

DYNAMIC FISTA FOR CONVEX COMPOSITE BILEVEL OPTIMIZATION*

ROEY MERHAV[†], SHOHAM SABACH[†], AND MARC TEBOULLE[‡]

Abstract. In this paper, we study convex bilevel optimization problems where both the inner and outer levels are given as a composite convex minimization. We propose the Fast Bilevel Proximal Gradient (FBI-PG) algorithm, which can be interpreted as applying FISTA to a dynamic regularized composite objective function. The dynamic nature of the regularization parameters allows one to achieve an optimal fast convergence rate of $O(1/k^2)$ in terms of the inner objective function. This is the fastest known convergence rate under no additional restrictive assumptions. We also show that FBI-PG achieves sublinear simultaneous rates in terms of both the inner and outer objective functions. Moreover, we show that under a Hölderian type error bound assumption on the inner objective function, the FBI-PG algorithm achieves improved simultaneous rates and converges to an optimal solution of the bilevel optimization problem. Finally, we present numerical experiments demonstrating the performance of the proposed scheme.

Key words. convex minimization, bilevel optimization, first order methods, dynamic proximal gradient, fast convergence rate, global convergence

MSC codes. 90C25, 65K05

DOI. 10.1137/25M1767893

1. Introduction. Bilevel optimization emerges as a dynamic and evolving research field within the broad optimization area. In this paper, we consider the bilevel convex minimization problem, which aims at minimizing a convex function, called an outer objective function, over the optimal solution set of another convex minimization of a function called the inner objective. We are interested in studying a convex bilevel model where both objectives in the inner and outer levels are instances of composite convex minimization problems. Our focus is on achieving faster convergence rates and corresponding pointwise global convergence for a first-order algorithm to this important class of bilevel composite convex optimization problems.

Following the terminology of inner and outer problems, we are interested in considering an outer problem given by the classical convex composite model

$$(OP) \quad \min_{\mathbf{x} \in X^*} \{\omega(\mathbf{x}) := \sigma(\mathbf{x}) + \psi(\mathbf{x})\},$$

where X^* is the assumed nonempty set of minimizers of the inner level problem, which is also described via another convex composite model

$$(P) \quad \min_{\mathbf{x} \in \mathbb{R}^n} \{\varphi(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}.$$

*Received by the editors June 11, 2025; accepted for publication (in revised form) February 23, 2026; published electronically July 15, 2026.

<https://doi.org/10.1137/25M1767893>

Funding: The research of the second author was partially supported by the Israel Science Foundation (ISF) grant 2480-21. The research of the third author was partially supported by the Israel Science Foundation (ISF) grant 2619-20.

[†]Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 3200003, Israel (merhav.roey@campus.technion.ac.il, ssabach@technion.ac.il).

[‡]School of Mathematical Sciences, Tel-Aviv University, Ramat-Aviv 69978, Israel (teboulle@tauex.tau.ac.il).

The functions $f, \sigma : \mathbb{R}^n \rightarrow \mathbb{R}$ are smooth and convex, while the functions $g, \psi : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ are nonsmooth and convex. We will provide the exact details of the problem's data in the forthcoming section. Moreover, we refer to φ and ω as the inner and outer objective functions, respectively. Throughout this paper, we denote by $X' \subset X^*$ the set of all optimal solutions of the outer optimization problem (OP).

This type of bilevel minimization problem is also known sometimes under the name *simple bi-level programming* as being a particular case of the more general bilevel programming model, which involves, in both the inner/outer problems, a parametric representation featuring two variables. This simpler model, however, is very flexible and, as such, relevant in modeling various modern applications; see, for instance, [11] for a thorough overview of bilevel optimization problems as well as their applications, and the literature therein. Very recently, simple bilevel problems have also been studied in the more challenging setting of variational inequalities at both levels (see [15, 16, 21], and references therein).

