bregman Bilevel Optimizer

```
$T, K$; stepsizes $\alpha, \eta > 0$; distance generating functions $\phi_t(\boldsymbol{\lambda}) : \mathcal{X} \mapsto \mathbb{R}$; window $w \geq 1$.
Initialize $\boldsymbol{\beta}_1 \in \mathbb{R}^{d_2}$ and $\boldsymbol{\lambda}_1 \in \mathcal{X}$
for $t = 1, \ldots, T$ do
    Retrieve information about $f_t$ and $g_t$
    $\boldsymbol{\omega}_t^0 \leftarrow \boldsymbol{\beta}_t$
    for $k = 1, \ldots, K$ do
        $\boldsymbol{\omega}_t^k \leftarrow \boldsymbol{\omega}_t^{k-1} - \eta \nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1})$
    end
    Get $\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K) := \frac{\partial f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}$ from (18) and store in memory
    Compute $\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1})$ from (6) for $\boldsymbol{\beta}_{t+1} = \boldsymbol{\omega}_t^K$
    $\boldsymbol{\lambda}_{t+1} \leftarrow \arg\min_{\boldsymbol{\lambda}\in\mathcal{X}} \left\{ \left\langle \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}), \boldsymbol{\lambda} \right\rangle + h(\boldsymbol{\lambda}) + \frac{1}{\alpha} \mathcal{D}_{\phi_t}(\boldsymbol{\lambda}, \boldsymbol{\lambda}_t) \right\}$
end
return $\boldsymbol{\lambda}_{T+1}, \boldsymbol{\beta}_{T+1}$
```

In Theorem 2.7 of [17], time smoothing is shown to be necessary in the gradient step of an online algorithm to achieve a sublinear rate of local regret. Following such, we introduce time-smoothing into our OBBO algorithm by defining and utilizing the gradient estimator for $w \geq 1$ of

$$\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta_{t+1}}) := \frac{1}{w} \sum_{i=0}^{w-1} \widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\beta}_{t+1-i}), \quad f_t = 0 \, \forall t \leq 0. \tag{6}$$

Stochastic Algorithm (SOBBO) We choose to analyze our stochastic online Bregman bilevel optimizer SOBBO with a stochastic gradient estimator common in offline bilevel optimization, see [11] and [20]. In particular, the stochastic gradient of $\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{E}_t)$ provides an estimate of $\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1})$ and is constructed $\forall t \in [1, T]$ via Lemma 13. As in (6), we introduce time-smoothing with the estimator $\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w})$ defined for all $t \in [1, T]$, window size $w \geq 1$, and independent samples $\mathcal{Z}_{t,w} = \{\mathcal{E}_{t-i}\}_{i=0}^{w-1}$ as

$$\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) := \frac{1}{w} \sum_{i=0}^{w-1} \widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\beta}_{t+1-i}, \mathcal{E}_{t-i}), \quad f_t = 0 \, \forall t \leq 0 \tag{7}$$

```
**Algorithm 2** SOBBO Stochastic Online Bregman Bilevel Optimizer
$T, K, m, s; \alpha, \eta > 0$; distance generating function $\phi_t(\boldsymbol{\lambda}) : \mathcal{X} \mapsto \mathbb{R}$; window $w \geq 1$.
Initialize $\boldsymbol{\beta}_1 \in \mathbb{R}^{d_2}$ and $\boldsymbol{\lambda}_1 \in \mathcal{X}$
for $t = 1, \ldots, T$ do
    Retrieve information about $f_t$ and $g_t$
    $\boldsymbol{\omega}_t^0 \leftarrow \boldsymbol{\beta}_t$
    for $k = 1, \ldots, K$ do
        Draw $s$ independent samples of $\zeta$, Set $\bar{\zeta}_{t,k} := \left\{\zeta_{t,i}^{k-1}\right\}_{i=1}^{s}$
        $\boldsymbol{\omega}_t^k \leftarrow \boldsymbol{\omega}_t^{k-1} - \eta \nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1}, \bar{\zeta}_{t,k})$
    end
    $\boldsymbol{\beta}_{t+1} \leftarrow \boldsymbol{\omega}_t^K$
    Draw $m$ independent samples of $\zeta$, Set $\mathcal{E}_t = \left\{\epsilon_t, \zeta_t^0, \ldots, \zeta_t^{m-1}\right\}$
    Get $\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{E}_t)$ from (20) and store in memory
    Compute $\widetilde{\nabla} f_{t,\boldsymbol{w}}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w})$ from (7)
    $\boldsymbol{\lambda}_{t+1} \leftarrow \arg\min_{\boldsymbol{\lambda} \in \mathcal{X}} \left\{ \left\langle \widetilde{\nabla} f_{t,\boldsymbol{w}}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \boldsymbol{\lambda} \right\rangle + h(\boldsymbol{\lambda}) + \frac{1}{\alpha} \mathcal{D}_{\phi_t}(\boldsymbol{\lambda}, \boldsymbol{\lambda}_t) \right\}$
end
return $\boldsymbol{\lambda}_{T+1}, \boldsymbol{\beta}_{T+1}$
```

4. Bilevel Local Regret Minimization

OBBO Regret Our objective is to show the deterministic OBBO algorithm of Algorithm 1 achieves a sublinear rate of bilevel local regret. We first rely on a novel decomposition of the hypergradient error incurred by OBBO at time t presented in Lemma 17. The hypergradient error decomposition of $\left\| \frac{\partial f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_t(\boldsymbol{\lambda}_t) \right\|^2$ introduced with OBBO is novel, as the upper bound is expressed in terms of the previous bilevel local regret and time-smoothed hypergradient error up to time $t-1$, demonstrating how errors propagate over time. In contrast, the expansion in SOBOW, see Theorem 5.6 of [24], is limited to the cumulative difference of outer-level variables. This result, from Bregman divergences, is essential for the improved sublinear rate (see Table 1) discussed next.

Theorem 2 *Suppose Assumptions A, B, D, E, and F. Let inner step size of* $\eta < \min\left(\frac{1}{\ell_{g,1}}, \frac{1}{\mu_g}\right)$, *outer step size of* $\alpha \le \min\{\frac{3\rho}{4\ell_{F,1}}, \frac{\rho\sqrt{(1-\nu)}}{\kappa_g\sqrt{108C_{\mu_g}L_{\boldsymbol{\beta}}}}\}$*, and inner iteration count* $K = \frac{\log(T)}{\log((1-\eta\mu_g)^{-1})} + 1$. *Then the bilevel local regret of our OBBO algorithm satisfies*

$$BLR_w(T) := \sum_{t=1}^{T} \|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\|^2 \le O\left(\frac{T}{w} + V_{1,T} + \kappa_g^2 H_{2,T}\right) \tag{8}$$

Analogously to OBO ([24]), we are interested in sublinear comparator sequences, e.g., $V_{1,T} = o(T)$ and $H_{2,T} = o(T)$. It is a weak assumption that still allows for the amount of nonstationarity to grow up to a rate of time itself. Using the above assumption with a properly selected sublinear window size $w = o(T)$ results in the sublinear rate of bilevel local regret presented in Theorem 2.

SOBBO Regret To study the bilevel local regret in the stochastic framework, we rely on a novel decomposition of the expected hypergradient error of SOBBO algorithm in Lemma 22. We show that the expected hypergradient error $\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\left\|\tilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}) - \nabla F_t(\boldsymbol{\lambda}_t)\right\|^2\right]$ at time t is upper bounded by terms from the deterministic setting and a variance term $\sigma_{g_{\boldsymbol{\beta}}}^2$, motivating use of variance reduction techniques. Using this, the next theorem shows a sublinear rate of bilevel local regret.

Theorem 3 *Suppose Assumptions A, B, C, D, E, and F. Let the inner step size of* $\eta \le \frac{2}{\ell_{g,1}+\mu_g}$, *inner iteration count of* $K \ge 1$, *outer step size of* $\alpha \le \min\{\frac{3\rho}{4\ell_{F,1}}, \frac{\rho\sqrt{(1-\nu)}}{\kappa_g^2\sqrt{72C_{\mu_g}}}\}$*, and batch sizes of* $s = w$ *and* $m = \log(w)/\log\left(1 - \frac{\mu_g}{\ell_{g,1}}\right) + 1$. *Then the bilevel local regret of SOBBO satisfies*

$$BLR_w(T) \le O\left(\frac{T}{w}\left(1 + \kappa_g^2 + \sigma_f^2 + \kappa_g^2\sigma_{g_{\boldsymbol{\beta}}}^2\right) + V_{1,T} + \kappa_g^2 H_{2,T}\right) \tag{9}$$

As in the deterministic setting, we consider sublinear $V_{1,T} = o(T)$ and $H_{2,T} = o(T)$. For a properly chosen window and batch size of $w = o(T)$ and $s = o(T)$, (9) results in a sublinear rate of bilevel local regret. Note our deterministic rate is a special case of (9) when $\sigma_f^2 = 0, \sigma_{g_{\boldsymbol{\beta}}}^2 = 0$.

Algorithm	$BLR_w(T)$
OAGD	$O(T/w + H_{1,T} + \boldsymbol{\kappa_g^4} H_{2,T})$
SOBOW	$O(T/w + V_{1,T} + \boldsymbol{\kappa_g^3} H_{2,T})$
OBBO	$O(T/w + V_{1,T} + \boldsymbol{\kappa_g^2} H_{2,T})$
SOBBO	$O\left(T/w\left(1 + \kappa_{\mathbf{g}}^{\mathbf{2}} + \sigma_f^2 + \kappa_{\mathbf{g}}^{\mathbf{2}}\sigma_{g_{\boldsymbol{\beta}}}^2\right) + V_{1,T} + \kappa_{\mathbf{g}}^{\mathbf{2}} H_{2,T}\right)$

Table 1: Bilevel local regret, $BLR_w(T)$, of OBBO and SOBBO vs. online bilevel benchmarks SOBOW from [24] and OAGD from [34]. Full derivation of benchmarks can be found in Appendix D

References

[1] Sergul Aydore, Tianhao Zhu, and Dean P Foster. Dynamic local regret for non-convex online forecasting. *Advances in Neural Information Processing Systems*, 32, 2019.

[2] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations Research*, 63(5), 2015.

[3] Jerome Bracken and James T. McGill. Mathematical programs with optimization problems in the constraints. *Operations Research*, 21(1), 1973. ISSN 0030364X, 15265463.

[4] L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3), 1967.

[5] Tianyi Chen, Yuejiao Sun, Quan Xiao, and Wotao Yin. A single-timescale method for stochastic bilevel optimization. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, 2022.

[6] Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. *Proceedings of the 25th Annual Conference on Learning Theory*, 2012.

[7] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61), 2011.

[8] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*, 2017.

[9] Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. Online meta-learning. *International Conference on Machine Learning*, 2019.

[10] Luca Franceschi, Michele Donini, Paolo Frasconi, and Massimiliano Pontil. Forward and reverse gradient-based hyperparameter optimization. *International Conference on Machine Learning*, 2017.

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

[12] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155 (1), 2016.

[13] David Grangier, Pierre Ablin, and Awni Hannun. Adaptive training distributions with scalable online bilevel optimization. *arXiv preprint arXiv:2311.11973*, 2023.

[14] Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. *International Conference on Machine Learning*, 2020.

[15] Eric Hall and Rebecca Willett. Dynamical models and tracking regret in online convex programming. *International Conference on Machine Learning*, 2013.

[16] Nadav Hallak, Panayotis Mertikopoulos, and Volkan Cevher. Regret minimization in stochastic non-convex learning via a proximal-gradient approach. *International Conference on Machine Learning*, 2021.

[17] Elad Hazan, Karan Singh, and Cyril Zhang. Efficient regret minimization in non-convex games. *International Conference on Machine Learning*, 2017.

[18] Feihu Huang, Junyi Li, and Shangqian Gao. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.

[19] Feihu Huang, Junyi Li, and Heng Huang. Super-Adam: Faster and universal framework of adaptive gradients. *Advances in Neural Information Processing Systems*, 34, 2021.

[20] Feihu Huang, Junyi Li, Shangqian Gao, and Heng Huang. Enhanced bilevel optimization via Bregman distance. *Advances in Neural Information Processing Systems*, 35, 2022.

[21] Kaiyi Ji, Junjie Yang, and Yingbin Liang. Bilevel optimization: Convergence analysis and enhanced design. *International Conference on Machine Learning*, 2021.

[22] Prashant Khanduri, Siliang Zeng, Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A near-optimal algorithm for stochastic bilevel optimization via double-momentum. *Advances in Neural Information Processing Systems*, 34, 2021.

[23] Brian Kulis and Peter L. Bartlett. Implicit online learning. *International Conference on Machine Learning*, 2010.

[24] Sen Lin, Daouda Sow, Kaiyi Ji, Yingbin Liang, and Ness Shroff. Non-convex bilevel optimization with time-varying objective functions. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. *International Conference on Artificial Intelligence and Statistics*, 2020.

[26] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. *International Conference on Machine Learning*, 2015.

[27] Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 15, 2011.

[28] H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010.

[29] Saypraseuth Mounsaveng, Issam Laradji, Ismail Ben Ayed, David Vazquez, and Marco Pedersoli. Learning data augmentation with online bilevel optimization for image classification. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021.

[30] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in Neural Information Processing Systems*, 31, 2018.

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

[32] Fabian Pedregosa. Hyperparameter optimization with approximate gradient. *International Conference on Machine Learning*, 2016.

[33] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. Meta-learning with implicit gradients. *Advances in Neural Information Processing Systems*, 32, 2019.

[34] Davoud Ataee Tarzanagh, Parvin Nazari, Bojian Hou, Li Shen, and Laura Balzano. Online bilevel optimization: Regret analysis of online alternating gradient methods. *International Conference on Artificial Intelligence and Statistics*, 2024.

[35] Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: Optimal dynamic regret of online learning with true and noisy gradient. *International Conference on Machine Learning*, 2016.

[36] Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. *Advances in Neural Information Processing Systems*, 33, 2020.

[37] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. AAAI Press, 2003. ISBN 1577351894.

Appendix A. Overview

A.1. Related Work

Online optimization is a well-developed area of research that extends the offline optimization problem to an objective function that can change over time. For a convex time-varying objective, gradient-based methods were proposed within [37] and achieved a sublinear rate of static regret. Improved rates of static regret are provided in [28] through utilizing a strongly-convex adaptive regularization function and are further generalized to a nonsmooth regularization term within [27]. In non-stationary and convex environments as in [15], [36], and [2], the static regret performance metric is too optimistic, and instead, a dynamic regret measure should be used to capture the changing comparator sequence. For a time-varying and nonconvex objective function, there has been recent interest in online optimization algorithms that achieve sublinear local regret, a stationary measure introduced within [17]. Recent work extends local regret to non-stationary environments, as in [1] and stochastic gradient oracles in [16]. The use of Bregman divergences has notably improved rates of regret in online optimization, as demonstrated by the seminal Adagrad algorithm [7], and further developed in implicit online learning algorithms in [23].

Offline bilevel optimization has a rich history with its introduction in constrained optimization of [3]. For applications in modern-day machine learning, gradient-based optimization methods are often preferred due to their ease of implementation and scalability. One class of methods is approximate implicit differentiation techniques, detailed in [32] and [25], which construct a hypergradient approximation through a conjugate gradient or fixed point approach. Another common method is to use iterative differentiation techniques and their efficient computational implementations to form an approximate hypergradient (see e.g., [26] and [10]). Convergence improvements are achieved in the Bio-BreD algorithm of [20] through incorporating Bregman divergences in the outer level gradient step. Moreover, implications of the Bregman-based optimizers of [20] allow for integration of adaptive learning rates and momentum in offline bilevel optimization, such as in the BiAdam framework of [18]. Due to the inherent uncertainty of utilizing sample batches, stochastic formulations of offline bilevel optimization are of recent interest. A stochastic approximation algorithm is introduced in [11] that utilizes a single stochastic gradient sample at each iteration. Convergence rates are improved in Stoc-BiO within [21] through an improved stochastic hypergradient estimator. Sample-efficient stochastic algorithms of SUSTAIN and STABLE are discussed in [22] and [5] by employing momentum-based variance reduction recurrences.

OBO is underdeveloped relative to single-level online optimization and offline bilevel optimization. The online alternating gradient descent algorithm (OAGD) of [34] is shown to achieve sublinear bilevel local regret for a nonconvex time-varying objective. The single-loop online bilevel optimizer with window averaging (SOBOW) of [24] offers a computational improvement while maintaining a sublinear rate. The general problem formulation of an OBO framework fits time-varying machine learning tasks not within the constraints of offline bilevel optimization, such as learning optimal dataset augmentations in [29] and online adaptation of pre-trained distributions of [13].

A.2. Concluding Remarks

In this work, we develop a novel family of algorithms parameterized by Bregman divergences. We prove that in addition to achieving sublinear bilevel local regret, our algorithms offer improvements in convergence rates by adapting to the underlying geometry of the problem. Furthermore, we introduce the first stochastic online bilevel optimizer and show that by utilizing a weighted average of recent stochastic approximated hypergradients, our algorithm achieves a sublinear rate of bilevel local regret in a sample efficient manner. Empirical results on machine learning tasks such as online hyperparameter optimization and online meta-learning validate our theoretical contributions and the practicality of our proposed methods.

Appendix B. Notation and Preliminaries

The next two lemmas introduce known results for smooth and strongly convex functions.

Lemma 4 *Suppose a function $g(\boldsymbol{\beta}) : \mathbb{R}^{d_2} \mapsto \mathbb{R}$ is ℓ-Lipschitz continuous with respect to $\|\cdot\|$. Then the following inequality holds $\forall \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^{d_2}$*

$$|g(\boldsymbol{\beta}_2) - g(\boldsymbol{\beta}_1)| \leq \ell \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\| \tag{10}$$

Lemma 5 *A function $g(\boldsymbol{\beta})$ that is μ_g strongly convex with respect to $\|\cdot\|$ satisfies the inequality $\forall \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^{d_2}$ of*

$$g(\boldsymbol{\beta}_2) - g(\boldsymbol{\beta}_1) \geq \langle \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1, \nabla g(\boldsymbol{\beta}_1) \rangle + \frac{\mu_g}{2} \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\|^2 \tag{11}$$

Lemma 6 *(Lemma 12 in [34]) For any set of vectors $\{\boldsymbol{\beta}_i\}_{i=1}^m$, it holds that*

$$\left\| \sum_{i=1}^m \boldsymbol{\beta}_i \right\|^2 \leq m \sum_{i=1}^m \|\boldsymbol{\beta}_i\|^2 \tag{12}$$

The following lemma provides progress bounds for gradient descent applied to a μ_g-strongly convex and twice differentiable function $g(\boldsymbol{\beta})$.