The implicit constraints imposed by the inner minimization problem make such problems difficult to develop simple and efficient algorithms. In recent years, there has been a growing interest in deriving theoretical guarantees beyond mere convergence results for convex bilevel optimization problems under various types of assumptions on the problem's data. For some of the most recent advances in this area, we refer readers to [19] and the recent paper [17], which both provide ample details on the past and current developments. Despite many of these interesting results, a grand challenge in this field remains open: *developing a first-order algorithm capable of achieving fast convergence rates for convex bilevel optimization problems* akin to the existing faster rates of first-order schemes for the traditional (single level) convex composite minimization [18, 2, 4].

Motivated by the remarkable success of fast first-order algorithms for traditional convex composite optimization problems just alluded to above, our objective is to derive similar accelerated rates for the more challenging setting of bilevel convex composite optimization problems. Toward this goal, we propose employing the Tikhonov regularization technique, which involves approaching bilevel optimization as the minimization of either a single *fixed* regularized (penalized) or a *dynamic* regularized composite convex objective. The convex composite structure of the regularized objective thus naturally calls for applying any fast first-order scheme on it. However, at this point, we faced a hurdle. As we shall see below, we cannot apply existing convergence rate results directly due to either the lack of knowledge of the fixed regularization parameter or the dynamic change of the regularized composite objective. Thus, to tackle this inherent difficulty, a fresh analysis/approach is required.

Literature. Before concluding this section, we briefly discuss several recent works related to our objectives. The literature on algorithms for simple bilevel optimization problems can broadly be categorized into two classes: *direct* methods and *nested* methods. The key distinction lies in computational complexity. Roughly speaking, direct methods avoid solving an inner optimization problem at each iteration, typically by considering an adequately regularized problem; in contrast, nested methods use an approximation that requires additional computations involving another optimization problem.

Given our focus on developing the simplest possible algorithms, we do not discuss nested approaches in detail. For instance, recent works such as [14, 8, 26, 27] propose nested algorithms with convergence guarantees for bilevel problems where the inner constraint set Z is compact and convex. However, compactness is essential to their analysis, and the guarantees generally fail without it. Moreover, these methods require

computing projections onto the intersection of Z and a hyperplane, a task that is often computationally infeasible in practice.

Turning to direct methods, we highlight the recent works [10, 20] that address a regularized version of the problem using a fixed regularization parameter determined by a prespecified iteration count K . While this approach enables a trivial analysis (see more details below), it does not guarantee recovery of a solution to the original bilevel problem.¹ More generally, as recognized in the literature, choosing an appropriate regularization parameter is a fundamentally difficult and often impractical task. Although convergence analysis under fixed regularization is typically straightforward (as we discuss below), it offers limited insight into the solution of the original bilevel formulation.

This paper is organized as follows. The subsequent section will provide a comprehensive discussion of the Tikhonov technique and its strong connections to bilevel optimization problems. This will lead us to propose the Fast Bilevel Proximal Gradient (FBi-PG) algorithm that can be simply interpreted as applying FISTA [4] to the dynamic regularized objective. The algorithm FBi-PG is equipped with a *dynamic* update of the regularization parameter, which is itself characterized by the choice of another fixed parameter allowing us to devise accelerated schemes. This is developed in section 3, where we will prove our main rate of convergence results. These analyses highlight the significance and crucial roles played by the dynamic regularization parameter, which is user-determined in regulating the performance of the FBi-PG scheme, a method which (for the first time) is proven to be capable of achieving the fast convergence rate of $O(1/k^2)$ in terms of the inner optimization problem; see Theorem 3.7. Additionally, we establish convergence rates concerning the outer objective function and simultaneous rates concerning both functions. In the classical context of convex composite minimization, it is well-known that the sequence generated by FISTA does not necessarily converge to an optimal solution. This issue was, however, successfully addressed in [9] through a smart analysis of a variant of FISTA. The natural question is then whether such a result can be extended to the more involved convex composite bilevel setting. In section 4, we positively address this question by adapting and extending the analysis of [9] to this case, and under a Hölder type error bound assumption [6], we prove the convergence of the generated sequences by FBi-PG to an optimal solution of the bilevel problem (OP). Finally, in section 5 we present numerical experiments demonstrating the significant rate/improvement achieved by FBi-PG in comparison to other current state-of-the-art schemes.

2. The FBi proximal gradient algorithm.

2.1. Motivation and difficulties. Several algorithmic approaches exist for addressing bilevel optimization problems. One of the earliest ideas dates back to the book of Tikhonov and Arsenin [25], who introduced what is now known as the Tikhonov regularization technique. In their original work, the goal was to find a minimal norm solution (meaning the outer objective function is $\omega \equiv \sigma = \|\cdot\|^2$) to a linear optimization problem. Its primary advantage lies in simplifying the bilevel problem into a single regularized formulation, which avoids the need to handle two optimization problems. In the context of our paper, it means to tackle the following penalized problem:

¹See our numerical section for a concrete example illustrating this limitation in a real-world application.

$$(2.1) \quad \min_{\mathbf{x} \in \mathbb{R}^n} \{F_\alpha(\mathbf{x}) \equiv \varphi(\mathbf{x}) + \alpha\omega(\mathbf{x})\},$$

where $\alpha > 0$ is the regularization parameter. In this context, we mention the work [13], which studies, among other questions, the connections between bilevel problems and their regularized counterpart in a general convex setting. They proved that there is a small enough $\alpha > 0$ such that an optimal solution of the regularized problem (2.1) is an optimal solution of the bilevel problem (OP). Thus, a seemingly natural approach to obtain rates of convergence for the convex composite bilevel problem, based on the later idea of a priori fixing the regularization parameter α , would be to simply apply any first-order scheme on the regularized objective function F_α , for some predefined parameter α , to obtain rates on the *regularized* objective function F_α . This approach has recently been formalized in [10, 20], though the resulting analysis is straightforward and follows directly from standard results on first-order methods. Indeed, for instance, an application of FISTA [4, Theorem 4.4] on minimizing F_α with say an L_α -Lipschitz gradient part of the objective will generate a sequence $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ which satisfies

$$F_\alpha(\mathbf{x}^k) - F_\alpha(\mathbf{x}') \leq \frac{2L_\alpha \|\mathbf{x}^0 - \mathbf{x}'\|^2}{(k+1)^2} \quad \forall \mathbf{x}' \in X'.$$

Using the definition of F_α , and since $\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') \geq 0$ for all $k \in \mathbb{N}$, one immediately deduces the following fast rate of convergence for the outer problem

$$(2.2) \quad \omega(\mathbf{x}^k) - \omega(\mathbf{x}') \leq \frac{2L_\alpha \|\mathbf{x}^0 - \mathbf{x}'\|^2}{\alpha(k+1)^2} \quad \forall \mathbf{x}' \in X'.$$

At this juncture, it is important to note that a generic result can be easily proved for the case of a fixed regularizing parameter by showing that independently of the structure of the bilevel optimization problem and for any single-level optimization algorithm, the arguments above remain valid. For the sake of completeness, the proof of this fact appears in Appendix A.

Unfortunately, finding/knowing the value of a fixed regularization parameter α , for which a solution of the regularized problem recovers an optimal solution of the bilevel problem, remains a challenging task. Furthermore, finding a corresponding rate for the inner optimization problem also depends on the unknown value of α . Therefore, obtaining rates of convergence for bilevel optimization problems *cannot* rely on existing results for single-level optimization problems, and requires a fresh, specially designed approach.

The first work to provide an efficient use of the Tikhonov regularization technique tailored for bilevel optimization problems is [22]. It was confined to the convex bilevel problem with a smooth and convex outer objective ω and with the inner objective function φ being the sum of a smooth and convex function and an indicator of a closed and convex set. The resulting projected gradient algorithm applied on F_α proposed in [22] suggests that at each iteration k we regularize the outer objective function ω with a *dynamic* parameter α_k , and it is proved that if $\alpha_k \rightarrow 0$ as $k \rightarrow \infty$ as long as $\sum_{k=1}^\infty \alpha_k = \infty$, then any limit point of the generated sequence is an optimal solution of the bilevel optimization problem. Therefore, following [22], using a dynamic regularization parameter at each iteration seems to be the right approach to overcome the lack of knowledge of the “right” regularization parameter α . Yet, one of the interesting open questions about this approach, which we address in this work, pertains to establishing convergence rates, particularly fast convergent rates, as well

as the corresponding pointwise global convergence results for the more general convex composite bilevel problem (OP).