Lemma 7 *Let $g(\boldsymbol{\omega})$ be a twice differentiable and μ_g-strongly convex function with $\nabla g(\boldsymbol{\omega})$ satisfying $\ell_{g,1}$-Lipschitz continuity. Further assume $g(\boldsymbol{\omega})$ has a global minimizer $\widehat{\boldsymbol{\omega}}$ over the domain $\mathbb{R}^{d_2}$. Then under the gradient descent method of*

$$\boldsymbol{\omega}^k = \boldsymbol{\omega}^{k-1} - \eta \nabla g(\boldsymbol{\omega}^{k-1}),$$

the following satisfies for $\eta \leq \frac{1}{\ell_{g,1}}$

$$\left\| \boldsymbol{\omega}^k - \widehat{\boldsymbol{\omega}} \right\|^2 \leq (1 - \eta\mu_g) \left\| \boldsymbol{\omega}^{k-1} - \widehat{\boldsymbol{\omega}} \right\|^2,$$

The following two lemmas characterize useful properties of the generalized projection $\mathcal{G}_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{q}, \alpha)$.

Lemma 8 *(Lemma 1 in [12]) Let $\boldsymbol{\lambda}^+$ be from (3). Then $\forall \boldsymbol{u} \in \mathcal{X}$, $\boldsymbol{q} \in \mathbb{R}^{d_1}$, and $\alpha > 0$ we have*

$$\langle \boldsymbol{q}, \mathcal{G}_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{q}, \alpha) \rangle \geq \rho \|\mathcal{G}_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{q}, \alpha)\|^2 + \frac{1}{\alpha} \left(h(\boldsymbol{\lambda}^+) - h(\boldsymbol{u}) \right) \tag{13}$$

such that $\rho > 0$ is the strong convexity parameter of the distance generating function $\phi(\boldsymbol{\lambda})$.

Notation	**Description**
t	Time index
w	Window size
g_t	Inner level objective at t
f_t	Outer level objective at t
F_t	Reparameterized outer level objective at t
h	Convex and potentially nonsmooth regularization term
$\boldsymbol{\beta}_t$	Inner level variable at time t
$\boldsymbol{\lambda}_t$	Outer level variable at time t
ϵ	Error random variable for f_t
ζ	Error random variable for g_t
$\Vert\cdot\Vert$	The ℓ_2 norm for vectors (spectral norm for matrices)
$\mathcal{X}$	Decision set for outer level variable
S	Diameter of $\mathcal{X}$: $S = \max_{\boldsymbol{\lambda},\boldsymbol{\lambda}' \in \mathcal{X}} \left\Vert\boldsymbol{\lambda} - \boldsymbol{\lambda}'\right\Vert$
Q	Upper bound on F_t: $\sup_{\boldsymbol{\lambda} \in \mathcal{X}} \lvert F_t(\boldsymbol{\lambda})\rvert \leq Q$
$\nabla F_t(\boldsymbol{\lambda})$	Gradient of F_t w.r.t. $\boldsymbol{\lambda}$, i.e., the hypergradient
$\widetilde{\nabla} f_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$	Gradient estimate for $\nabla F_t(\boldsymbol{\lambda})$ given any $\boldsymbol{\lambda} \in \mathcal{X}$ and $\boldsymbol{\beta} \in \mathbb{R}^{d_2}$
$\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{\boldsymbol{t+1}})$	Time-smoothed $\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1})$ across a window of size w
s	Batch size for $\nabla_{\boldsymbol{\beta}} g_t(\cdot, \cdot, \bar{\zeta})$, that is $\left\lvert\bar{\zeta}\right\rvert = s$
m	Batch size for $\widetilde{\nabla} f_t(\cdot, \cdot, \mathcal{E})$, that is $\lvert\mathcal{E}\rvert = m$
α	Outer step size
η	Inner step size
K	Inner iteration count
$\ell_{f,0}$	Lipschitz constant for f_t
$\ell_{f,1}$	Lipschitz constant for ∇f_t
$\ell_{F,1}$	Lipschitz constant for ∇F_t
$\ell_{g,1}$	Lipschitz constant for $\nabla_{\boldsymbol{\beta}} g_t$
$\ell_{g,2}$	Lipschitz constant for $\nabla^2_{\boldsymbol{\beta},\boldsymbol{\beta}} g_t$
μ_g	Strong convexity parameter of g_t
κ_g	Condition number of g_t: $\kappa_g = \ell_{g,1}/\mu_g$
$\phi(\boldsymbol{\lambda})$	Continuously differentiable and strongly-convex distance generating function
$D_\phi(\cdot, \cdot)$	Bregman Divergence defined by $\phi_t(\boldsymbol{\lambda})$
ρ	Strong convexity parameter of $\phi_t(\boldsymbol{\lambda})$
L_1	First hypergradient error constant: $L_1 := \kappa_g(\ell_{g,1} + \mu_g)$
L_2	Second hypergradient error constant: $L_2 := \frac{2\ell_{f,0}\ell_{g,2}}{\mu_g}(1 + \kappa_g)$
L_3	Third hypergradient error constant: $L_3 := \ell_{f,0}\kappa_g$
$L_{\boldsymbol{\beta}}$	Hypergradient constant w.r.t. $\boldsymbol{\beta}_t$: $L_{\boldsymbol{\beta}} := L_1^2(1 - \eta\mu_g)^K + L_2^2(1 - \eta\mu_g)^{K-1}$
$\sigma^2_{g_{\boldsymbol{\beta}}}$	Finite variance of $\nabla_{\boldsymbol{\beta}} g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)$
σ^2_f	Finite variance of $\widetilde{\nabla} f_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \mathcal{E})$
$\mathcal{G}_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{q}, \alpha)$	Generalized projection $\forall \boldsymbol{u} \in \mathcal{X}, \boldsymbol{q} \in \mathbb{R}^{d_1}$, and $\alpha > 0$

Table 2: Summary of Notation

Lemma 9 *(Proposition 1 in [12]) Let $\mathcal{G}_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{q}, \alpha)$ be the generalized projection. Then $\forall \boldsymbol{q}_1, \boldsymbol{q}_2 \in \mathbb{R}^{d_1}$, $\forall \boldsymbol{u} \in \mathcal{X}$, $\forall \alpha > 0$, we have*

$$\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{q}_1, \alpha) - \mathcal{G}_{\mathcal{X}}(\boldsymbol{u}, \boldsymbol{q}_2, \alpha)\| \leq \frac{1}{\rho} \|\boldsymbol{q}_1 - \boldsymbol{q}_2\| \,. \tag{14}$$

The next Lemma provides useful bounds on the hypergradient $\nabla F_t(\boldsymbol{\lambda})$, gradient estimate $\nabla f_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$, and optimal inner level variables $\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})$ in the deterministic online bilevel optimization problem.

Lemma 10 *(Lemma 3 in [34]) Under assumptions A and B, it holds for all $t \in [1, T]$, $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \mathcal{X}$, and $\boldsymbol{\beta} \in \mathbb{R}^{d_2}$ that*

$$\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_1) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_2)\right\| \leq \kappa_g \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| \,, \tag{15}$$

where $\kappa_g := \frac{\ell_{g,1}}{\mu_g} = O(\kappa_g)$, the gradient estimator $\widetilde{\nabla} f_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ satisfies

$$\|\widetilde{\nabla} f_t(\boldsymbol{\lambda}, \boldsymbol{\beta}) - \nabla F_t(\boldsymbol{\lambda})\| \leq M_f \left\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})\right\| , \tag{16}$$

where $M_f := \ell_{f,1} + \ell_{f,1}\kappa_g + \frac{\ell_{f,0}\ell_{g,2}}{\mu_g}(1 + \kappa_g) = O(\kappa_g^2)$, and

$$\|\nabla F_t(\boldsymbol{\lambda}_1) - \nabla F_t(\boldsymbol{\lambda}_2)\| \leq \ell_{F,1} \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| \,, \tag{17}$$

where $\ell_{F,1} := \ell_{f,1}(1 + \kappa_g) + \frac{\ell_{f,0}\ell_{g,2}}{\mu_g}(1 + \kappa_g) + M_f\kappa_g = O(\kappa_g^3)$.

Lemma 11 provides an analytical form to compute the hypergradient via iterative differentiation.

Lemma 11 *(Proposition 2 in [21]) The partial $\frac{\partial f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}$ takes an analytical form of $\frac{\partial f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} =$*

$$\nabla_{\boldsymbol{\lambda}} f_t\left(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K\right) - \eta \sum_{k=0}^{K-1} \nabla^2_{\boldsymbol{\lambda},\boldsymbol{\omega}} g_t\left(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^k\right) H_{\boldsymbol{\omega},\boldsymbol{\omega}} \nabla_{\boldsymbol{\omega}} f_t\left(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K\right) , \tag{18}$$

where $H_{\boldsymbol{\omega},\boldsymbol{\omega}} := \prod_{j=k+1}^{K-1}\left(I_{d_2} - \eta\nabla^2_{\boldsymbol{\omega},\boldsymbol{\omega}} g_t\left(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^j\right)\right)$, the d_2-identity matrix is denoted I_{d_2}, with $\eta > 0$ and K as the step size and number of iterations for the inner loop.

Lemma 12 provides an upper bound on the hypergradient error when utilizing an iterative differentiation approach for estimation.

Lemma 12 *(Lemma 6 in [21]) Suppose Assumptions A and B are satisfied with $\eta < \frac{1}{\ell_{g,1}}$ and $K \geq 1$. Then we have $\forall t \in [1, T]$*

$$\begin{aligned} &\left\|\frac{\partial f_t(\boldsymbol{\lambda}, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_t(\boldsymbol{\lambda})\right\| \\ &\leq \left(L_1(1-\eta\mu_g)^{\frac{K}{2}} + L_2(1-\eta\mu_g)^{\frac{K-1}{2}}\right)\left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})\right\| + L_3(1-\eta\mu_g)^K, \end{aligned} \tag{19}$$

where $L_1 = \kappa_g(\ell_{g,1} + \mu_g)$, $L_2 = \frac{2\ell_{f,0}\ell_{g,2}}{\mu_g}(1 + \kappa_g)$, and $L_3 = \ell_{f,0}\kappa_g$.

Lemma 13 provides a analytical formula to compute a stochastic hypergradient estimator.

Lemma 13 *(Algorithm 3 in [11]) For a sample upper bound of m and independent $\mathcal{E}_t = \{\epsilon_t, \zeta_t^0, \ldots, \zeta_t^{m-1}\}$, the stochastic gradient of $\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{E}_t)$ provides an estimate of $\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1})$ and is constructed $\forall t \in [1, T]$ as*

$$\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{E}_t) := \nabla_{\boldsymbol{\lambda}} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \epsilon_t) - \nabla^2_{\boldsymbol{\lambda},\boldsymbol{\beta}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \zeta_t^0)$$
$$\times \left[\frac{m}{\ell_{g,1}} \prod_{j=1}^{\widetilde{m}} \left(I_{d_2} - \frac{1}{\ell_{g,1}} \nabla^2_{\boldsymbol{\beta}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \zeta_t^j) \right) \right] \nabla_{\boldsymbol{\beta}} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \epsilon_t), \tag{20}$$

where $\widetilde{m} \sim \mathcal{U}(0, 1, \ldots, m-1)$ and $\prod_{j=1}^{m=0}(\cdot) = I_{d_2}$.

The next Lemma characterizes the bias of the stochastic hypergradient estimate $\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{E}_t)$.

Lemma 14 *(Lemma 2.1 in [22]) Suppose Assumptions A,B, and E. For any $m \geq 1$ the gradient estimator of (20) satisfies*

$$\left\| \widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}) - \mathbb{E}_{\mathcal{E}_t} \left[\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{E}_t) \right] \right\| \leq \ell_{f,1} \kappa_g \left(1 - \frac{\mu_g}{\ell_{g,1}} \right)^m \tag{21}$$

Appendix C. Proof of Main Results

C.1. Deterministic Setting

First, we introduce some required lemmas. The following Lemma provides an upper bound on the cumulative difference between the time-smoothed outer level objective $F_{t,w}(\boldsymbol{\lambda})$ evaluated at $\boldsymbol{\lambda}_t$ and $\boldsymbol{\lambda}_{t+1}$ in terms of the outer level objective upper bound Q, window size w, and the comparator sequence on subsequent function evaluations $V_{1,T}$.

Lemma 15 *Suppose Assumption E. If our OBBO algorithm in Algorithm 1 is applied with window size $w \geq 1$ to generate the sequence $\{\boldsymbol{\lambda}_t\}_{t=1}^T$, then we have*

$$\sum_{t=1}^T \left(F_{t,w}(\boldsymbol{\lambda}_t) - F_{t,w}(\boldsymbol{\lambda}_{t+1}) \right) \leq \frac{2TQ}{w} + V_{1,T}.$$

where $V_{1,T} := \sum_{t=1}^T \sup_{\boldsymbol{\lambda} \in \mathcal{X}} \left[F_{t+1}(\boldsymbol{\lambda}) - F_t(\boldsymbol{\lambda}) \right]$

Proof By definition, in the deterministic setting, we have $F_t(\boldsymbol{\lambda}) \triangleq f_t\left(\boldsymbol{\lambda}, \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})\right)$. Then it holds

$$\sum_{t=1}^T \left(F_{t,w}(\boldsymbol{\lambda}_t) - F_{t,w}(\boldsymbol{\lambda}_{t+1}) \right)$$
$$= \sum_{t=1}^T \frac{1}{w} \sum_{i=0}^{w-1} \left(f_{t-i}\left(\boldsymbol{\lambda}_{t-i}, \widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-i})\right) - f_{t-i}\left(\boldsymbol{\lambda}_{t+1-i}, \widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t+1-i})\right) \right)$$

Which is equivalent to

$$\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\left(f_{t-i}\left(\boldsymbol{\lambda}_{t-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-i})\right)-f_{t-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t+1-i})\right)\right)$$

$$=\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\left(f_{t-i}\left(\boldsymbol{\lambda}_{t-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-i})\right)-f_{t+1-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t+1-i}(\boldsymbol{\lambda}_{t+1-i})\right)\right) \tag{22}$$

$$+\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\left(f_{t+1-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t+1-i}(\boldsymbol{\lambda}_{t+1-i})\right)-f_{t-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t+1-i})\right)\right) \tag{23}$$

For (22), we can write

$$\begin{aligned}&\frac{1}{w}\sum_{i=0}^{w-1}\left(f_{t-i}\left(\boldsymbol{\lambda}_{t-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-i})\right)-f_{t+1-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t+1-i}(\boldsymbol{\lambda}_{t+1-i})\right)\right)\\
&=\frac{1}{w}\left[f_t\left(\boldsymbol{\lambda}_t,\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right)+\ldots+f_{t+1-w}\left(\boldsymbol{\lambda}_{t+1-w},\widehat{\boldsymbol{\beta}}_{t+1-w}(\boldsymbol{\lambda}_{t+1-w})\right)\right]\\
&-\frac{1}{w}\left[f_{t+1}\left(\boldsymbol{\lambda}_{t+1},\widehat{\boldsymbol{\beta}}_{t+1}(\boldsymbol{\lambda}_{t+1})\right)+\ldots+f_{t+2-w}\left(\boldsymbol{\lambda}_{t+2-w},\widehat{\boldsymbol{\beta}}_{t+2-w}(\boldsymbol{\lambda}_{t+2-w})\right)\right]\\
&=\frac{1}{w}\left[f_{t+1-w}\left(\boldsymbol{\lambda}_{t+1-w},\widehat{\boldsymbol{\beta}}_{t+1-w}(\boldsymbol{\lambda}_{t+1-w})\right)-f_{t+1}\left(\boldsymbol{\lambda}_{t+1},\widehat{\boldsymbol{\beta}}_{t+1}(\boldsymbol{\lambda}_{t+1})\right)\right]\\
&=\frac{1}{w}\left(F_{t+1-w}(\boldsymbol{\lambda}_{t+1-w})-F_{t+1}(\boldsymbol{\lambda}_{t+1})\right)\leq\frac{2Q}{w},\end{aligned} \tag{24}$$

where the last inequality comes from Assumption E. Note (23) can be bounded through

$$\begin{aligned}&\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\left(f_{t+1-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t+1-i}(\boldsymbol{\lambda}_{t+1-i})\right)-f_{t-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t+1-i})\right)\right)\\
&\leq\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\sup_{\boldsymbol{\lambda}\in\mathcal{X}}\left[f_{t+1-i}\left(\boldsymbol{\lambda},\widehat{\boldsymbol{\beta}}_{t+1-i}(\boldsymbol{\lambda})\right)-f_{t-i}\left(\boldsymbol{\lambda},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda})\right)\right]\leq V_{1,T}\end{aligned} \tag{25}$$

Combining (24) and (25) results in the upper bound of

$$\sum_{t=1}^{T}\left(F_{t,w}(\boldsymbol{\lambda}_t)-F_{t,w}(\boldsymbol{\lambda}_{t+1})\right)\leq\frac{2TQ}{w}+V_{1,T}.$$

■

The next Lemma provides an upper bound on the error of $\left\|\boldsymbol{\beta}_t-\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2$ for all $t\in[1,T]$ in terms of an initial error, the cumulative differences of the outer level variable, and the cumulative differences of the optimal inner level variables.

Lemma 16 *Suppose Assumptions A and B. Choose the inner step size of η and inner iteration count of K to satisfy*

$$\eta<\min\left(\frac{1}{\ell_{g,1}},\frac{1}{\mu_g}\right),\ \textit{and}\quad K\geq 1,$$

and define the decay parameter ν, inner level variable error constant C_{μ_g}, and initial error $\Delta_{\boldsymbol{\beta}}$ respectively as

$$\nu := \left(1 - \frac{\eta\mu_g}{2}\right)(1-\eta\mu_g)^{K-1}, \text{ and } \quad C_{\mu_g} := \left(1 + \frac{2}{\eta\mu_g}\right), \\ \text{and} \quad \Delta_{\boldsymbol{\beta}} := \left\|\boldsymbol{\beta}_1 - \widehat{\boldsymbol{\beta}}_1(\boldsymbol{\lambda}_1)\right\|^2.$$

Then our OBBO algorithm in Algorithm 1 guarantees $\forall t \in [1, T]$

$$\left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 \leq \nu^{t-1}\Delta_{\boldsymbol{\beta}} \\ +2C_{\mu_g}\kappa_g^2 \sum_{j=0}^{t-2} \nu^j \left\|\boldsymbol{\lambda}_{t-1-j} - \boldsymbol{\lambda}_{t-j}\right\|^2 + 2C_{\mu_g}\sum_{j=0}^{t-2}\nu^j \left\|\widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j}) - \widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j})\right\|^2. \quad (26)$$

Proof By definition for $t = 1$, we have $\left\|\boldsymbol{\beta}_1 - \widehat{\boldsymbol{\beta}}_1(\boldsymbol{\lambda}_1)\right\|^2 = \Delta_{\boldsymbol{\beta}}$. Then $\forall t \in [2, T]$

$$\left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 = \left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) + \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2, \quad (27)$$

which can be expanded based on the Young's Inequality for any $\delta > 0$ as

$$\left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) + \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 \\ \leq (1+\delta)\left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2 \\ + \left(1 + \frac{1}{\delta}\right)\left\|\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2.$$

Now it holds that

$$\left\|\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 \leq 2\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 + 2\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2$$

which through Lemma 10 can be further upper bounded with the Lipschitz constant of κ_g as

$$\left\|\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 \leq 2\kappa_g^2\left\|\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t\right\|^2 + 2\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2$$

Combining above, we see that $\forall \delta > 0$, (27) is upper bounded as

$$\left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 \leq (1+\delta)\left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2 \\ +2\left(1+\frac{1}{\delta}\right)\kappa_g^2\left\|\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t\right\|^2 + 2\left(1+\frac{1}{\delta}\right)\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2. \quad (28)$$

As $\eta < \frac{1}{\ell_{g,1}}$, we apply Lemma 7 to see

$$(1+\delta)\left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2 \leq (1+\delta)(1-\eta\mu_g)^K\left\|\boldsymbol{\beta}_{t-1} - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2$$

Now setting $\delta = \frac{\eta\mu_g}{2} > 0$ implies that

$$(1+\delta)(1-\eta\mu_g)^K = (1+\frac{\eta\mu_g}{2})(1-\eta\mu_g)^K < \left(1-\frac{\eta\mu_g}{2}\right)(1-\eta\mu_g)^{K-1} < 1$$

Using $\nu := \left(1-\frac{\eta\mu_g}{2}\right)(1-\eta\mu_g)^{K-1}$ in (28), we get

$$\nu \left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 \le \nu^2 \left\|\boldsymbol{\beta}_{t-1} - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2$$
$$+2C_{\mu_g}\nu\kappa_g^2 \left\|\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t\right\|^2 + 2C_{\mu_g}\nu \left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2,$$

where $C_{\mu_g} = \left(1+\frac{2}{\eta\mu_g}\right)$. Starting at $t=T$ and unrolling backward to $t=1$, results in the upper bound of

$$\left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 \le \nu^{t-1}\Delta_{\boldsymbol{\beta}}$$
$$+2C_{\mu_g}\kappa_g^2 \sum_{j=0}^{t-2} \nu^j \left\|\boldsymbol{\lambda}_{t-1-j} - \boldsymbol{\lambda}_{t-j}\right\|^2 + 2C_{\mu_g} \sum_{j=0}^{t-2} \nu^j \left\|\widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j}) - \widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j})\right\|^2.$$

■

The next Lemma utilizes Lemma 12 and Lemma 16 to derive an upper bound on the hypergradient error $\forall t \in [1,T]$ in terms of discounted variations of the (i) cumulative time-smoothed hypergradient error; (ii) bilevel local regret; and (iii) cumulative difference between optimal inner-level variables. A final term is included, composed of a discounted initial error and smoothness term of the inner objective.

Lemma 17 *Suppose Assumptions A, B, D, and F. Choose the inner step size of η and inner iteration count of K to satisfy*

$$\eta < \min\left(\frac{1}{\ell_{g,1}}, \frac{1}{\mu_g}\right), \textit{ and } \quad K \ge 1.$$

Using the definitions of ν, C_{μ_g}, and $\Delta_{\boldsymbol{\beta}}$ from Lemma 16 as well as the further definition of

$$L_{\boldsymbol{\beta}} := L_1^2(1-\eta\mu_g)^K + L_2^2(1-\eta\mu_g)^{K-1},$$

then the hypergradient error from our OBBO algorithm in Algorithm 1 is bounded $\forall t \in [1,T]$ as

$$\left\|\frac{\partial f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_t(\boldsymbol{\lambda}_t)\right\|^2 \le \delta_t + A\sum_{j=0}^{t-2}\nu^j \left\|\frac{\partial f_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\omega}_{t-1-j}^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j})\right\|^2$$
$$+B\sum_{j=0}^{t-2}\nu^j \left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha\right)\right\|^2 + C\sum_{j=0}^{t-2}\nu^j \left\|\widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j}) - \widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j})\right\|^2, \tag{29}$$

where $\delta_t = 3L_3^2(1-\eta\mu_g)^{2K} + 3L_{\boldsymbol{\beta}}\nu^{t-1}\Delta_{\boldsymbol{\beta}}$ and $A = \frac{12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}}\kappa_g^2}{\rho^2}$, $B = 12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}}\kappa_g^2$, and $C = 6L_{\boldsymbol{\beta}}C_{\mu_g}$.

Proof Note that Lemma 12 implies $\forall t \in [1, T]$

$$\left\|\frac{\partial f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_t(\boldsymbol{\lambda}_t)\right\|^2 \leq 3L_{\boldsymbol{\beta}} \left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 + 3L_3^2(1-\eta\mu_g)^{2K}.$$

Substituting the upper bound on $\left\|\boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2$ from Lemma 16, we have

$$\begin{aligned}\left\|\frac{\partial f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_t(\boldsymbol{\lambda}_t)\right\|^2 &\leq 3L_3^2(1-\eta\mu_g)^{2K} \\ +3L_{\boldsymbol{\beta}} &\left(\nu^{t-1}\Delta_{\boldsymbol{\beta}} + 2C_{\mu_g}\kappa_g^2 \sum_{j=0}^{t-2} \nu^j \left\|\boldsymbol{\lambda}_{t-1-j} - \boldsymbol{\lambda}_{t-j}\right\|^2\right) \\ +6L_{\boldsymbol{\beta}} C_{\mu_g} &\sum_{j=0}^{t-2} \nu^j \left\|\widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j}) - \widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j})\right\|^2,\end{aligned}$$

By definition, we have $\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_{t-1-j}, \frac{\partial f_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\omega}_{t-1-j}^K)}{\partial \boldsymbol{\lambda}}, \alpha\right) := \frac{1}{\alpha}\left(\boldsymbol{\lambda}_{t-1-j} - \boldsymbol{\lambda}_{t-j}\right)$

$$\begin{aligned}\sum_{j=0}^{t-2} \nu^j \left\|\boldsymbol{\lambda}_{t-1-j} - \boldsymbol{\lambda}_{t-j}\right\|^2 &= \alpha^2 \sum_{j=0}^{t-2} \nu^j \left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_{t-1-j}, \frac{\partial f_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\omega}_{t-1-j}^K)}{\partial \boldsymbol{\lambda}}, \alpha\right)\right\|^2 \\ &\leq 2\alpha^2 \sum_{j=0}^{t-2} \nu^j \left(\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha)\right\|^2\right) \\ +2\alpha^2 \sum_{j=0}^{t-2} \nu^j &\left(\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_{t-1-j}, \frac{\partial f_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\omega}_{t-1-j}^K)}{\partial \boldsymbol{\lambda}}, \alpha\right) - \mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha\right)\right\|^2\right) \\ &\leq 2\alpha^2 \sum_{j=0}^{t-2} \nu^j \left(\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha)\right\|^2\right) \\ +2\alpha^2 \sum_{j=0}^{t-2} \nu^j &\left(\frac{1}{\rho^2}\left\|\frac{\partial f_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\omega}_{t-1-j}^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j})\right\|^2\right)\end{aligned} \tag{30}$$

such that the last inequality comes from Lemma 9. Rearranging terms, we have decomposed the hypergradient error term at t in terms of the cumulative hypergradient error from $j = 1, \ldots, t-1$

$$\begin{aligned} \left\| \frac{\partial f_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_t(\boldsymbol{\lambda}_t) \right\|^2 &\leq 3L_3^2(1-\eta\mu_g)^{2K} + 3L_{\boldsymbol{\beta}}\nu^{t-1}\Delta_{\boldsymbol{\beta}} \\ &+ 12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}} \kappa_g^2 \sum_{j=0}^{t-2} \nu^j \left\| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha) \right\|^2 \\ + \frac{12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}} \kappa_g^2}{\rho^2} \sum_{j=0}^{t-2} \nu^j &\left\| \frac{\partial f_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\omega}_{t-1-j}^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}) \right\|^2 \\ &+ 6L_{\boldsymbol{\beta}} C_{\mu_g} \sum_{j=0}^{t-2} \nu^j \left\| \widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j}) - \widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j}) \right\|^2, \end{aligned}$$

∎

The next Lemma provides an upper bound on the cumulative time-smoothed hypergradient error using the result of Lemma 17.

Lemma 18 *Suppose Assumptions A, B, D, and F. Choose the inner step size of* $\eta < \min\left(\frac{1}{\ell_{g,1}}, \frac{1}{\mu_g}\right)$, *the outer step size* $\alpha \leq \frac{\rho\sqrt{(1-\nu)}}{\kappa_g\sqrt{108 C_{\mu_g} L_{\boldsymbol{\beta}}}}$, *and inner iteration count* $K = \frac{\log(T)}{\log((1-\eta\mu_g)^{-1})} + 1$. *Then the cumulative time-smoothed hypergradient error from our OBBO algorithm in Algorithm 1 satisfies*

$$\begin{aligned} \sum_{t=1}^{T} \left\| \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}(\boldsymbol{\lambda}_t) \right\|^2 &\leq \frac{27}{8}\left(\frac{\Delta_{\boldsymbol{\beta}} L_{\boldsymbol{\beta}}}{(1-\nu)} + L_3^2 \right) \\ + \frac{\rho^2}{8} \sum_{t=1}^{T} \left\| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha) \right\|^2 &+ \frac{27 L_{\boldsymbol{\beta}} C_{\mu_g}}{2(1-\nu)} \sum_{t=2}^{T} \left\| \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2, \end{aligned}$$

Proof Note by definition of the time-smoothed outer level objective and application of Young's inequality we have

$$\begin{aligned}
\left\|\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}} - \nabla F_{t,w}(\boldsymbol{\lambda}_t)\right\|^2 &= \left\|\frac{1}{w}\sum_{i=0}^{w-1}\left[\frac{\partial f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\omega}_{t-i}^K)}{\partial\boldsymbol{\lambda}} - \nabla F_{t-i}(\boldsymbol{\lambda}_{t-i})\right]\right\|^2 \\
= \left[\sum_{i=0}^{w-1}\frac{1}{w}\sum_{j=0}^{w-1}\frac{1}{w}\left\langle\frac{\partial f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\omega}_{t-i}^K)}{\partial\boldsymbol{\lambda}} - \nabla F_{t-i}(\boldsymbol{\lambda}_{t-i}), \frac{\partial f_{t-j}(\boldsymbol{\lambda}_{t-j},\boldsymbol{\omega}_{t-j}^K)}{\partial\boldsymbol{\lambda}} - \nabla F_{t-j}(\boldsymbol{\lambda}_{t-j})\right\rangle\right]& \\
&\leq [\sum_{i=0}^{w-1}\frac{1}{w}\sum_{j=0}^{w-1}\frac{1}{w}(\frac{1}{2}\left\|\frac{\partial f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\omega}_{t-i}^K)}{\partial\boldsymbol{\lambda}} - \nabla F_{t-i}(\boldsymbol{\lambda}_{t-i})\right\|^2 \\
&\quad +\frac{1}{2}\left\|\frac{\partial f_{t-j}(\boldsymbol{\lambda}_{t-j},\boldsymbol{\omega}_{t-j}^K)}{\partial\boldsymbol{\lambda}} - \nabla F_{t-j}(\boldsymbol{\lambda}_{t-j})\right\|^2)] \\
&= \frac{1}{w}\sum_{i=0}^{w-1}\left\|\frac{\partial f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\omega}_{t-i}^K)}{\partial\boldsymbol{\lambda}} - \nabla F_{t-i}(\boldsymbol{\lambda}_{t-i})\right\|^2
\end{aligned} \tag{31}$$

Substituting the upper bound on $\left\|\frac{\partial f_t(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}} - \nabla F_t(\boldsymbol{\lambda}_t)\right\|^2$ from Lemma 17 and re-indexing the bilevel local regret and the cumulative time-smoothed hypergradient error, we construct the upper bound of

$$\begin{aligned}
&\sum_{t=1}^{T}\left\|\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}} - \nabla F_{t,w}(\boldsymbol{\lambda}_t)\right\|^2 \\
&\leq \sum_{t=1}^{T}\frac{1}{w}\left[\sum_{i=0}^{w-1}\left(3L_3^2(1-\eta\mu_g)^{2K} + 3L_{\boldsymbol{\beta}}\nu^{t-i-1}\Delta_{\boldsymbol{\beta}}\right)\right] \\
&+\sum_{t=1}^{T}\frac{1}{w}\left[\sum_{i=0}^{w-1}\left(12\alpha^2 C_{\mu_g}L_{\boldsymbol{\beta}}\kappa_g^2\sum_{j=0}^{t-i-2}\nu^j\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-i-j},\nabla F_{t-i-j,w}(\boldsymbol{\lambda}_{t-i-j}),\alpha)\right\|^2\right)\right] \\
&+\sum_{t=1}^{T}\frac{1}{w}\left[\sum_{i=0}^{w-1}\left(\frac{12\alpha^2 C_{\mu_g}L_{\boldsymbol{\beta}}\kappa_g^2}{\rho^2}\sum_{j=0}^{t-i-2}\nu^j\left\|\frac{\partial f_{t-i-j,w}(\boldsymbol{\lambda}_{t-i-j},\boldsymbol{\omega}_{t-i-j}^K)}{\partial\boldsymbol{\lambda}} - \nabla F_{t-i-j,w}(\boldsymbol{\lambda}_{t-i-j})\right\|^2\right)\right] \\
&+\sum_{t=2}^{T}\frac{1}{w}\left[\sum_{i=0}^{w-1}\left(6L_{\boldsymbol{\beta}}C_{\mu_g}\sum_{j=0}^{t-i-2}\nu^j\left\|\widehat{\boldsymbol{\beta}}_{t-i-j}(\boldsymbol{\lambda}_{t-i-1-j}) - \widehat{\boldsymbol{\beta}}_{t-i-1-j}(\boldsymbol{\lambda}_{t-i-1-j})\right\|^2\right)\right]
\end{aligned} \tag{32}$$

Given $\nu < 1$, it holds that $\sum_{j=0}^{t-2} \nu^j < \sum_{j=0}^{\infty} \nu^j = \frac{1}{1-\nu}$, which lets us upper bound (32) as

$$\begin{aligned}
&\sum_{t=1}^{T} \left\| \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}(\boldsymbol{\lambda}_t) \right\|^2 \\
&\leq \sum_{t=1}^{T} \frac{1}{w} \left[\sum_{i=0}^{w-1} \left(3L_3^2 (1-\eta\mu_g)^{2K} + 3L_{\boldsymbol{\beta}} \nu^{t-i-1} \Delta_{\boldsymbol{\beta}} \right) \right] \\
&+ \frac{12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}} \kappa_g^2}{(1-\nu)} \sum_{t=1}^{T} \frac{1}{w} \left[\sum_{i=0}^{w-1} \| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-i}, \nabla F_{t-i,w}(\boldsymbol{\lambda}_{t-i}), \alpha) \|^2 \right] \\
&+ \frac{12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}} \kappa_g^2}{\rho^2(1-\nu)} \sum_{t=1}^{T} \frac{1}{w} \left[\sum_{i=0}^{w-1} \left\| \frac{\partial f_{t-i,w}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\omega}_{t-i}^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t-i,w}(\boldsymbol{\lambda}_{t-i}) \right\|^2 \right] \\
&+ \frac{6L_{\boldsymbol{\beta}} C_{\mu_g}}{(1-\nu)} \sum_{t=2}^{T} \frac{1}{w} \left[\sum_{i=0}^{w-1} \left\| \widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-i-1}) - \widehat{\boldsymbol{\beta}}_{t-i-1}(\boldsymbol{\lambda}_{t-i-1}) \right\|^2 \right].
\end{aligned}$$

and further

$$\begin{aligned}
&\sum_{t=1}^{T} \left\| \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}(\boldsymbol{\lambda}_t) \right\|^2 \\
&\leq \sum_{t=1}^{T} \left(3L_3^2 (1-\eta\mu_g)^{2K} + 3L_{\boldsymbol{\beta}} \nu^{t-1} \Delta_{\boldsymbol{\beta}} \right) + \frac{12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}} \kappa_g^2}{(1-\nu)} \sum_{t=1}^{T} \| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha) \|^2 \\
&+ \frac{12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}} \kappa_g^2}{\rho^2(1-\nu)} \sum_{t=1}^{T} \left\| \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}(\boldsymbol{\lambda}_t) \right\|^2 + \frac{6L_{\boldsymbol{\beta}} C_{\mu_g}}{(1-\nu)} \sum_{t=2}^{T} \left\| \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2
\end{aligned}$$

which implies that

$$\begin{aligned}
&\left(1 - \frac{12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}} \kappa_g^2}{\rho^2(1-\nu)} \right) \sum_{t=1}^{T} \left\| \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}(\boldsymbol{\lambda}_t) \right\|^2 \\
&\leq \frac{3\Delta_{\boldsymbol{\beta}} L_{\boldsymbol{\beta}}}{1-\nu} + \sum_{t=1}^{T} \left(3L_3^2 (1-\eta\mu_g)^{2K} \right) + \frac{12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}} \kappa_g^2}{(1-\nu)} \sum_{t=1}^{T} \| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha) \|^2 \\
&+ \frac{6L_{\boldsymbol{\beta}} C_{\mu_g}}{(1-\nu)} \sum_{t=2}^{T} \left\| \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2,
\end{aligned}$$

Setting $K = \log(T)/\log\left((1-\eta\mu_g)^{-1}\right) + 1$ and $0 < \alpha \leq \frac{\rho\sqrt{(1-\nu)}}{\kappa_g \sqrt{108 C_{\mu_g} L_{\boldsymbol{\beta}}}}$

$$\left(1 - \frac{12\alpha^2 C_{\mu_g} L_{\boldsymbol{\beta}} \kappa_g^2}{\rho^2(1-\nu)} \right) \geq \frac{8}{9}$$

implies the upper bound of

$$\sum_{t=1}^{T}\left\|\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}}-\nabla F_{t,w}(\boldsymbol{\lambda}_t)\right\|^2\leq\frac{27}{8}\left(\frac{\Delta_{\boldsymbol{\beta}}L_{\boldsymbol{\beta}}}{(1-\nu)}+L_3^2\right)$$
$$+\frac{\rho^2}{8}\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\|^2+\frac{27L_{\boldsymbol{\beta}}C_{\mu_g}}{2(1-\nu)}\sum_{t=2}^{T}\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1})-\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2,$$

■

The next theorem presents the theoretical contribution for our OBBO algorithm in Algorithm 1. For suitably chosen step sizes, the sequence of iterates $\{\boldsymbol{\lambda}_t\}_{t=1}^T$ achieves sublinear bilevel local regret.