A first attempt in this direction was taken in the thesis [24]. In that thesis, the fast scheme based on the accelerated scheme of [2] (see also [23] and references therein), which differs from FISTA, was used. However, that study did not cover the general convex composite bilevel problem (OP), the derived fast rate of convergence result was limited only to the inner problem, and no simultaneous rate was established. Moreover, pointwise global convergence was not derived. Inspired by [24] and using FISTA as our basis, we can significantly extend both the rate and the global convergence results beyond those presented in [24] in various scenarios; see below for a description of our main results.

2.2. Problem's statement and the FBi-PG algorithm. Motivated by the above discussion, this paper extends the framework of FISTA proposed in [4] to introduce the Fast Bilevel Proximal Gradient (FBi-PG) algorithm designed specifically for solving convex bilevel optimization problems. Our approach incorporates the Tikhonov regularization technique, utilizing a *dynamically* updated regularization parameter to enhance the algorithm's performance in achieving a fast scheme. More precisely, given $\gamma > 0$, we will use the following sequence of regularizing parameters $\{\alpha_k\}_{k \in \mathbb{N}}$, which is defined for an $a \in \mathbb{N}$ such that $a \geq 2$, by

$$(2.3) \quad \alpha_k = (k + a)^{-\gamma}, \quad k \in \mathbb{N}.$$

For the sake of simplicity of the developments below, we define $\alpha_{-1} := 0$. Note that we could have set $a \equiv 1$. The additional parameter a is introduced here to simplify/unify our presentation, which also allows us to handle pointwise global convergence, as seen in section 4.

Before presenting the algorithm, we consider the inner and outer optimization problems (P) and (OP). Throughout this paper, we will use the following standard assumption on the problem's data.

Assumption 2.1.

- (i) The functions $f, \sigma : \mathbb{R}^n \rightarrow \mathbb{R}$ are convex, continuously differentiable, and have Lipschitz continuous gradients with the Lipschitz parameters β_f and β_σ , respectively. We denote $\beta \equiv \beta_f + \beta_\sigma \in \mathbb{R}_{++}$.
- (ii) The functions $g, \psi : \mathbb{R}^n \rightarrow (-\infty, \infty]$ are proper, lower semicontinuous, and convex.
- (iii) The set X^* of all optimal solutions of problem (P) is nonempty.
- (iv) $\inf_{\mathbf{x} \in \mathbb{R}^n} \omega(\mathbf{x}) \equiv \omega^* > -\infty$.

Moreover, throughout this paper we systematically define, for all $k \in \mathbb{N}$, the smooth regularized function $f_k \equiv f + \alpha_k \sigma$, the nonsmooth regularized function $g_k \equiv g + \alpha_k \psi$, and the regularized function $F_k \equiv \varphi + \alpha_k \omega$.

Remark 2.2. The functions f and σ having Lipschitz continuous gradients with the Lipschitz parameters β_f and β_σ , respectively. Hence, for any $k \in \mathbb{N}$, the function f_k also has a Lipschitz continuous gradient with the parameter $\beta_f + \alpha_k \beta_\sigma$. Since $\alpha_k \leq 1$, then $\beta_f + \alpha_k \beta_\sigma \leq \beta_f + \beta_\sigma = \beta$.

Based on this remark, we see that the regularized objective function F_k also has a composite structure of the sum of a smooth function f_k with a Lipschitz continuous gradient and a nonsmooth function g_k . We can thus apply FISTA [4] on the dynamic objective F_k , $k \in \mathbb{N}$, to produce the FBi-PG algorithm, which is described below.

Algorithm 2.1. Fast Bilevel Proximal Gradient.

Input: $\gamma > 0$ and $2 \leq a \in \mathbb{N}$.