Theorem 19 *Suppose Assumptions A, B, D, E, F. Choose the inner step size of* $\eta<\min\left(\frac{1}{\ell_{g,1}},\frac{1}{\mu_g}\right)$*, the outer step size of* $\alpha\leq\min\left\{\frac{3\rho}{4\ell_{F,1}},\frac{\rho\sqrt{(1-\nu)}}{\kappa_g\sqrt{108C_{\mu_g}L_{\boldsymbol{\beta}}}}\right\}$*, and inner iteration count* $K=\frac{\log(T)}{\log((1-\eta\mu_g)^{-1})}+1$*. Then the bilevel local regret of our OBBO algorithm in Algorithm 1 satisfies the bound of*

$$BLR_w(T):=\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\|^2\leq O\left(\frac{T}{w}+V_{1,T}+\kappa_g^2H_{2,T}\right),\tag{33}$$

Proof Note, with Assumption A we have the upper bound of

$$\begin{aligned}F_{t,w}(\boldsymbol{\lambda}_{t+1})-F_{t,w}(\boldsymbol{\lambda}_t)&=\frac{1}{w}\sum_{i=0}^{w-1}F_{t-i}(\boldsymbol{\lambda}_{t+1-i})-\frac{1}{w}\sum_{i=0}^{w-1}F_{t-i}(\boldsymbol{\lambda}_{t-i})\\&=\frac{1}{w}\sum_{i=0}^{w-1}\left[F_{t-i}(\boldsymbol{\lambda}_{t+1-i})-F_{t-i}(\boldsymbol{\lambda}_{t-i})\right]\\&\leq\frac{1}{w}\sum_{i=0}^{w-1}\left[\langle\nabla F_{t-i}(\boldsymbol{\lambda}_{t-i}),\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\rangle+\frac{\ell_{F,1}}{2}\|\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\|^2\right]\\&=\langle\nabla F_{t,w}(\boldsymbol{\lambda}_t),\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\rangle+\frac{\ell_{F,1}}{2}\|\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\|^2.\end{aligned}$$

Substituting in $\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t,\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}},\alpha\right):=\frac{1}{\alpha}(\boldsymbol{\lambda}_t-\boldsymbol{\lambda}_{t+1})$,

$$\begin{aligned}&F_{t,w}(\boldsymbol{\lambda}_{t+1})-F_{t,w}(\boldsymbol{\lambda}_t)\leq\langle\nabla F_{t,w}(\boldsymbol{\lambda}_t),\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\rangle+\frac{\ell_{F,1}}{2}\|\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\|^2\\&=-\alpha\left\langle\nabla F_{t,w}(\boldsymbol{\lambda}_t),\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t,\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}},\alpha\right)\right\rangle+\frac{\alpha^2\ell_{F,1}}{2}\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t,\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}},\alpha\right)\right\|^2,\\&=-\alpha\left\langle\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}},\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t,\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}},\alpha\right)\right\rangle\\&+\alpha\left\langle\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}}-\nabla F_{t,w}(\boldsymbol{\lambda}_t),\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t,\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}},\alpha\right)\right\rangle\\&+\frac{\alpha^2\ell_{F,1}}{2}\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t,\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}},\alpha\right)\right\|^2.\end{aligned}\tag{34}$$

Using Lemma 8 with $\boldsymbol{q} = \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}$, note that

$$\alpha \left\langle \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \mathcal{G}_{\mathcal{X}} \left(\boldsymbol{\lambda}_t, \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \alpha \right) \right\rangle \\ \geq \alpha\rho \left\| \mathcal{G}_{\mathcal{X}} \left(\boldsymbol{\lambda}_t, \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \alpha \right) \right\|^2 + h(\boldsymbol{\lambda}_{t+1}) - h(\boldsymbol{\lambda}_t) \tag{35}$$

and further we get the following based on a variation of Young's Inequality

$$\left\langle \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right), \mathcal{G}_{\mathcal{X}} \left(\boldsymbol{\lambda}_t, \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \alpha \right) \right\rangle \\ \leq \frac{1}{\rho} \left\| \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right) \right\|^2 + \frac{\rho}{4} \left\| \mathcal{G}_{\mathcal{X}} \left(\boldsymbol{\lambda}_t, \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \alpha \right) \right\|^2 \tag{36}$$

Using (35) and (36) in (34) we get

$$F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right) - F_{t,w}\left(\boldsymbol{\lambda}_t\right) \leq \left(\frac{\alpha^2 \ell_{F,1}}{2} - \frac{3\alpha\rho}{4} \right) \left\| \mathcal{G}_{\mathcal{X}} \left(\boldsymbol{\lambda}_t, \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \alpha \right) \right\|^2 \\ + \frac{\alpha}{\rho} \left\| \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right) \right\|^2 + h(\boldsymbol{\lambda}_t) - h(\boldsymbol{\lambda}_{t+1}) \tag{37}$$

which as $0 < \alpha \leq \frac{3\rho}{4\ell_{F,1}}$ results in the further upper bound of

$$F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right) - F_{t,w}\left(\boldsymbol{\lambda}_t\right) \leq -\frac{3\alpha\rho}{8} \left\| \mathcal{G}_{\mathcal{X}} \left(\boldsymbol{\lambda}_t, \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \alpha \right) \right\|^2 \\ + \frac{\alpha}{\rho} \left\| \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right) \right\|^2 + h(\boldsymbol{\lambda}_t) - h(\boldsymbol{\lambda}_{t+1}) \tag{38}$$

Further note we can upper bound the local regret as

$$\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\|^2 \leq 2 \left\| \mathcal{G}_{\mathcal{X}} \left(\boldsymbol{\lambda}_t, \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \alpha \right) \right\|^2 \\ + 2 \left\| \mathcal{G}_{\mathcal{X}} \left(\boldsymbol{\lambda}_t, \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \alpha \right) - \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha) \right\|^2 \\ \leq 2 \left\| \mathcal{G}_{\mathcal{X}} \left(\boldsymbol{\lambda}_t, \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \alpha \right) \right\|^2 + \frac{2}{\rho^2} \left\| \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right) \right\|^2,$$

where the last inequality comes from Lemma 9. This then implies that

$$-\left\| \mathcal{G}_{\mathcal{X}} \left(\boldsymbol{\lambda}_t, \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}}, \alpha \right) \right\|^2 \leq -\frac{1}{2} \|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\|^2 \\ + \frac{1}{\rho^2} \left\| \frac{\partial f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^K)}{\partial \boldsymbol{\lambda}} - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right) \right\|^2 \tag{39}$$

Substituting (39) into (38) gives us

$$F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)-F_{t,w}\left(\boldsymbol{\lambda}_t\right)\leq-\frac{3\alpha\rho}{16}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\right\|^2$$
$$+\left(\frac{\alpha}{\rho}+\frac{3\alpha}{8\rho}\right)\left\|\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}}-\nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2+h(\boldsymbol{\lambda}_t)-h(\boldsymbol{\lambda}_{t+1}). \tag{40}$$

Rearranging we see

$$\frac{3\alpha\rho}{16}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\right\|^2\leq F_{t,w}\left(\boldsymbol{\lambda}_t\right)-F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)$$
$$+\frac{11\alpha}{8\rho}\left\|\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}}-\nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2+h(\boldsymbol{\lambda}_t)-h(\boldsymbol{\lambda}_{t+1}). \tag{41}$$

Summing from $1,\ldots,T$ and telescoping $h(\boldsymbol{\lambda}_t)$

$$\frac{3\alpha\rho}{16}\sum_{t=1}^{T}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\right\|^2\leq\sum_{t=1}^{T}\left(F_{t,w}\left(\boldsymbol{\lambda}_t\right)-F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)\right)$$
$$+\frac{11\alpha}{8\rho}\sum_{t=1}^{T}\left(\left\|\frac{\partial f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\omega}_t^K)}{\partial\boldsymbol{\lambda}}-\nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2\right)+\Delta_h,$$

where $\Delta_h:=h(\boldsymbol{\lambda}_1)-h(\boldsymbol{\lambda}_{T+1})$ Then we can substitute Lemma 18 to get

$$\frac{3\alpha\rho}{16}\sum_{t=1}^{T}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\right\|^2\leq\sum_{t=1}^{T}\left(F_{t,w}\left(\boldsymbol{\lambda}_t\right)-F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)\right)$$
$$+\frac{11\alpha}{8\rho}\left(\frac{27}{8}\left(\frac{\Delta_{\boldsymbol{\beta}}L_{\boldsymbol{\beta}}}{(1-\nu)}+L_3^2\right)+\frac{\rho^2}{8}\sum_{t=1}^{T}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\right\|^2\right)$$
$$+\frac{11\alpha}{8\rho}\left(\frac{27L_{\boldsymbol{\beta}}C_{\mu_g}}{2(1-\nu)}\sum_{t=2}^{T}\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1})-\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2\right)+\Delta_h.$$

Rearranging we have

$$\frac{12\alpha\rho}{64}\sum_{t=1}^{T}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\right\|^2\leq\sum_{t=1}^{T}\left(F_{t,w}\left(\boldsymbol{\lambda}_t\right)-F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)\right)$$
$$+\frac{11\alpha\rho}{64}\sum_{t=1}^{T}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\right\|^2+\frac{11\alpha}{8\rho}\left(\frac{27}{8}\left(\frac{\Delta_{\boldsymbol{\beta}}L_{\boldsymbol{\beta}}}{(1-\nu)}+L_3^2\right)\right)$$
$$+\frac{11\alpha}{8\rho}\left(\frac{27L_{\boldsymbol{\beta}}C_{\mu_g}}{2(1-\nu)}\sum_{t=2}^{T}\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1})-\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2\right)+\Delta_h,$$

or more succinctly

$$\sum_{t=1}^{T}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\right\|^2\leq\frac{64}{\alpha\rho}\sum_{t=1}^{T}\left(F_{t,w}\left(\boldsymbol{\lambda}_t\right)-F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)\right)$$
$$+\frac{88}{\rho^2}\left(\frac{27}{8}\left(\frac{\Delta_{\boldsymbol{\beta}}L_{\boldsymbol{\beta}}}{(1-\nu)}+L_3^2\right)\right)+\frac{88}{\rho^2}\frac{27L_{\boldsymbol{\beta}}C_{\mu_g}}{2(1-\nu)}\sum_{t=2}^{T}\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1})-\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2+\frac{64\Delta_h}{\alpha\rho}. \tag{42}$$

Applying Lemma 15 we see

$$\sum_{t=1}^{T}\left(F_{t,w}\left(\boldsymbol{\lambda}_t\right)-F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)\right)\leq\frac{2TQ}{w}+V_{1,T},\tag{43}$$

which by using (43) in (42) we get for $L_{\boldsymbol{\beta}}=O(\kappa_g^2)$

$$\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\|^2\leq\frac{64}{\alpha\rho}\left(\frac{2TQ}{w}+V_{1,T}\right)+\frac{297}{\rho^2}\left(\frac{\Delta_{\boldsymbol{\beta}}L_{\boldsymbol{\beta}}}{(1-\nu)}+L_3^2\right)$$
$$+\frac{64\Delta_h}{\alpha\rho}+\frac{1188L_{\boldsymbol{\beta}}C_{\mu_g}}{\rho^2(1-\nu)}H_{2,T},\tag{44}$$

which dividing by T and recalling we imposed regularity constraints of $H_{2,T}=o(T)$, as well as $V_{1,T}=o(T)$, implies the bilevel local regret of our OBBO algorithm is sublinear on the order of

$$BLR_w(T):=\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\|^2\leq O\left(\frac{T}{w}+V_{1,T}+\kappa_g^2H_{2,T}\right).\tag{45}$$

■

C.2. Stochastic Setting

The next Lemma upper bounds the expected cumulative difference between the time-smoothed outer level objective $F_{t,w}(\boldsymbol{\lambda})$ evaluated at $\boldsymbol{\lambda}_t$ and $\boldsymbol{\lambda}_{t+1}$ in terms of the outer level objective upper bound m, window size w, and a comparator sequence on subsequent function evaluations $V_{1,T}$.

Lemma 20 *Suppose Assumption E. If our SOBBO algorithm in Algorithm 2 is applied with window size $w\geq 1$ to generate the sequence $\{\boldsymbol{\lambda}_t\}_{t=1}^T$, then we have the upper bound in expectation of*

$$\sum_{t=1}^{T}\left(F_{t,w}(\boldsymbol{\lambda}_t)-F_{t,w}(\boldsymbol{\lambda}_{t+1})\right)\leq\frac{2TQ}{w}+V_{1,T}.$$

where $V_{1,T}:=\sum_{t=1}^{T}\sup_{\boldsymbol{\lambda}\in\mathcal{X}}\left[F_{t+1}\left(\boldsymbol{\lambda}\right)-F_t\left(\boldsymbol{\lambda}\right)\right]$.

Proof By definition in the stochastic setting, we have $F_t(\boldsymbol{\lambda})\triangleq\mathbb{E}_{\epsilon}\left[f_t(\boldsymbol{\lambda},\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}),\epsilon)\right]$. Then it holds, with the linearity of expectation that

$$\sum_{t=1}^{T}\left(F_{t,w}(\boldsymbol{\lambda}_t)-F_{t,w}(\boldsymbol{\lambda}_{t+1})\right)=\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\left(F_{t-i}(\boldsymbol{\lambda}_{t-i})-F_{t-i}(\boldsymbol{\lambda}_{t+1-i})\right)$$
$$=\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\left(\mathbb{E}_{\epsilon}\left[f_{t-i}\left(\boldsymbol{\lambda}_{t-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-i}),\epsilon\right)\right]-\mathbb{E}_{\epsilon}\left[f_{t-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t+1-i}),\epsilon\right)\right]\right)$$
$$=\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\mathbb{E}_{\epsilon}\left[f_{t-i}\left(\boldsymbol{\lambda}_{t-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-i}),\epsilon\right)-f_{t-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t+1-i}),\epsilon\right)\right]$$

Which with the linearity of expectation is equivalent to

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\mathbb{E}_\epsilon\left[f_{t-i}\left(\boldsymbol{\lambda}_{t-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-i}),\epsilon\right)-f_{t-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t+1-i}),\epsilon\right)\right]\\
&=\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\mathbb{E}_\epsilon\left[f_{t-i}\left(\boldsymbol{\lambda}_{t-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-i}),\epsilon\right)-f_{t+1-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t+1-i}(\boldsymbol{\lambda}_{t+1-i}),\epsilon\right)\right] \qquad (46)\\
&+\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\mathbb{E}_\epsilon\left[f_{t+1-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t+1-i}(\boldsymbol{\lambda}_{t+1-i}),\epsilon\right)-f_{t-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t+1-i}),\epsilon\right)\right] \qquad (47)
\end{aligned}$$

For (46), with linearity of expectation, we have

$$\begin{aligned}
&\frac{1}{w}\sum_{i=0}^{w-1}\mathbb{E}_\epsilon\left[f_{t-i}\left(\boldsymbol{\lambda}_{t-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-i}),\epsilon\right)-f_{t+1-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t+1-i}(\boldsymbol{\lambda}_{t+1-i}),\epsilon\right)\right]\\
&=\frac{1}{w}\mathbb{E}_\epsilon\left[f_t\left(\boldsymbol{\lambda}_t,\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t),\epsilon\right)+\ldots+f_{t+1-w}\left(\boldsymbol{\lambda}_{t+1-w},\widehat{\boldsymbol{\beta}}_{t+1-w}(\boldsymbol{\lambda}_{t+1-w}),\epsilon\right)\right]\\
&-\frac{1}{w}\mathbb{E}_\epsilon\left[f_{t+1}\left(\boldsymbol{\lambda}_{t+1},\widehat{\boldsymbol{\beta}}_{t+1}(\boldsymbol{\lambda}_{t+1})\right)+\ldots+f_{t+2-w}\left(\boldsymbol{\lambda}_{t+2-w},\widehat{\boldsymbol{\beta}}_{t+2-w}(\boldsymbol{\lambda}_{t+2-w}),\epsilon\right)\right]\\
&=\frac{1}{w}\mathbb{E}_\epsilon\left[f_{t+1-w}\left(\boldsymbol{\lambda}_{t+1-w},\widehat{\boldsymbol{\beta}}_{t+1-w}(\boldsymbol{\lambda}_{t+1-w}),\epsilon\right)-f_{t+1}\left(\boldsymbol{\lambda}_{t+1},\widehat{\boldsymbol{\beta}}_{t+1}(\boldsymbol{\lambda}_{t+1}),\epsilon\right)\right]\\
&=\frac{1}{w}\left(F_{t+1-w}(\boldsymbol{\lambda}_{t+1-w})-F_{t+1}(\boldsymbol{\lambda}_{t+1})\right)\leq\frac{2Q}{w}, \qquad (48)
\end{aligned}$$

where the last inequality comes from Assumption E. Note (47) can be bounded through

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\mathbb{E}_\epsilon\left[f_{t+1-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t+1-i}(\boldsymbol{\lambda}_{t+1-i}),\epsilon\right)-f_{t-i}\left(\boldsymbol{\lambda}_{t+1-i},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t+1-i}),\epsilon\right)\right]\\
&\leq\sum_{t=1}^{T}\frac{1}{w}\sum_{i=0}^{w-1}\sup_{\boldsymbol{\lambda}}\mathbb{E}_\epsilon\left[f_{t+1-i}\left(\boldsymbol{\lambda},\widehat{\boldsymbol{\beta}}_{t+1-i}(\boldsymbol{\lambda}),\epsilon\right)-f_{t-i}\left(\boldsymbol{\lambda},\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}),\epsilon\right)\right]\\
&=\sum_{t=1}^{T}\sup_{\boldsymbol{\lambda}\in\mathcal{X}}\left[F_{t+1}\left(\boldsymbol{\lambda}\right)-F_t\left(\boldsymbol{\lambda}\right)\right]:=V_{1,T} \qquad (49)
\end{aligned}$$

Combining (47) and (49) results in the upper bound of

$$\sum_{t=1}^{T}\left(F_{t,w}(\boldsymbol{\lambda}_t)-F_{t,w}(\boldsymbol{\lambda}_{t+1})\right)\leq\frac{2TQ}{w}+V_{1,T}.$$

■

The next Lemma provides an upper bound on the expected error of $\mathbb{E}\left[\left\|\boldsymbol{\beta}_t-\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2\right]$ for all $t\in[1,T]$ in terms of an expected initial error, the expected cumulative differences of the outer level variable, the expected cumulative differences of the optimal inner level variables, and a variance term arising from the stochasticity of $g_t(\boldsymbol{\lambda},\boldsymbol{\beta},\zeta)$.

Lemma 21 *Suppose Assumptions A, B, and C. Choose the inner step size of η and the inner iteration count K as*

$$0 < \eta \leq \frac{2}{\ell_{g,1} + \mu_g}, \text{ and } \quad K \geq 1,$$

and define the decay parameter ν, the inner level variable error constant C_{μ_g}, the initial error $\Delta_{\boldsymbol{\beta}}$, and the inner level variable error variance C_K respectively as

$$\nu := \left(1 - \frac{\eta \ell_{g,1} \mu_g}{\ell_{g,1} + \mu_g}\right)\left(1 - \frac{2\eta \ell_{g,1} \mu_g}{\ell_{g,1} + \mu_g}\right)^{K-1}, \quad C_{\mu_g} := \left(1 + \frac{\ell_{g,1} + \mu_g}{\eta \ell_{g,1} \mu_g}\right),$$
$$\Delta_{\boldsymbol{\beta}} := \left\|\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_1(\boldsymbol{\lambda}_1)\right\|^2, \quad \text{and} \quad C_K := \sum_{k=1}^{K}\left(1 - \frac{2\eta \ell_{g,1} \mu_g}{\ell_{g,1} + \mu_g}\right)^k. \tag{50}$$

Then we have $\forall t \in [1, T]$,

$$\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\left\|\boldsymbol{\beta}_{t+1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2\right] \leq \nu^{t-1}\Delta_{\boldsymbol{\beta}} + 2C_{\mu_g}\kappa_g^2 \sum_{j=0}^{t-2} \nu^{j+1}\left[\|\boldsymbol{\lambda}_{t-1-j} - \boldsymbol{\lambda}_{t-j}\|^2\right]$$
$$+2C_{\mu_g}\sum_{j=0}^{t-2}\nu^{j+1}\left[\left\|\widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j}) - \widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j})\right\|^2\right] + \frac{C_K \eta^2 \sigma_{g_{\boldsymbol{\beta}}}^2}{s}\sum_{j=0}^{t-2}\nu^j.$$

Proof Note $\forall k \in [1, K]$ the following expansion holds

$$\begin{aligned}
&\left\|\boldsymbol{\omega}_t^k - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 \\
&= \left\|\boldsymbol{\omega}_t^k - \boldsymbol{\omega}_t^{k-1}\right\|^2 + 2\left\langle \boldsymbol{\omega}_t^k - \boldsymbol{\omega}_t^{k-1}, \boldsymbol{\omega}_t^{k-1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\rangle + \left\|\boldsymbol{\omega}_t^{k-1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2 \\
&= \eta^2\left\|\nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1}, \bar{\zeta}_{t,k})\right\|^2 - 2\eta\left\langle \nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1}, \bar{\zeta}_{t,k}), \boldsymbol{\omega}_t^{k-1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\rangle \\
&\quad + \left\|\boldsymbol{\omega}_t^{k-1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2.
\end{aligned}$$

Using the definition of variance of

$$\begin{aligned}
&VAR_{\bar{\zeta}_{t,k}}\left[\left\|\nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1}, \bar{\zeta}_{t,k})\right\|\right] \\
&= \mathbb{E}_{\bar{\zeta}_{t,k}}\left[\left\|\nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1}, \bar{\zeta}_{t,k})\right\|^2\right] - \mathbb{E}_{\bar{\zeta}_{t,k}}\left[\left\|\nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1}, \bar{\zeta}_{t,k})\right\|\right]^2,
\end{aligned}$$

and conditioning on $\boldsymbol{\omega}_t^{k-1}$, we take expectation to provide the upper bound of

$$\mathbb{E}_{\bar{\zeta}_{t,k}}\left[\left\|\boldsymbol{\omega}_t^k - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2\right] \leq \eta^2\left(\frac{\sigma_{g_{\boldsymbol{\beta}}}^2}{s} + \left\|\nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1})\right\|^2\right)$$
$$-2\eta\left\langle \nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1}), \boldsymbol{\omega}_t^{k-1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\rangle + \left\|\boldsymbol{\omega}_t^{k-1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2. \tag{51}$$

The above upper bound is deterministic, and as such we can utilize the μ_g-strong convexity of g_t to bound

$$\begin{aligned} -2\eta \left\langle \nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1}), \boldsymbol{\omega}_t^{k-1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\rangle \\ \leq -2\eta \left(\frac{\ell_{g,1}\mu_g}{\ell_{g,1}+\mu_g} \left\| \boldsymbol{\omega}_t^{k-1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 + \frac{1}{\ell_{g,1}+\mu_g} \left\| \nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1}) \right\|^2 \right), \end{aligned}$$

which we can substitute in (51) to get

$$\begin{aligned} \mathbb{E}_{\bar{\zeta}_{t,k}} \left[\left\| \boldsymbol{\omega}_t^k - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \right] \leq \frac{\eta^2 \sigma_{g\boldsymbol{\beta}}^2}{s} - \eta \left(\frac{2}{\ell_{g,1}+\mu_g} - \eta \right) \left\| \nabla_{\boldsymbol{\omega}} g_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}_t^{k-1}) \right\|^2 \\ + \left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1}+\mu_g} \right) \left\| \boldsymbol{\omega}_t^{k-1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2. \end{aligned}$$

As $\eta \leq \frac{2}{\ell_{g,1}+\mu_g}$ this provides the upper bound to (51) of

$$\mathbb{E}_{\bar{\zeta}_{t,k}} \left[\left\| \boldsymbol{\omega}_t^k - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \right] \leq \left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1}+\mu_g} \right) \left\| \boldsymbol{\omega}_t^{k-1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 + \frac{\eta^2 \sigma_{g\boldsymbol{\beta}}^2}{s}.$$

This can be unrolled, through iterative conditioning, from $k = K, \ldots, 1$

$$\mathbb{E}_{\bar{\zeta}_{t,K+1}} \left[\left\| \boldsymbol{\omega}_t^K - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \right] \leq \left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1}+\mu_g} \right)^K \mathbb{E}_{\bar{\zeta}_{t,1}} \left\| \boldsymbol{\omega}_t^0 - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 + \frac{C_K \eta^2 \sigma_{g\boldsymbol{\beta}}^2}{s},$$

for $C_K := \sum_{k=1}^K \left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1}+\mu_g} \right)^k$. By definition of $\boldsymbol{\beta}_{t+1} = \boldsymbol{\omega}_t^K$ and $\boldsymbol{\omega}_t^0 = \boldsymbol{\beta}_t$ gives us

$$\mathbb{E}_{\bar{\zeta}_{t,K+1}} \left[\left\| \boldsymbol{\beta}_{t+1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \right] \leq \left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1}+\mu_g} \right)^K \mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 + \frac{C_K \eta^2 \sigma_{g\boldsymbol{\beta}}^2}{s}.$$

Note we can decompose

$$\mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 = \mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) + \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2,$$

which can be expanded based on Young's Inequality and the linearity of expectation for any $\delta > 0$ as

$$\begin{aligned} \mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) + \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \\ \leq (1+\delta) \mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2 \\ + \left(1 + \frac{1}{\delta} \right) \mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2. \end{aligned} \tag{52}$$

Now it holds through linearity of expectation that

$$\begin{aligned} \mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \leq 2\mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \\ + 2\mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2 \end{aligned} \tag{53}$$

which through Lemma 10 can be further upper bounded with the Lipschitz constant of κ_g as

$$\begin{aligned}
&\mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \\
&\leq 2\kappa_g^2 \mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t \right\|^2 + 2\mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left\| \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2 \\
&= 2\kappa_g^2 \left\| \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t \right\|^2 + 2 \left\| \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2
\end{aligned} \tag{54}$$

where the last line comes from the non-randomness of $\left\| \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t \right\|^2$ and $\left\| \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2$ with respect to $\bar{\zeta}_{t,k}$. Combining (53) and (52), we have $\forall \delta > 0$

$$\begin{aligned}
\mathbb{E}_{\bar{\zeta}_{t,K+1}} \left[\left\| \boldsymbol{\beta}_{t+1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \right] &\leq \left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1} + \mu_g}\right)^K (1+\delta) \mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left[\left\| \boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2 \right] \\
&+2\left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1} + \mu_g}\right)^K \left(1 + \frac{1}{\delta}\right) \kappa_g^2 \left\| \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t \right\|^2 \\
+2\left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1} + \mu_g}\right)^K & \left(1 + \frac{1}{\delta}\right) \left\| \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2 + \frac{C_K \eta^2 \sigma_{g_{\boldsymbol{\beta}}}^2}{s}.
\end{aligned}$$

Now setting $\delta = \frac{\eta\ell_{g,1}\mu_g}{\ell_{g,1}+\mu_g} > 0$ implies the upper bound of

$$(1+\delta)\left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1} + \mu_g}\right)^K < \left(1 - \frac{\eta\ell_{g,1}\mu_g}{\ell_{g,1} + \mu_g}\right)\left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1} + \mu_g}\right)^{K-1} < 1,$$

which defining $\nu := \left(1 - \frac{\eta\mu_g\ell_{g,1}}{\ell_{g,1}+\mu_g}\right)\left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1}+\mu_g}\right)^{K-1}$ and $\delta > 0$ implies

$$\left(1 - \frac{2\eta\ell_{g,1}\mu_g}{\ell_{g,1} + \mu_g}\right)^K < \nu,$$

Using the definition of ν, we get

$$\begin{aligned}
\nu \mathbb{E}_{\bar{\zeta}_{t,K+1}} \left[\left\| \boldsymbol{\beta}_{t+1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \right] &\leq \nu^2 \mathbb{E}_{\bar{\zeta}_{t-1,K+1}} \left[\left\| \boldsymbol{\beta}_t - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2 \right] \\
+2C_{\mu_g}\nu^2\kappa_g^2 \left\| \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t \right\|^2 + 2C_{\mu_g}\nu^2 & \left\| \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1}) \right\|^2 + \frac{\nu C_K \eta^2 \sigma_{g_{\boldsymbol{\beta}}}^2}{s},
\end{aligned}$$

where $C_{\mu_g} = \left(1 + \frac{\ell_{g,1}+\mu_g}{\eta\ell_{g,1}\mu_g}\right)$. Starting at $t = T$, and unrolling to $t = 1$, we can write

$$\begin{aligned}
\mathbb{E}_{\bar{\zeta}_{t,K+1}} \left[\left\| \boldsymbol{\beta}_{t+1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t) \right\|^2 \right] &\leq \nu^{t-1}\Delta_{\boldsymbol{\beta}} + 2C_{\mu_g}\kappa_g^2 \sum_{j=0}^{t-2} \nu^{j+1} \left[\left\| \boldsymbol{\lambda}_{t-1-j} - \boldsymbol{\lambda}_{t-j} \right\|^2 \right] \\
+2C_{\mu_g} \sum_{j=0}^{t-2} \nu^{j+1} & \left[\left\| \widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j}) - \widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j}) \right\|^2 \right] + \frac{C_K \eta^2 \sigma_{g_{\boldsymbol{\beta}}}^2}{s} \sum_{j=0}^{t-2} \nu^j.
\end{aligned}$$

■

The next Lemma utilizes Lemma 10 and Lemma 21 to derive an upper bound on the expected hypergradient error $\forall t \in [1, T]$ with respect to $\bar{\zeta}_{t,k}$ in terms of discounted variations of the (i) cumulative time-smoothed hypergradient error; (ii) bilevel local regret; and (iii) cumulative difference between optimal inner-level variables. There is a term composed of a discounted initial error and smoothness term of the inner objective, as well as an additional term arising from the variance of the stochastic gradients of $g_t(\boldsymbol{\lambda}, \boldsymbol{\beta}, \zeta)$.

Lemma 22 *Suppose Assumptions A, B, C, D, and F. Choose the inner step size of η and inner iteration count K as*

$$0 < \eta \leq \frac{2}{\ell_{g,1} + \mu_g}, \text{ and } \quad K \geq 1.$$

With the definitions of ν, C_{μ_g}, $\Delta_{\boldsymbol{\beta}}$, and C_K from Lemma 21, the expected hypergradient error can be bounded as

$$\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\left\|\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}) - \nabla F_t\left(\boldsymbol{\lambda}_t\right)\right\|^2\right] \leq \delta_t + A \sum_{j=0}^{t-2} \nu^{j+1} \left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha)\right\|^2$$
$$+ B \sum_{j=0}^{t-2} \nu^{j+1} \left\|\widetilde{\nabla} f_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\beta}_{t-j}, \mathcal{Z}_{t-1-j,w}) - \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j})\right\|^2$$
$$+ C \sum_{j=0}^{t-2} \nu^{j+1} \left[\left\|\widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j}) - \widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j})\right\|^2\right] + \frac{D \sigma_{g_{\boldsymbol{\beta}}}^2}{s}.$$

where $\delta_t = \kappa_g^2 \nu^{t-1} \Delta_{\boldsymbol{\beta}}$ and $A = 4 C_{\mu_g} \kappa_g^4 \alpha^2$, $B = \frac{4 C_{\mu_g} \kappa_g^4 \alpha^2}{\rho^2}$, $C = 2 C_{\mu_g} \kappa_g^2$, and $D = C_K \kappa_g^2 \eta^2 \sum_{j=0}^{t-2} \nu^j$.

Proof First, from Lemma 10 we have that $\forall \boldsymbol{\lambda} \in \mathcal{X}$ and $\boldsymbol{\beta} \in \mathbb{R}^{d_2}$

$$\left\|\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}) - \nabla F_t\left(\boldsymbol{\lambda}_t\right)\right\|^2 \leq \kappa_g^2 \left\|\boldsymbol{\beta}_{t+1} - \widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_t)\right\|^2. \tag{55}$$

Taking expectation of (55) with respect to $\bar{\zeta}_{t,K+1}$ and substituting the upper bound of Lemma 21, note

$$\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\left\|\widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}) - \nabla F_t\left(\boldsymbol{\lambda}_t\right)\right\|^2\right]$$
$$\leq \kappa_g^2 \left(\nu^{t-1} \Delta_{\boldsymbol{\beta}} + 2 C_{\mu_g} \kappa_g^2 \sum_{j=0}^{t-2} \nu^{j+1} \left\|\boldsymbol{\lambda}_{t-1-j} - \boldsymbol{\lambda}_{t-j}\right\|^2\right) \tag{56}$$
$$+ \kappa_g^2 \left(2 C_{\mu_g} \sum_{j=0}^{t-2} \nu^{j+1} \left\|\widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j}) - \widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j})\right\|^2 + \frac{C_K \eta^2 \sigma_{g_{\boldsymbol{\beta}}}^2}{s} \sum_{j=0}^{t-2} \nu^j\right). \tag{57}$$

Focusing on the second term of (56) we see by definition

$$\sum_{j=0}^{t-2} \nu^{j+1} \left\|\boldsymbol{\lambda}_{t-1-j} - \boldsymbol{\lambda}_{t-j}\right\|^2$$
$$= \sum_{j=0}^{t-2} \nu^{j+1} \alpha^2 \left\| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \widetilde{\nabla} f_{t-1-j,\boldsymbol{w}}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\beta}_{t-j}, \mathcal{Z}_{t-1-j,w}), \alpha) \right\|^2. \tag{58}$$

Using Lemma 9 we have $\forall j \in [0, t-2]$

$$\left\| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \widetilde{\nabla} f_{t-1-j,\boldsymbol{w}}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\beta}_{t-j}, \mathcal{Z}_{t-1-j,w}), \alpha) \right\|^2 \leq 2 \left\| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha) \right\|^2$$
$$+ 2 \left\| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha) - \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \widetilde{\nabla} f_{t-1-j,\boldsymbol{w}}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\beta}_{t-j}, \mathcal{Z}_{t-1-j,w}), \alpha) \right\|^2$$
$$\leq 2 \left\| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha) \right\|^2$$
$$+ \frac{2}{\rho^2} \left\| \widetilde{\nabla} f_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\beta}_{t-j}, \mathcal{Z}_{t-1-j,w}) - \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}) \right\|^2.$$

We can write an upper bound to (58) as

$$\sum_{j=0}^{t-2} \nu^{j+1} \left\|\boldsymbol{\lambda}_{t-1-j} - \boldsymbol{\lambda}_{t-j}\right\|^2 \leq 2\alpha^2 \sum_{j=0}^{t-2} \nu^{j+1} \left\| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha) \right\|^2$$
$$+ \frac{2\alpha^2}{\rho^2} \sum_{j=0}^{t-2} \nu^{j+1} \left(\left\| \widetilde{\nabla} f_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\beta}_{t-j}, \mathcal{Z}_{t-1-j,w}) - \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}) \right\|^2 \right). \tag{59}$$

Using (59), we get

$$\mathbb{E}_{\bar{\zeta}_{t,K+1}} \left[\left\| \widetilde{\nabla} f_t(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}) - \nabla F_t\left(\boldsymbol{\lambda}_t\right) \right\|^2 \right] \leq \delta_t + A \sum_{j=0}^{t-2} \nu^{j+1} \left\| \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-1-j}, \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}), \alpha) \right\|^2$$
$$+ B \sum_{j=0}^{t-2} \nu^{j+1} \left\| \widetilde{\nabla} f_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}, \boldsymbol{\beta}_{t-j}, \mathcal{Z}_{t-1-j,w}) - \nabla F_{t-1-j,w}(\boldsymbol{\lambda}_{t-1-j}) \right\|^2$$
$$+ C \sum_{j=0}^{t-2} \nu^{j+1} \left[\left\| \widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j}) - \widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j}) \right\|^2 \right] + \frac{D \sigma_{g_{\boldsymbol{\beta}}}^2}{s}.$$

where $\delta_t = \kappa_g^2 \nu^{t-1} \Delta_{\boldsymbol{\beta}}$ and $A = 4C_{\mu_g} \kappa_g^4 \alpha^2$, $B = \frac{4C_{\mu_g} \kappa_g^4 \alpha^2}{\rho^2}$, $C = 2C_{\mu_g} \kappa_g^2$, and $D = C_K \kappa_g^2 \eta^2 \sum_{j=0}^{t-2} \nu^j$. ■

Lemma 23 provides an upper bound on the expected cumulative time-smoothed hypergradient error in terms of an initial error, expected bilevel local regret, expected cumulative differences of optimal inner level variables, as well as variance terms from the stochastic approximated gradients.