Initialization: $\alpha_{-1} = t_{-1} = 0$, $t_0 = 1$, and $\mathbf{x}^0 = \mathbf{x}^{-1} \in \mathbb{R}^n$.

for $k \geq 0$ **do**

$$(2.4) \quad \alpha_k = (k + a)^{-\gamma} \quad f_k \equiv f + \alpha_k \sigma \quad g_k \equiv g + \alpha_k \psi,$$

$$(2.5) \quad \mathbf{y}^k = \mathbf{x}^k + t_k^{-1} (t_{k-1} - 1) (\mathbf{x}^k - \mathbf{x}^{k-1}),$$

$$(2.6) \quad \mathbf{x}^{k+1} = \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^n} \left\{ g_k(\mathbf{x}) + \mathbf{x}^T \nabla f_k(\mathbf{y}^k) + \frac{\beta}{2} \|\mathbf{x} - \mathbf{y}^k\|^2 \right\},$$

$$(2.7) \quad t_{k+1} = \frac{k + 1 + a}{a}$$

end for

The FBi-PG algorithm is associated with a parameter $\gamma > 0$, via the regularization parameter $\alpha_k = (k + a)^{-\gamma}$, which enables the capture of a broad range of behaviors within a unified framework. Throughout this paper, we establish three types of theoretical guarantees for the FBi-PG algorithm:

- (i) individual convergence rates for the inner or outer problems;
- (ii) simultaneous convergence rates for both levels;
- (iii) pointwise convergence of the iterates to an optimal solution.

Although simultaneous convergence is the ultimate goal in bilevel optimization, individual convergence guarantees remain valuable. They offer insight into the algorithm's behavior across regimes, reflect the historical progression of the field, which focused initially on inner rates, and often serve as foundational steps toward establishing full simultaneous results.

All our rate results are provided in section 3. We first present our results under the basic Assumption 2.1, and then show how they can be strengthened under a Hölderian error bound for the inner objective function. To help the reader navigate the results, we provide a summary table outlining the convergence guarantees, underlying assumptions, and corresponding theorems.²

TABLE 1
Summary of convergence rate results under various parameter regimes.

Parameter regime	Inner convergence	Outer convergence lower, upper bound	Simul.	Relevant Thms./Props.	Note
$\gamma \in (0, 1)$	$O(1/k^\gamma)$	N/A, $O(1/k^{2-\gamma})$	No	Thm. 3.9, Prop. 3.8	Outer best iterate
$\gamma = 1$	$O(\log k/k)$	N/A, $O(1/k)$	Yes	Thm 3.10	Both best iterate ²
$\gamma \in (1, 2)$	$O(1/k^\gamma)$	N/A, $O(1/k^{2-\gamma})$	No	Thm 3.9, Prop 3.8	Outer best iterate
$\gamma \in (1, 2)$	$O(1/k^2)$	$O(1/k)$, $O(1/k^{2-\gamma})$	Yes	Thm 3.14	Hölderian EB
$\gamma = 2$	$O(\log(k)/k^2)$	N/A	No	Prop 3.8	
$\gamma > 2$	$O(1/k^2)$	N/A	No	Thm 3.7	

Before concluding this section, in the following remark, we briefly discuss the algorithm's main computational step.

²Both best iterate refers to $\operatorname{argmin}\{\omega(\mathbf{x}) : \mathbf{x} \in \{\mathbf{x}^k, \bar{\mathbf{x}}^k\}\}$, where $\bar{\mathbf{x}}^k = \sum_{s=1}^k \mathbf{x}^s/k$.

Remark 2.3 (computation of the prox of the sum of two functions—lifted reformulation). The FBi-PG algorithm needs to compute, at each iteration $k \in \mathbb{N}$, the proximal mapping of the sum $g + \alpha_k \psi$, where the two functions g, ψ are assumed prox-friendly. Most often, this is a difficult and computationally demanding task, even when both functions are prox-friendly. The nice observation made in [12] provides a way to bypass this difficulty. The lifting technique proposed in [12] allows us to eliminate this computational bottleneck by reformulating the bilevel optimization problem as follows:

$$\min_{\mathbf{w} \in \mathbb{R}^{n \times n}} \left\{ \bar{\omega}(\mathbf{w}) : \mathbf{w} \in \underset{\mathbf{w} \in \mathbb{R}^{n \times n}}{\operatorname{argmin}} \bar{\varphi}(\mathbf{w}) \right\},$$

where $\bar{\omega}(\mathbf{w}) \equiv \sigma(\mathbf{x}) + \psi(\mathbf{z})$, and $\bar{\varphi}(\mathbf{w}) \equiv f(\mathbf{x}) + g(\mathbf{x}) + \|\mathbf{x} - \mathbf{z}\|^2/2$.

The equivalence immediately follows, since in the inner problem any minimizer $\mathbf{w}' = (\mathbf{x}', \mathbf{z}')$ must satisfy $\mathbf{z}' = \mathbf{x}'$, with $\mathbf{x}'$ a minimizer of $\bar{\varphi}$. Thanks to this lifted reformulation, applying the Fbi-PG now reduces to computing the prox of the *separable* sum $g(\mathbf{x}) + \alpha_k \psi(\mathbf{z})$, i.e., it reduces to computing separately each of the prox friendly functions g and ψ . However, it is important to note that the lifted reformulation doubles the ambient dimension. Due to the bilevel setting considered here, this increase does not translate into a significant computational burden. The structure of the problem is highly favorable: the outer objective remains block-separable, and the only coupling between the lifted variables in the inner problem is through a simple quadratic term. Consequently, the resulting computations remain efficient while allowing us to bypass the original proximal bottleneck.

3. Rate of convergence analysis of FBi-PG. The section is divided into three parts, beginning with a few key results that will serve as a foundation for deriving the main results. In section 3.2, we will first prove (last iterate) individual inner rates for three different regimes: $0 < \gamma < 2$, $\gamma = 2$, and $\gamma > 2$. In the last regime, we obtain the fast inner rate, which is the best current known acceleration result. Then, we will provide (best iterate) outer rate of convergence result for $0 < \gamma < 2$. We conclude this part with our main result, which is a simultaneous (best iterate) rate of both the inner and outer problems for $\gamma = 1$. In section 3.3, we will prove an improved (last iterate) simultaneous rate for $1 < \gamma < 2$ under the Hölderian error bound condition for the inner objective function.

3.1. Main pillars of the analysis. To obtain our basic inequality for the development of all the announced results, we will need the following classical proximal inequality (see [5, Lemma 2.6]).

PROPOSITION 3.1. *Let $s : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex and continuously differentiable function having a Lipschitz continuous gradient with the Lipschitz parameter $L > 0$, and let $q : \mathbb{R}^n \rightarrow (-\infty, \infty]$ be a proper, lower semicontinuous, and convex function. For a given vector $\mathbf{y} \in \mathbb{R}^n$, we define*

$$\mathbf{x}^+ := \underset{\mathbf{x} \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ q(\mathbf{x}) + \mathbf{x}^T \nabla s(\mathbf{y}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2 \right\}.$$

Then, for all $\mathbf{u} \in \operatorname{dom} q$, we have

$$q(\mathbf{x}^+) + s(\mathbf{x}^+) - (q(\mathbf{u}) + s(\mathbf{u})) \leq \frac{L}{2} \left(\|\mathbf{u} - \mathbf{y}\|^2 - \|\mathbf{u} - \mathbf{x}^+\|^2 \right).$$

At each iteration $k \in \mathbb{N}$, the steps of FBi-PG can be seen as a direct application of the well-known FISTA algorithm [4] on the regularized function F_k . The next

technical result can be obtained from, e.g., [5, Lemma 2.7]. However, for the sake of completeness, we include a simple proof of this result.

PROPOSITION 3.2. *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG. For all $k \in \mathbb{N}$ and $\mathbf{x} \in \mathbb{R}^n$, we have*

$$(3.1) \quad t_k^2 (F_k(\mathbf{x}^{k+1}) - F_k(\mathbf{x})) + \frac{\beta}{2} \|\mathbf{z}^{k+1} - \mathbf{x}\|^2 \leq (t_k^2 - t_k) (F_k(\mathbf{x}^k) - F_k(\mathbf{x})) + \frac{\beta}{2} \|\mathbf{z}^k - \mathbf{x}\|^2,$$

where we define $\mathbf{z}^k := (1 - t_{k-1})\mathbf{x}^{k-1} + t_{k-1}\mathbf{x}^k$.