Lemma 23 *Suppose Assumptions A, B, C, D, and F. Choose the inner step size of η, the inner iteration count K, and the outer step size α respectively as*

$$0 < \eta \leq \frac{2}{\ell_{g,1} + \mu_g}, \quad K \geq 1, \text{ and } \quad \alpha < \frac{\rho\sqrt{(1-\nu)}}{\kappa_g^2 \sqrt{72 C_{\mu_g}}}.$$

Then $\forall t \in [1, T]$ *the expected cumulative time-smoothed hypergradient error with respect to independent samples* $Z_{t,w}$ *from SOBBO satisfies*

$$\mathbb{E}_{Z_{t,w}}\left[\sum_{t=1}^{T}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2\right] \leq \frac{9T\sigma_f^2}{2w} + \frac{9T\ell_{f,1}^2\kappa_g^2}{2}\left(1 - \frac{\mu_g}{\ell_{g,1}}\right)^{2m}$$
$$+\frac{9\kappa_g^2\Delta_{\boldsymbol{\beta}}}{4(1-\nu)} + \frac{\rho^2}{8}\sum_{t=1}^{T}\mathbb{E}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\right\|^2$$
$$+\frac{9C_{\mu_g}\kappa_g^2}{2(1-\nu)}\sum_{t=2}^{T}\mathbb{E}\left[\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2\right] + \frac{9TC_K\kappa_g^2\eta^2\sigma_{g_{\boldsymbol{\beta}}}^2}{4s(1-\nu)}.$$

Proof With the linearity of expectation and by definition of (7) we have

$$\mathbb{E}_{Z_{t,w}}\left[\sum_{t=1}^{T}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2\right]$$
$$= \sum_{t=1}^{T}\mathbb{E}_{Z_{t,w}}\left[\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2\right]$$
$$= \frac{1}{w^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\left\|\sum_{i=0}^{w-1}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\beta}_{t+1-i}, \mathcal{E}_{t-i}) - \nabla F_{t-i}\left(\boldsymbol{\lambda}_{t-i}\right)\right]\right\|^2\right]. \tag{60}$$

Note that we can upper bound (60) as

$$\frac{1}{w^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\left\|\sum_{i=0}^{w-1}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\beta}_{t+1-i}, \mathcal{E}_{t-i}) - \nabla F_{t-i}\left(\boldsymbol{\lambda}_{t-i}\right)\right]\right\|^2\right]$$
$$\leq \frac{2}{w^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\left\|\sum_{i=0}^{w-1}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\beta}_{t+1-i}, \mathcal{E}_{t-i}) - \mathbb{E}_{\mathcal{E}_{t-i}}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\beta}_{t+1-i}, \mathcal{E}_{t-i})\right]\right]\right\|^2\right] \tag{61}$$
$$+\frac{2}{w^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\left\|\sum_{i=0}^{w-1}\left[\mathbb{E}_{\mathcal{E}_{t-i}}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\beta}_{t+1-i}, \mathcal{E}_{t-i})\right] - \nabla F_{t-i}\left(\boldsymbol{\lambda}_{t-i}\right)\right]\right\|^2\right]. \tag{62}$$

The linearity of expectation, definition of variance, and independence of $Z_{t,w} := \{\mathcal{E}_{t-i}\}_{i=0}^{w-1}$ $\forall t \in [1, T]$ implies for $y_i = \widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\beta}_{t+1-i}, \mathcal{E}_{t-i})$ with finite variance σ_f^2, we have

$$\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\left\|\sum_{i=0}^{w-1}\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\beta}_{t+1-i}, \mathcal{E}_{t-i}) - \mathbb{E}_{\mathcal{E}_{t-i}}\left[\sum_{i=0}^{w-1}\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i}, \boldsymbol{\beta}_{t+1-i}, \mathcal{E}_{t-i})\right]\right\|^2\right]$$
$$\leq \sum_{i=0}^{w-1}\sigma_f^2 = w\sigma_f^2. \tag{63}$$

Expanding (62) we have

$$\frac{2}{w^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\left\|\sum_{i=0}^{w-1}\left[\mathbb{E}_{\mathcal{E}_{t-i}}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i},\mathcal{E}_{t-i})\right]-\nabla F_{t-i}\left(\boldsymbol{\lambda}_{t-i}\right)\right]\right\|^2\right]$$

$$\leq \frac{4}{w^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\left\|\sum_{i=0}^{w-1}\left[\mathbb{E}_{\mathcal{E}_{t-i}}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i},\mathcal{E}_{t-i})\right]-\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i})\right]\right\|^2\right] \tag{64}$$

$$+\frac{4}{w^2}\sum_{t=1}^{T}\left[\left\|\sum_{i=0}^{w-1}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i})-\nabla F_{t-i}\left(\boldsymbol{\lambda}_{t-i}\right)\right]\right\|^2\right] \tag{65}$$

Utilizing Lemmas 6 and 14for (64) gives us the expected stochastic gradient bias

$$\frac{4}{w^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\left\|\sum_{i=0}^{w-1}\left[\mathbb{E}_{\mathcal{E}_{t-i}}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i},\mathcal{E}_{t-i})\right]-\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i})\right]\right\|^2\right]$$

$$\leq \frac{4}{w^2}\sum_{t=1}^{T}\mathbb{E}_{\mathcal{Z}_{t,w}}\left[w\sum_{i=0}^{w-1}\left\|\mathbb{E}_{\mathcal{E}_{t-i}}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i},\mathcal{E}_{t-i})\right]-\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i})\right\|^2\right]$$

$$\leq \frac{4}{w^2}\sum_{t=1}^{T}\left(w^2\ell_{f,1}^2\kappa_g^2\left(1-\frac{\mu_g}{\ell_{g,1}}\right)^{2m}\right)=4T\ell_{f,1}^2\kappa_g^2\left(1-\frac{\mu_g}{\ell_{g,1}}\right)^{2m} \tag{66}$$

Applying Lemma 6 with linearity of expectation to (65) results in

$$\frac{4}{w^2}\sum_{t=1}^{T}\left[\left\|\sum_{i=0}^{w-1}\left[\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i})-\nabla F_{t-i}\left(\boldsymbol{\lambda}_{t-i}\right)\right]\right\|^2\right]$$

$$\leq \frac{4}{w^2}\sum_{t=1}^{T}w\sum_{i=0}^{w-1}\left[\left\|\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i})-\nabla F_{t-i}\left(\boldsymbol{\lambda}_{t-i}\right)\right\|^2\right]$$

$$=\frac{4}{w}\sum_{t=1}^{T}\sum_{i=0}^{w-1}\left[\left\|\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i})-\nabla F_{t-i}\left(\boldsymbol{\lambda}_{t-i}\right)\right\|^2\right] \tag{67}$$

Combining (61), (62), (63), and (67), we have the upper bound of

$$\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\sum_{t=1}^{T}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_{t},\boldsymbol{\beta}_{t+1},\mathcal{Z}_{t,w})-\nabla F_{t,w}\left(\boldsymbol{\lambda}_{t}\right)\right\|^2\right]\leq\frac{4T\sigma_f^2}{w}$$

$$+4T\ell_{f,1}^2\kappa_g^2\left(1-\frac{\mu_g}{\ell_{g,1}}\right)^{2m}+\frac{4}{w}\sum_{t=1}^{T}\sum_{i=0}^{w-1}\left[\left\|\widetilde{\nabla} f_{t-i}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i})-\nabla F_{t-i}\left(\boldsymbol{\lambda}_{t-i}\right)\right\|^2\right], \tag{68}$$

Taking expectation with respect to $\bar{\zeta}_{t,K+1}$, we utilize the upper bound from Lemma 22. By iterative conditioning and re-indexing the expected cumulative hypergradient error as well as dropping expectation for non-random quantities, we derive an upper bound on (68) as

$$
\begin{aligned}
&\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\sum_{t=1}^{T}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\beta}_{t+1},\mathcal{Z}_{t,w})-\nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2\right]\right] \\
&\leq \frac{4T\sigma_f^2}{w}+4T\ell_{f,1}^2\kappa_g^2\left(1-\frac{\mu_g}{\ell_{g,1}}\right)^{2m}+\frac{2}{w}\sum_{t=1}^{T}\sum_{i=0}^{w-1}\left(\kappa_g^2\nu^{t-i-1}\Delta_{\boldsymbol{\beta}}\right) \\
&+\frac{2}{w}\sum_{t=1}^{T}\sum_{i=0}^{w-1}\left(4C_{\mu_g}\kappa_g^4\alpha^2\sum_{j=0}^{t-i-2}\nu^{j+1}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-i-j},\nabla F_{t-i-j,w}(\boldsymbol{\lambda}_{t-i-j}),\alpha)\right\|^2\right) \\
&+\frac{2}{w}\sum_{t=1}^{T}\sum_{i=0}^{w-1}\left(\frac{4C_{\mu_g}\kappa_g^4\alpha^2}{\rho^2}\sum_{j=0}^{t-i-2}\nu^{j+1}A_{t-i,j}\right) \\
&+\frac{2}{w}\sum_{t=2}^{T}\sum_{i=0}^{w-1}\left(2C_{\mu_g}\kappa_g^2\sum_{j=0}^{t-i-2}\nu^{j+1}B_{t-i,j}\right)+\frac{2}{w}\sum_{t=1}^{T}\sum_{i=0}^{w-1}\left(\frac{C_K\kappa_g^2\eta^2\sigma_{g_{\boldsymbol{\beta}}}^2}{s}\sum_{j=0}^{t-i-2}\nu^j\right),
\end{aligned}
\tag{69}
$$

where

$$
\begin{aligned}
A_{t,j}&:=\mathbb{E}_{\bar{\zeta}_{t-j,K+1}}\left[\mathbb{E}_{\mathcal{Z}_{t-j,w}}\left[\left\|\widetilde{\nabla} f_{t-j,w}(\boldsymbol{\lambda}_{t-j},\boldsymbol{\beta}_{t+1-j},\mathcal{Z}_{t-j,w})-\nabla F_{t-j,w}(\boldsymbol{\lambda}_{t-j})\right\|^2\right]\right] \\
B_{t,j}&:=\left\|\widehat{\boldsymbol{\beta}}_{t-j}(\boldsymbol{\lambda}_{t-1-j})-\widehat{\boldsymbol{\beta}}_{t-1-j}(\boldsymbol{\lambda}_{t-1-j})\right\|^2.
\end{aligned}
$$

Given $\nu<1$, it holds that $\sum_{j=0}^{t-2}\nu^j<\sum_{j=0}^{\infty}\nu^j=\frac{1}{1-\nu}$, which lets us upper bound (69) as

$$
\begin{aligned}
&\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\sum_{t=1}^{T}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\beta}_{t+1},\mathcal{Z}_{t,w})-\nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2\right]\right] \leq \frac{4T\sigma_f^2}{w}+4T\ell_{f,1}^2\kappa_g^2\left(1-\frac{\mu_g}{\ell_{g,1}}\right)^{2m} \\
&+\frac{2}{w}\sum_{t=1}^{T}\sum_{i=0}^{w-1}\left(\kappa_g^2\nu^{t-i-1}\Delta_{\boldsymbol{\beta}}+\frac{4C_{\mu_g}\kappa_g^4\alpha^2}{1-\nu}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_{t-i},\nabla F_{t-i,w}(\boldsymbol{\lambda}_{t-i}),\alpha)\right\|^2\right) \\
&+\frac{2}{w}\sum_{t=1}^{T}\sum_{i=0}^{w-1}\left(\frac{4C_{\mu_g}\kappa_g^4\alpha^2}{(1-\nu)\rho^2}\mathbb{E}_{\bar{\zeta}_{t-i,K+1}}\left[\mathbb{E}_{\mathcal{Z}_{t-i,w}}\left[\left\|\widetilde{\nabla} f_{t-i,w}(\boldsymbol{\lambda}_{t-i},\boldsymbol{\beta}_{t+1-i},\mathcal{Z}_{t-i,w})-\nabla F_{t-i,w}(\boldsymbol{\lambda}_{t-i})\right\|^2\right]\right]\right) \\
&+\frac{2}{w}\sum_{t=2}^{T}\sum_{i=0}^{w-1}\left(\frac{2C_{\mu_g}\kappa_g^2}{1-\nu}\left\|\widehat{\boldsymbol{\beta}}_{t-i}(\boldsymbol{\lambda}_{t-1-i})-\widehat{\boldsymbol{\beta}}_{t-1-i}(\boldsymbol{\lambda}_{t-1-i})\right\|^2\right)+\frac{2}{w}\sum_{t=1}^{T}\sum_{i=0}^{w-1}\left(\frac{C_K\kappa_g^2\eta^2\sigma_{g_{\boldsymbol{\beta}}}^2}{(1-\nu)s}\right).
\end{aligned}
\tag{70}
$$

Next we derive the upper bound of (70) as

$$\begin{aligned}
\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\sum_{t=1}^{T}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\beta}_{t+1},\mathcal{Z}_{t,w})-\nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2\right]\right] &\le \frac{4T\sigma_f^2}{w}+4T\ell_{f,1}^2\kappa_g^2\left(1-\frac{\mu_g}{\ell_{g,1}}\right)^{2m} \\
&\quad+\frac{2\kappa_g^2\Delta_{\boldsymbol{\beta}}}{(1-\nu)}+\frac{8C_{\mu_g}\kappa_g^4\alpha^2}{(1-\nu)}\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\|^2 \\
&\quad+\frac{8C_{\mu_g}\kappa_g^4\alpha^2}{\rho^2(1-\nu)}\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\sum_{t=1}^{T}\left\|\widetilde{\nabla} f_{t,w}\left(\boldsymbol{\lambda}_t,\boldsymbol{\beta}_{t+1},\mathcal{Z}_{t,w}\right)-\nabla F_{t,w}(\boldsymbol{\lambda}_t)\right\|^2\right]\right] \\
&\quad+\frac{4C_{\mu_g}\kappa_g^2}{(1-\nu)}\sum_{t=2}^{T}\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1})-\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2+\frac{2TC_K\kappa_g^2\eta^2\sigma_{g_{\boldsymbol{\beta}}}^2}{s(1-\nu)}.
\end{aligned}$$

which implies through linearity of expectation that

$$\begin{aligned}
&\left(1-\frac{8C_{\mu_g}\kappa_g^4\alpha^2}{\rho^2(1-\nu)}\right)\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\sum_{t=1}^{T}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\beta}_{t+1},\mathcal{Z}_{t,w})-\nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2\right]\right] \\
&\le \frac{4T\sigma_f^2}{w}+4T\ell_{f,1}^2\kappa_g^2\left(1-\frac{\mu_g}{\ell_{g,1}}\right)^{2m}+\frac{2\kappa_g^2\Delta_{\boldsymbol{\beta}}}{(1-\nu)}+\frac{8C_{\mu_g}\kappa_g^4\alpha^2}{(1-\nu)}\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\|^2 \\
&\quad+\frac{4C_{\mu_g}\kappa_g^2}{(1-\nu)}\sum_{t=2}^{T}\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1})-\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2+\frac{2TC_K\kappa_g^2\eta^2\sigma_{g_{\boldsymbol{\beta}}}^2}{s(1-\nu)}.
\end{aligned}$$

As $0<\alpha\le\frac{\rho\sqrt{(1-\nu)}}{\kappa_g^2\sqrt{72C_{\mu_g}}}$

$$\left(1-\frac{8C_{\mu_g}\kappa_g^4}{\rho^2(1-\nu)}\right)\ge\frac{8}{9},$$

we have the upper bound of

$$\begin{aligned}
&\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\sum_{t=1}^{T}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t,\boldsymbol{\beta}_{t+1},\mathcal{Z}_{t,w})-\nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2\right]\right] \\
&\le \frac{9T\sigma_f^2}{2w}+\frac{9T\ell_{f,1}^2\kappa_g^2}{2}\left(1-\frac{\mu_g}{\ell_{g,1}}\right)^{2m}+\frac{9\kappa_g^2\Delta_{\boldsymbol{\beta}}}{4(1-\nu)}+\frac{\rho^2}{8}\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\|^2 \\
&\quad+\frac{9C_{\mu_g}\kappa_g^2}{2(1-\nu)}\sum_{t=2}^{T}\left[\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1})-\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2\right]+\frac{9TC_K\kappa_g^2\eta^2\sigma_{g_{\boldsymbol{\beta}}}^2}{4s(1-\nu)}.
\end{aligned}$$

■

The following theorem presents the proof of SOBBO achieving a sublinear rate of bilevel local regret.

Theorem 24 *Suppose Assumptions A, B, C, D, E, and F. Choose the inner step size of η, the inner iteration count of K, the outer step size of α, and the batch sizes of s and m to respectively satisfy*

$$\eta \leq \frac{2}{\ell_{g,1}+\mu_g}, \quad K \geq 1, \quad \alpha \leq \min\left\{\frac{3\rho}{4\ell_{F,1}}, \frac{\rho\sqrt{(1-\nu)}}{\kappa_g^2\sqrt{72C_{\mu_g}}}\right\},$$
$$s = w, \quad \textit{and} \quad m = \log(w)/\log\left(1-\frac{\mu_g}{\ell_{g,1}}\right)+1. \tag{71}$$

Then the expected bilevel local regret of our SOBBO Algorithm 2 satisfies

$$\begin{aligned} BLR_w(T) &:= \sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\|^2 \\ &\leq O\left(\frac{T}{w}\left(1+\sigma_f^2+\kappa_g^2\sigma_{g_\beta}^2\right)+V_{1,T}+\kappa_g^2 H_{2,T}\right) \end{aligned} \tag{72}$$

which is a sublinear rate of bilevel local regret when the regularity constraints of $V_{1,T}=o(T)$ and $H_{2,T}=o(T)$ are imposed.

Proof Note, with Assumption A we have the upper bound of

$$\begin{aligned} F_{t,w}(\boldsymbol{\lambda}_{t+1}) - F_{t,w}(\boldsymbol{\lambda}_t) &= \frac{1}{w}\sum_{i=0}^{w-1}F_{t-i}(\boldsymbol{\lambda}_{t+1-i}) - \frac{1}{w}\sum_{i=0}^{w-1}F_{t-i}(\boldsymbol{\lambda}_{t-i}) \\ &= \frac{1}{w}\sum_{i=0}^{w-1}\left[F_{t-i}(\boldsymbol{\lambda}_{t+1-i}) - F_{t-i}(\boldsymbol{\lambda}_{t-i})\right] \\ &\leq \frac{1}{w}\sum_{i=0}^{w-1}\left[\langle\nabla F_{t-i}(\boldsymbol{\lambda}_{t-i}), \boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\rangle + \frac{\ell_{F,1}}{2}\|\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\|^2\right] \\ &= \langle\nabla F_{t,w}(\boldsymbol{\lambda}_t), \boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\rangle + \frac{\ell_{F,1}}{2}\|\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\|^2. \end{aligned}$$

Substituting in $\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right) := \frac{1}{\alpha}(\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t+1})$,

$$\begin{aligned} F_{t,w}(\boldsymbol{\lambda}_{t+1}) - F_{t,w}(\boldsymbol{\lambda}_t) &\leq \langle\nabla F_{t,w}(\boldsymbol{\lambda}_t), \boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\rangle + \frac{\ell_{F,1}}{2}\|\boldsymbol{\lambda}_{t+1}-\boldsymbol{\lambda}_t\|^2 \\ &= -\alpha\left\langle\nabla F_{t,w}(\boldsymbol{\lambda}_t), \mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\rangle \\ &\quad + \frac{\alpha^2\ell_{F,1}}{2}\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\|^2, \\ &= -\alpha\left\langle\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\rangle \\ &\quad +\alpha\left\langle\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}(\boldsymbol{\lambda}_t), \mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\rangle \\ &\quad + \frac{\alpha^2\ell_{F,1}}{2}\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\|^2. \end{aligned} \tag{73}$$

With Lemma 8, note for $\boldsymbol{q} = \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w})$

$$\begin{aligned}
&\alpha \left\langle \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\rangle \\
&\geq \alpha\rho \left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\|^2 + h(\boldsymbol{\lambda}_{t+1}) - h(\boldsymbol{\lambda}_t) \qquad (74)
\end{aligned}$$

and further we get the following based on a variation of Young's Inequality

$$\begin{aligned}
&\left\langle \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right), \mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\rangle \\
&\leq \frac{1}{\rho}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2 \\
&\quad + \frac{\rho}{4}\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\|^2. \qquad (75)
\end{aligned}$$

Combining (74) and (75) in (73) we get

$$\begin{aligned}
F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right) - F_{t,w}\left(\boldsymbol{\lambda}_t\right) &\leq \left(\frac{\alpha^2 \ell_{F,1}}{2} - \frac{3\alpha\rho}{4}\right)\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\|^2 \\
&\quad + \frac{\alpha}{\rho}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2 + h(\boldsymbol{\lambda}_t) - h(\boldsymbol{\lambda}_{t+1}),
\end{aligned}$$

which as $0 < \alpha \leq \frac{3\rho}{4\ell_{F,1}}$ results in the further upper bound of

$$\begin{aligned}
F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right) - F_{t,w}\left(\boldsymbol{\lambda}_t\right) &\leq \frac{3\alpha\rho}{8}\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\|^2 \\
&\quad + \frac{\alpha}{\rho}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2 + h(\boldsymbol{\lambda}_t) - h(\boldsymbol{\lambda}_{t+1}). \qquad (76)
\end{aligned}$$

Further, we have

$$\begin{aligned}
\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\right\|^2 &\leq 2\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\|^2 \\
&\quad + 2\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right) - \mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\right\|^2 \\
&\leq 2\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\|^2 + \frac{2}{\rho^2}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2,
\end{aligned}$$

where the last inequality comes from through Lemma 9. Then we have

$$\begin{aligned}
-\left\|\mathcal{G}_{\mathcal{X}}\left(\boldsymbol{\lambda}_t, \widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}), \alpha\right)\right\|^2 &\leq -\frac{1}{2}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\right\|^2 \\
&\quad + \frac{1}{\rho^2}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2. \qquad (77)
\end{aligned}$$

Substituting (77) in (76)

$$\begin{aligned}
F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right) - F_{t,w}\left(\boldsymbol{\lambda}_t\right) &\leq -\frac{3\alpha\rho}{16}\left\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\right\|^2 \\
&\quad + \left(\frac{\alpha}{\rho} + \frac{3\alpha}{8\rho}\right)\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2 + h(\boldsymbol{\lambda}_t) - h(\boldsymbol{\lambda}_{t+1})
\end{aligned}$$

Telescoping $t = 1, \ldots, T$ and taking expectation with respect to $\bar{\zeta}_{t,k}$ and $\mathcal{Z}_{t,w}$ gives us

$$\begin{aligned}\frac{3\alpha\rho}{16}\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\|^2 &\leq \sum_{t=1}^{T}\left(F_{t,w}\left(\boldsymbol{\lambda}_t\right) - F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)\right) \\ &+ \frac{11\alpha}{8\rho}\mathbb{E}_{\bar{\zeta}_{t,K+1}}\left[\mathbb{E}_{\mathcal{Z}_{t,w}}\left[\sum_{t=1}^{T}\left\|\widetilde{\nabla} f_{t,w}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}_{t+1}, \mathcal{Z}_{t,w}) - \nabla F_{t,w}\left(\boldsymbol{\lambda}_t\right)\right\|^2\right]\right] + \Delta_h, \end{aligned} \tag{78}$$

where $\Delta_h := h(\boldsymbol{\lambda}_1) - h(\boldsymbol{\lambda}_{T+1})$. Substituting the result of Lemma 23 in (78)

$$\begin{aligned}\frac{3\alpha\rho}{16}\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\|^2 &\leq \sum_{t=1}^{T}\left(F_{t,w}\left(\boldsymbol{\lambda}_t\right) - F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)\right) + \Delta_h \\ &+ \frac{11\alpha}{8\rho}\left(\frac{9T\sigma_f^2}{2w} + \frac{9T\ell_{f,1}^2\kappa_g^2}{2}\left(1 - \frac{\mu_g}{\ell_{g,1}}\right)^{2m} + \frac{9\kappa_g^2\Delta_{\boldsymbol{\beta}}}{4(1-\nu)} + \frac{9TC_K\kappa_g^2\eta^2\sigma_{g\boldsymbol{\beta}}^2}{4S(1-\nu)}\right) \\ &+ \frac{11\alpha}{8\rho}\left(\frac{\rho^2}{8}\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\|^2 + \frac{9C_{\mu_g}\kappa_g^2}{2(1-\nu)}\sum_{t=2}^{T}\left[\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2\right]\right) \end{aligned} \tag{79}$$

we have to rearrange

$$\begin{aligned}\frac{\alpha\rho}{64}\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\|^2 &\leq \sum_{t=1}^{T}\left(F_{t,w}\left(\boldsymbol{\lambda}_t\right) - F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)\right) + \Delta_h \\ &+ \frac{99\alpha}{32\rho}\left(\frac{2T\sigma_f^2}{w} + 2T\ell_{f,1}^2\kappa_g^2\left(1 - \frac{\mu_g}{\ell_{g,1}}\right)^{2m} + \frac{\kappa_g^2\Delta_{\boldsymbol{\beta}}}{(1-\nu)} + \frac{TC_K\kappa_g^2\eta^2\sigma_{g\boldsymbol{\beta}}^2}{s(1-\nu)}\right) \\ &+ \frac{99\alpha C_{\mu_g}\kappa_g^2}{16\rho(1-\nu)}\sum_{t=2}^{T}\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2 \end{aligned} \tag{80}$$

or more succinctly with the choice of $s = w$ and $m = \log{(w)}/\log{\left(1 - \frac{\mu_g}{\ell_{g,1}}\right)} + 1$, we have

$$\begin{aligned}\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t, \nabla F_{t,w}(\boldsymbol{\lambda}_t), \alpha)\|^2 &\leq \frac{64}{\alpha\rho}\left(\sum_{t=1}^{T}\left(F_{t,w}\left(\boldsymbol{\lambda}_t\right) - F_{t,w}\left(\boldsymbol{\lambda}_{t+1}\right)\right) + \Delta_h\right) \\ &+ \frac{198}{\rho^2}\frac{T}{w}\left(2\sigma_f^2 + 2\ell_{f,1}^2\kappa_g^2 + \frac{C_K\kappa_g^2\eta^2\sigma_{g\boldsymbol{\beta}}^2}{(1-\nu)}\right) + \frac{198}{\rho^2}\frac{\kappa_g^2\Delta_{\boldsymbol{\beta}}}{(1-\nu)} \\ &+ \frac{396C_{\mu_g}\kappa_g^2}{\rho^2(1-\nu)}\sum_{t=2}^{T}\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1}) - \widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2 \end{aligned} \tag{81}$$

Using the result of Lemma 20, we have

$$\sum_{t=1}^{T}\left(F_{t,w}(\boldsymbol{\lambda}_t) - F_{t,w}(\boldsymbol{\lambda}_{t+1})\right) \leq \frac{2TQ}{w} + V_{1,T} \tag{82}$$

or all together

$$\sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\|^2 \le \frac{64}{\alpha\rho}\left(\frac{2TQ}{w}+V_{1,T}+\Delta_h\right) + \frac{198}{\rho^2}\frac{T}{w}\left(2\sigma_f^2+2\ell_{f,1}^2\kappa_g^2+\frac{C_K\kappa_g^2\eta^2\sigma_{g_{\boldsymbol{\beta}}}^2}{(1-\nu)}\right)+\frac{198}{\rho^2}\frac{\kappa_g^2\Delta_{\boldsymbol{\beta}}}{(1-\nu)} + \frac{396C_{\mu_g}\kappa_g^2}{\rho^2(1-\nu)}\sum_{t=2}^{T}\left\|\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}_{t-1})-\widehat{\boldsymbol{\beta}}_{t-1}(\boldsymbol{\lambda}_{t-1})\right\|^2 \tag{83}$$

which dividing by T and recalling we imposed regularity constraints of $H_{2,T}=o(T)$, as well as $V_{1,T}=o(T)$, implies the bilevel local regret of our SOBBO algorithm is sublinear on the order of

$$BLR_w(T) := \sum_{t=1}^{T}\|\mathcal{G}_{\mathcal{X}}(\boldsymbol{\lambda}_t,\nabla F_{t,w}(\boldsymbol{\lambda}_t),\alpha)\|^2 \le O\left(\frac{T}{w}\left(1+\kappa_g^2+\sigma_f^2+\kappa_g^2\sigma_{g_{\boldsymbol{\beta}}}^2\right)+V_{1,T}+\kappa_g^2H_{2,T}\right) \tag{84}$$

■

Appendix D. Bilevel Local Regret Comparison

In this section, we provide a detailed comparison of the bilevel local regret achieved by OBBO relative to online bilevel benchmarks of OAGD ([34]) and SOBOW ([24]).

D.1. OAGD

We provide a restatement of the bilevel local regret presented in Theorem 9 of [34] using our notation.

Theorem 25 *(Theorem 9 in [34]) Suppose Assumptions A, B, D, and E. Then the OAGD algorithm of [34] for an inner step size of $\eta=\frac{2}{\ell_{g,1}+\mu_g}$, inner iteration count of $K=1$, and outer step size of $\alpha\le\min\left\{\frac{1}{8\ell_{F,1}},\frac{1}{2\sqrt{2}L_{\boldsymbol{\beta}}M_f(\kappa_g^2-1)^{1/2}}\right\}$ satisfies*

$$\sum_{t=1}^{T}\|\nabla F_{t,w}(\boldsymbol{\lambda}_t)\|^2 \le \frac{16}{\alpha}\left(\frac{2TQ}{w}+2Q+\ell_{f,0}H_{1,T}\right) + 10M_f^2(\kappa_g-1)^2\left(\frac{\left\|\boldsymbol{\beta}_1-\widehat{\boldsymbol{\beta}}_1(\boldsymbol{\lambda}_1\right\|^2}{2(\kappa_g+1)}+2H_{2,T}\right) = O\left(\frac{T}{w}+H_{1,T}+\kappa_g^4H_{2,T}\right) \tag{85}$$

Following the literature on dynamic regret, [34] considers the setting where the inner level path variation of $H_{1,T}$ and $H_{2,T}$ are sublinear (i.e., $H_{1,T}=o(T)$, $H_{2,T}=o(T)$) which enforces that the amount of nonstationarity cannot grow faster than or equal to time itself. In such a setting, the rate of bilevel local regret achieved by OAGD is sublinear when properly selecting the window size such that $w=o(T)$. Note as the constant M_f in [34] has a quadratic dependency on κ_g, that is $M_f=O(\kappa_g^2)$, the total dependency of the condition number κ_g in the sublinear rate of bilevel local regret of OAGD is fourth order.

D.2. SOBOW

We restate the bilevel local regret of Theorem 5.7 from [24] with our notation.

Theorem 26 *(Theorem 5.7 in [24]) Suppose Assumptions A, B, D, E and furthermore Assumption 5.4 of [24]. Let the inner step size of* $\eta \leq \frac{1}{\ell_{g,1}}$*, decay parameter of* $\nu \in \left(1 - \frac{\alpha\mu_g}{2}\right)$*, and outer step size of* $\alpha \leq \min\left\{\frac{1}{4\ell_{F,1}}, \frac{\mu_g^2 \ell_{F,1} W(1-\nu)(\nu-1+\alpha\mu_g/2}{24\ell_{g,1}^4 G_2 \nu}\right\}$*. Then we have*

$$\sum_{t=1}^{T} \|\nabla F_{t,w}(\boldsymbol{\lambda}_t)\|^2 \leq C_1 \left(\frac{2TQ}{w} + V_{1,T}\right)$$
$$+C_2 \left(\frac{1}{2} + \beta\ell_{F,1}\right) \frac{G_2 G_4}{G_3} H_{2,T} + C_3 = O\left(\frac{T}{w} + V_{1,T} + \kappa_g^3 H_{2,T}\right) \tag{86}$$

where $G_1 = O(\kappa_g^2), G_2 = O(\kappa_g^2), G_3 = O(1), G_4 = O(\kappa_g), \beta\ell_{F,1} = O(1)$ *and constants* $C_1, C_2, C_3 \in \mathbb{R}$ *are from Theorem 5.7 in [24].*

The work of [24], similarly inspired by the dynamic regret literature, considers the setting of sublinear $H_{2,T}$ and $V_{1,T}$– where the former term is second order inner level path variation and the latter term measures the variation in evaluations of the outer level objective function. In such a setting, the rate of bilevel local regret achieved by SOBOW is sublinear when the window size is properly selected such that $w = o(T)$. Recalling the dependencies of κ_g in the terms of $G_1, G_2, G_3, G_4, \beta\ell_{F,1}$ in [24], we remark the total dependency of the condition number κ_g in the sublinear rate of bilevel local regret of SOBOW is third order.

D.3. OBBO

We restate the sublinear rate of bilevel local regret achieved by OBBO in Theorem 2 below.

Theorem 27 *Suppose Assumptions A, B, D, E, and F. Let the inner step size of* $\eta < \min\left(\frac{1}{\ell_{g,1}}, \frac{1}{\mu_g}\right)$*, outer step size of* $\alpha \leq \min\left\{\frac{3\rho}{4\ell_{F,1}}, \frac{\rho\sqrt{(1-\nu)}}{\kappa_g\sqrt{108 C_{\mu_g} L_{\boldsymbol{\beta}}}}\right\}$*, and inner iteration count* $K = \frac{\log(T)}{\log((1-\eta\mu_g)^{-1})} + 1$*. For simplicity, assume* $\phi_t(\boldsymbol{\lambda}) = \phi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|^2$*, and* $h(\boldsymbol{\lambda}) = 0$*. Then the bilevel local regret of our OBBO algorithm satisfies*

$$\sum_{t=1}^{T} \|\nabla F_{t,w}(\boldsymbol{\lambda}_t)\|^2 \leq \frac{64}{\alpha\rho} \left(\frac{2TQ}{w} + V_{1,T}\right)$$
$$+\frac{297}{\rho^2}\left(\frac{\Delta_{\boldsymbol{\beta}} L_{\boldsymbol{\beta}}}{(1-\nu)} + L_3^2\right) + \frac{1188 L_{\boldsymbol{\beta}} C_{\mu_g}}{\rho^2(1-\nu)} H_{2,T} = O\left(\frac{T}{w} + V_{1,T} + \kappa_g^2 H_{2,T}\right) \tag{87}$$

As in the work of [24], we consider the setting of sublinear $H_{2,T}$ and $V_{1,T}$. In such a setting, the rate of bilevel local regret achieved by OBBO is sublinear when properly selecting the window size such that $w = o(T)$. Further, we remark that as $L_{\boldsymbol{\beta}} = O(\kappa_g^2)$, the total dependency of the condition number κ_g in the sublinear rate of bilevel local regret of OBBO is second order. Compared to the regret achieved by OAGD and SOBOW in Theorem 25 and 26 this is a second-order and first-order improvement respectively.

D.4. SOBBO

We restate the sublinear rate of bilevel local regret achieved by SOBBO in Theorem 3 below.

Theorem 28 *Suppose Assumptions A, B, C, D, E, and F. Let the inner step size $\eta \leq \frac{2}{\ell_{g,1}+\mu_g}$, the inner iteration count $K \geq 1$, the outer step size $\alpha \leq \min\left\{\frac{3\rho}{4\ell_{F,1}}, \frac{\rho\sqrt{(1-\nu)}}{\kappa_g^2\sqrt{72C_{\mu_g}}}\right\}$, and inner and outer level batch sizes of $s = w$ and $m = \log(w)/\log\left(1-\frac{\mu_g}{\ell_{g,1}}\right)+1$ respectively. For simplicity, assume $\phi_t(\boldsymbol{\lambda}) = \phi(\boldsymbol{\lambda}) = \frac{1}{2}\|\boldsymbol{\lambda}\|^2$, and $h(\boldsymbol{\lambda}) = 0$. Then the bilevel local regret of SOBBO satisfies*

$$\begin{aligned}\sum_{t=1}^{T}\|\nabla F_{t,w}(\boldsymbol{\lambda}_t)\|^2 &\leq \frac{64}{\alpha\rho}\left(\frac{2TQ}{w}+V_{1,T}+\Delta_h\right)\\ &+\frac{198}{\rho^2}\frac{T}{w}\left(2\sigma_f^2+2\ell_{f,1}^2\kappa_g^2+\frac{C_K\kappa_g^2\eta^2\sigma_{g_{\boldsymbol{\beta}}}^2}{(1-\nu)}\right)+\frac{198}{\rho^2}\frac{\kappa_g^2\Delta_{\boldsymbol{\beta}}}{(1-\nu)}+\frac{396C_{\mu_g}\kappa_g^2}{\rho^2(1-\nu)}\sum_{t=2}^{T}H_{2,T}\\ &= O\left(\frac{T}{w}\left(1+\kappa_g^2+\sigma_f^2+\kappa_g^2\sigma_{g_{\boldsymbol{\beta}}}^2\right)+V_{1,T}+\kappa_g^2H_{2,T}\right) \qquad (88)\end{aligned}$$

Following the setup for the deterministic case, we consider the setting of sublinear $H_{2,T}$ and $V_{1,T}$. In such a setting, the rate of bilevel local regret achieved by SOBBO is sublinear when properly selecting the window and *batch* sizes such that $w = o(T)$ and $s = o(T)$. Due to SOBBO only having access to noisy gradient samples, a sublinear rate of gradient samples is required. Assuming the above conditions are satisfied, SOBBO achieves a sublinear rate of bilevel local regret with the sublinear rate further generalizing the deterministic result to finite outer and inner variances $\sigma_f^2, \sigma_{g_{\boldsymbol{\beta}}}^2$.

Algorithm	$BLR_w(T)$
OAGD	$O(T/w + H_{1,T} + \boldsymbol{\kappa}_{\boldsymbol{g}}^{\mathbf{4}}H_{2,T})$
SOBOW	$O(T/w + V_{1,T} + \boldsymbol{\kappa}_{\boldsymbol{g}}^{\mathbf{3}}H_{2,T})$
OBBO	$O(T/w + V_{1,T} + \boldsymbol{\kappa}_{\boldsymbol{g}}^{\mathbf{2}}H_{2,T})$
SOBBO	$O\left(T/w\left(1+\kappa_{\mathbf{g}}^{\mathbf{2}}+\sigma_f^2+\kappa_{\mathbf{g}}^{\mathbf{2}}\sigma_{g_{\boldsymbol{\beta}}}^2\right)+V_{1,T}+\kappa_{\mathbf{g}}^{\mathbf{2}}H_{2,T}\right)$

Table 3: Bilevel local regret, $BLR_w(T)$, of OBBO vs. online bilevel benchmarks SOBOW from [24] and OAGD from [34]. Note the first and second order-wise improvement OBBO offers in terms of the condition number κ_g dependency. Bilevel local regret for SOBBO is also included, generalizing the deterministic case. The $BLR_w(T)$ is reported in online rounds T , window parameter w, comparator sequences $V_{1,T}, H_{1,T}, H_{2,T}$, condition number κ_g, and finite outer and inner variances $\sigma_f^2, \sigma_{g_{\boldsymbol{\beta}}}^2$, respectively.

Appendix E. Experimental Results

We provide two experiments to demonstrate the superior performance and efficiency of our algorithms relative to the online bilevel benchmarks of OAGD ([34]) and SOBOW ([24]). For our algorithms, we choose the reference function of $\phi_t(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T\mathbf{H}_t\boldsymbol{\lambda}$ such that $\mathbf{H}_t$ is an adaptive matrix of averaged gradient squares with coefficient 0.9, commonly applied in prior works ([20],[19]).

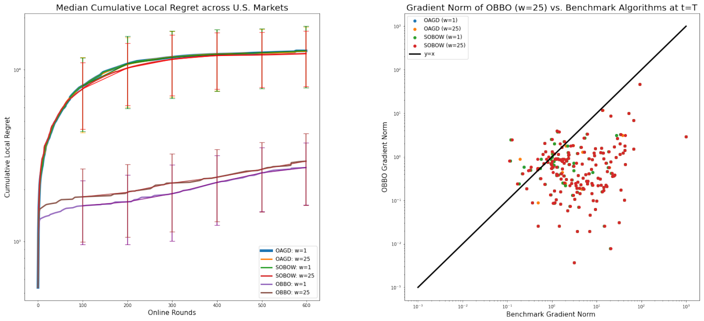

Figure 1: **Left Panel**: Median cumulative local regret of OBBO vs. benchmark algorithms and median deviation bars plotted every 100 rounds. **Right Panel**: Gradient norm of OBBO (w=25) vs. benchmark algorithms at $t = T$ with $y = x$ line plotted to visualize the improvement OBBO offers in achieving a solution with smaller gradient norm.

	Median Loss	Median Deviation
OBBO	**0.205**	0.150
OAGD	0.265	0.209
SOBOW	0.273	0.215

Table 4: Median test loss statistics for U.S. markets (w=25).

	Mean Loss	Standard Error
OBBO	**0.661**	0.055
OAGD	0.707	0.053
SOBOW	0.689	0.053

Table 5: Mean test loss statistics for U.S. markets (w=25).

E.1. Online Hyperparameter Optimization

Hyperparameter optimization has often been formulated as a bilevel optimization as in [32] and [25]. In hyperparameter optimization, the goal is to find optimal hyperparameter values on a validation dataset for optimal parameter values on a training dataset. Specifically, we consider an *online* hyperparameter optimization where the underlying data distribution can vary across time. Compared to the offline case, an online framework captures a larger class of hyperparameter optimization problems (e.g., nonstationarity in optimal hyperparameters).

In online hyperparameter optimization, at each time t, new data samples split into a training and validation set, that is $D_t := \{D_t^{tr}, D_t^{val}\}$, arrive from a potentially new distribution. The inner objective is a regularized training loss on D_t^{tr} of the form $\sum_{\boldsymbol{x} \in D_t^{tr}} L(\boldsymbol{\beta}, \boldsymbol{x}) + \Omega(\boldsymbol{\lambda}, \boldsymbol{\beta})$ for a loss function $L(\boldsymbol{\beta}, \boldsymbol{x})$ evaluated across samples $\boldsymbol{x} \in D_t^{tr}$ for parameters $\boldsymbol{\beta}$ and the regularization function of $\Omega(\boldsymbol{\lambda}, \boldsymbol{\beta})$. Given the optimal parameters $\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})$ from the inner optimization, the outer objective is the validation loss on D_t^{val} of the form $\sum_{\boldsymbol{x} \in D^{val}} L(\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}), \boldsymbol{x})$.

We consider a **Market Impact** dataset consisting of equity price time series. In particular, this dataset contains time series corresponding to 440 significant market impact events annotated by experts from the components of the S&P 500 index between January 2021 and December 2022. For each annotated event, there is a corresponding sequence of 600 training-validation subsets with equal length of 700 observations constructed on a rolling basis, see a sample of this time series split in Figure 2. Additionally, for each annotated event, there is a test set of 120 observations held out for evaluation such that the last corresponding training subset goes up to the annotation. We consider the task of time series forecasting on the post-annotation test set given the available training-validation subsets, and quantify performance by the mean-squared error.

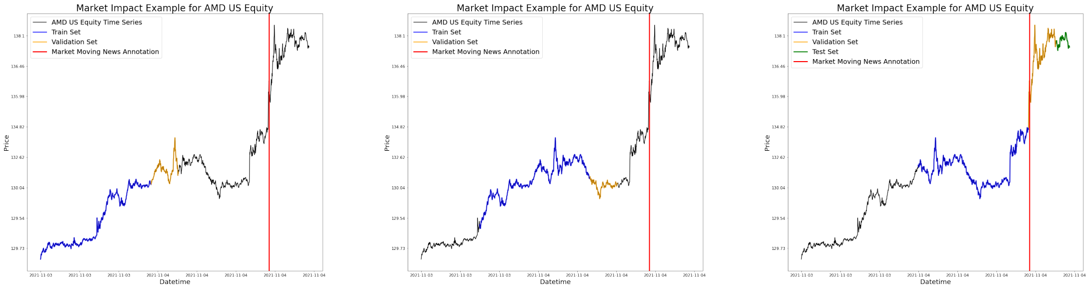

Figure 2: Sample training-validation subsets for AMD U.S. Equity with market event on 11-08-2021.

We consider a smoothing spline model of linear order where the inner level variables are B-spline coefficients and the outer level variable is a positive regularization hyperparameter, respectively fitted on the train and validation datasets. The simplicity of such a model allows us to use closed-form hypergradients, instead of an inner gradient descent loop. All algorithms and window configurations have the outer learning rate set at $\alpha = 0.001$. For OBBO we use the reference function of $\phi_t(\boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{\lambda}^T \mathbf{H}_t \boldsymbol{\lambda}$ such that $\mathbf{H}_t$ is an adaptive matrix of averaged gradient squares with coefficient 0.9, previously applied in prior works ([20],[19]). Default coefficients from PyTorch [31] are used for ADAM ($\beta_1 = 0.9, \beta_2 = 0.999$) as well as SGDM ($\beta_1 = 0.9$) with gradient clipping applied to all algorithms using the threshold on the gradient norm of $\|\nabla f_{t,w}\|^2 \leq 1000$.

In the left panel of Figure 3, note the significant improvement in the median cumulative local regret achieved by OBBO relative to online bilevel (OAGD, SOBOW) and offline general purpose (ADAM, SGDM) benchmark algorithms across 440 U.S. markets. This empirical improvement in cumulative local regret further justifies our theoretical results provided in Theorem 2 and Theorem (3). Further in the middle and right panels of Figure 3 we visualize stability metrics across algorithms of (i) the gradient norm at $t = T$, and (ii) the forecasting mean-squared error on a test set. Specifically we see OBBO often achieves a smaller gradient norm at $t = T$ and a smaller forecasting loss relative to online and offline benchmarks.

In Figure 4, we include forecasts generated from OBBO vs. benchmarks algorithms across a sample of training-validation subsets for the AMD U.S. equity time series. Note how OBBO is quicker to adapt to the annotated market impact event and achieves a better fit (i.e., smaller forecasting loss) relative to the benchmarks on the post-annotation test set. Both of the aforementioned improvements are exhibited across the Market Impact dataset and are not particularly sensitive to experiment design (e.g., number of subsets) or hyperparameters (e.g., window size). Descriptive statistics of forecasting loss aggregated across samples from the Market Impact dataset are in Table 6.

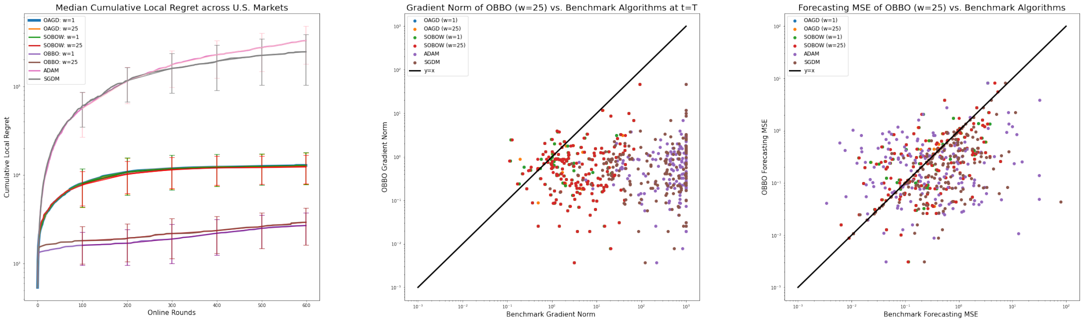

Figure 3: Left Panel: Median cumulative local regret of OBBO vs. online and offline benchmark algorithms across 440 U.S. markets with window size parameter $w = 1, 25$ and median deviation bars plotted every 100 rounds. **Middle Panel**: Gradient norm of OBBO (w=25) vs. online and offline benchmark algorithms at $t = T$ with $y = x$ line plotted to visualize the improvement OBBO offers in achieving a smaller gradient norm. **Right Panel**: Forecasting mean-squared error of OBBO (w=25) vs. online and offline benchmark algorithms with $y = x$ line plotted to visualize the improvement OBBO offers in forecasting loss on a test set.

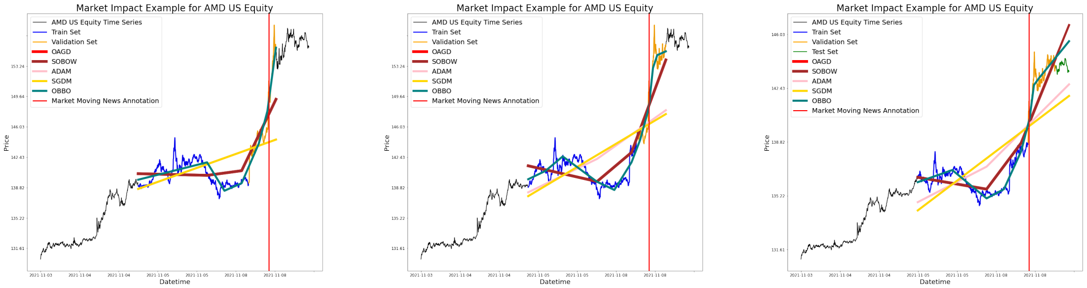

Figure 4: Example forecasts generated with OBBO vs. online (w=25) and offline benchmark algorithms. Note how OBBO achieves a better fit (i.e., smaller loss) relative to benchmarks on the post-annotation test set.

	Forecasting Loss across U.S. Markets			
Algorithm	Mean Loss	Standard Error	Median Loss	Median Absolute Deviation
OBBO	0.661	0.055	0.205	0.150
OAGD	0.707	0.053	0.265	0.209
SOBOW	0.689	0.053	0.273	0.215
Adam	1.265	0.176	0.267	0.230
SGDM	0.872	0.078	0.401	0.286

Table 6: Statistics of forecasting mean-squared error across U.S. markets for window parameter $w = 25$.

E.2. Online Meta-Learning

Meta-learning is frequently formulated as a bilevel optimization problem, see, e.g., [8] and [33]. The objective is to learn optimal meta-parameters in the outer optimization, which, after adaptation—typically via a gradient descent step—yield optimal parameters for the inner optimization. We consider the online meta-learning framework proposed in [9], which extends the traditional offline formulation to accommodate non-stationary task distributions in the online setting.

The online meta-learning problem is setup where at each time t, a new task D_t composed of training and validation samples, that is $D_t : \{D_t^{tr}, D_t^{val}\}$, arrives from a potentially nonstationary distribution. The goal of the learner in the inner objective is to learn task specific parameters $\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda})$ for task D_t through the minimization of $\sum_{\boldsymbol{x} \in D_t^{tr}} L(\boldsymbol{\beta}, \boldsymbol{x}) + \frac{\gamma}{2} \|\boldsymbol{\lambda} - \boldsymbol{\beta}\|^2$ for $\boldsymbol{\beta}$. Note the learner requires a loss function $L(\boldsymbol{\beta}, \boldsymbol{x})$, fixed meta-learned parameters $\boldsymbol{\lambda}$, and regularization constant γ that can also be learned. The meta-learner aims to learn optimal meta-parameters in the outer objective, that is $\sum_{\boldsymbol{x} \in D_t^{val}} L(\widehat{\boldsymbol{\beta}}_t(\boldsymbol{\lambda}), \boldsymbol{x})$, such that after task adaptation task-specific parameters are optimal.

We consider the **FC100** dataset from the CIFAR100 dataset for few-shot learning tasks. Originally introduced within [30], this dataset has been previously utilized in online meta-learning experiments such as in the OBO work of [34]. The dataset contains 100 classes, split into 60:20:20 classes for meta-training, meta-validation and meta-testing respectively. Samples are transformed into tasks via the procedure of [34] resulting in 20,000, 600, and 600 training-validation-test tasks.

As in the setup of [34], we consider a 4-layer convolutional neural network with each layer containing 64 filters. The CNN utilized has 4 convolutional blocks such that there is 3×3 convolution, batch normalization, ReLU activation, and 2×2 max pooling. Inner and outer learning rates are set as $\eta = 0.1$ and $\alpha = 1e - 4$. Following [34], we use the hypergradient estimate of $\widetilde{\nabla} f_t(\boldsymbol{\lambda}, \boldsymbol{\beta})$ computed via a fixed point approach as in [14]. For OBBO we use the function of $\phi_t(\boldsymbol{\lambda}) = \frac{1}{2} \boldsymbol{\lambda}^T \mathbf{H}_t \boldsymbol{\lambda}$ such that $\mathbf{H}_t$ is an adaptive matrix of averaged gradient squares with coefficient 0.9.

In the left panel of Figure 5, OBBO achieves a significant improvement in cumulative bilevel local regret relative to benchmarks OAGD and SOBOW across samples from the FC100 dataset. In the right panel of Figure 5, the histogram displays how OBBO achieves smaller evaluated gradient norms across iterations vs. the benchmark algorithms of OAGD and SOBOW. In the left panel of Figure 6, we report higher training accuracy achieved with OBBO. In the right panel of Figure 6, OBBO outperforms test accuracy relative to SOBOW while achieving OAGD performance with a 10x ($w = 10$) computationally cheaper update. All results are averaged across 5 random seeds.

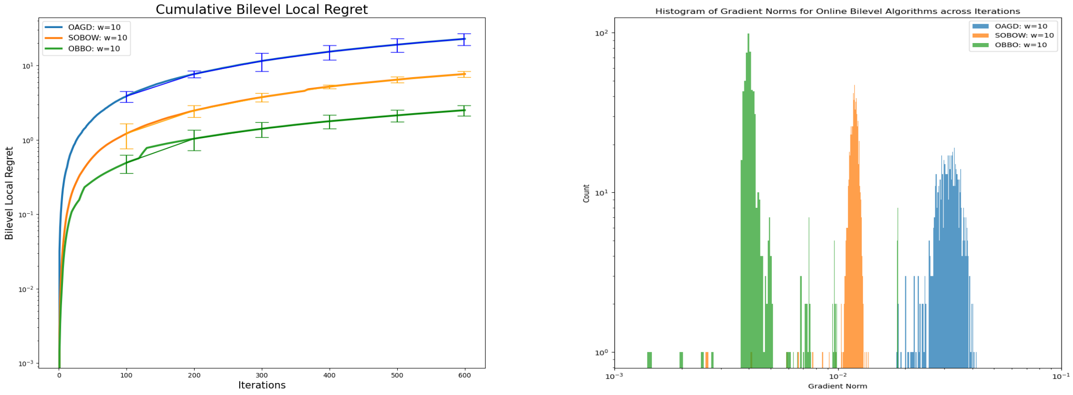

Figure 5: Left Panel: Significant improvement with OBBO on in cumulative bilevel local regret.
Right Panel: Gradient norms across iterations are smaller for OBBO than benchmarks with same initialization.

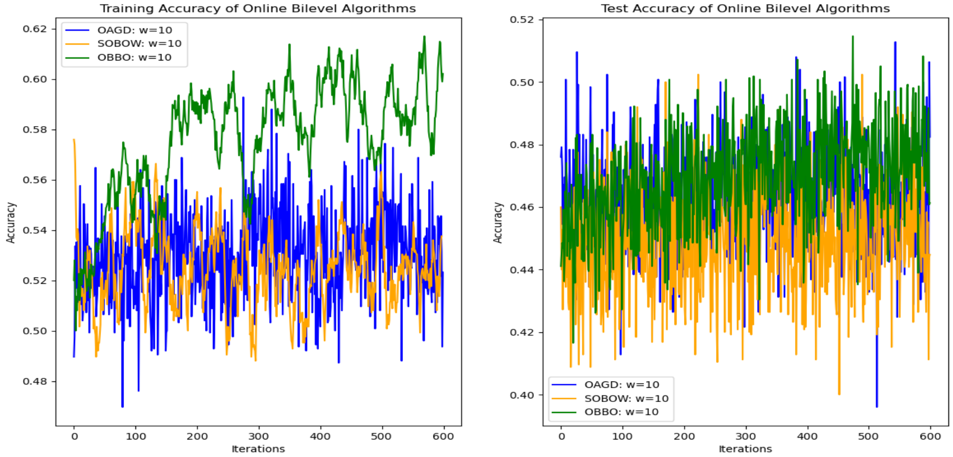

Figure 6: Left Panel: Higher training accuracy achieved with OBBO. **Right Panel**: Test accuracy: OBBO outperforms SOBOW while achieving OAGD performance with 10x ($w = 10$) computationally cheaper update.