Proof. Let $k \in \mathbb{N}$. From the $\mathbf{x}^{k+1}$ step (see (2.6)), using Proposition 3.1 with $s(\cdot) = f_k(\cdot)$, $q(\cdot) = g_k(\cdot)$, $\mathbf{y} = \mathbf{y}^k$, $\mathbf{x}^+ = \mathbf{x}^{k+1}$, $\mathbf{u} = (1 - t_k^{-1})\mathbf{x}^k + t_k^{-1}\mathbf{x}$, and $L = \beta$ we get for any $\mathbf{x} \in \mathbb{R}^n$ that (recall $F_k \equiv f_k + g_k$)

$$\begin{aligned} F_k(\mathbf{x}^{k+1}) &\leq F_k((1 - t_k^{-1})\mathbf{x}^k + t_k^{-1}\mathbf{x}) + \frac{\beta}{2} \left(\|(1 - t_k^{-1})\mathbf{x}^k + t_k^{-1}\mathbf{x} - \mathbf{y}^k\|^2 \right. \\ &\quad \left. - \|(1 - t_k^{-1})\mathbf{x}^k + t_k^{-1}\mathbf{x} - \mathbf{x}^{k+1}\|^2 \right) \\ &= F_k((1 - t_k^{-1})\mathbf{x}^k + t_k^{-1}\mathbf{x}) + \frac{\beta}{2t_k^2} \left(\|\mathbf{z}^k - \mathbf{x}\|^2 - \|\mathbf{z}^{k+1} - \mathbf{x}\|^2 \right), \end{aligned}$$

where the last equality follows from the definitions of $\mathbf{z}^k$ and $\mathbf{y}^k$ (see (2.5)). Since F_k is convex, the inequality above yields

$$(3.2) \quad F_k(\mathbf{x}^{k+1}) \leq (1 - t_k^{-1}) F_k(\mathbf{x}^k) + t_k^{-1} F_k(\mathbf{x}) + \frac{\beta}{2t_k^2} \left(\|\mathbf{z}^k - \mathbf{x}\|^2 - \|\mathbf{z}^{k+1} - \mathbf{x}\|^2 \right).$$

Subtracting $F_k(\mathbf{x})$ from both sides and multiplying both sides with t_k^2 , we obtain after rearranging that

$$\begin{aligned} t_k^2 (F_k(\mathbf{x}^{k+1}) - F_k(\mathbf{x})) + \frac{\beta}{2} \|\mathbf{z}^{k+1} - \mathbf{x}\|^2 &\leq (1 - t_k^{-1}) t_k^2 (F_k(\mathbf{x}^k) - F_k(\mathbf{x})) \\ &\quad + \frac{\beta}{2} \|\mathbf{z}^k - \mathbf{x}\|^2, \end{aligned}$$

which proves the desired result. $\square$

From now on, we can no longer apply the rate of convergence results of FISTA (see [4, Theorem 4.4] or [5, Theorem 2.4]) to obtain relevant faster rates for the algorithm FBi-PG, and we need a new line of analysis. Indeed, the *dynamic* nature of the regularized function F_k , $k \in \mathbb{N}$, precludes the possibility of using directly the rates established for FISTA on F_k , unless one either pick a fixed parameter $\alpha_k \equiv \alpha > 0$, for all $k \in \mathbb{N}$, or one fixes in advance the iteration counter at some predefined value $k \equiv K$. However, neither strategy is practically and theoretically viable. Indeed, as discussed earlier in section 2 and as also done in [10, 20], in the former case (among other troubles!) the value of the “right” regularizing parameter α is, in general, not available, and in the latter case it is immediate to see that from (2.2) with $\alpha_K := (K + 1)^{-\gamma}$ (for some $0 < \gamma < 2$) one obtains

$$\omega(\mathbf{x}^K) - \omega(\mathbf{x}') \leq \frac{2\beta \|\mathbf{x}^0 - \mathbf{x}'\|^2}{(K + 1)^{2-\gamma}} \quad \forall \mathbf{x}' \in X',$$

which shows that the fast rate is lost! Furthermore, in general, it makes no sense to fix in advance the number of iterations to be performed by the algorithm FBi-PG since,

eventually to warrant convergence of the sequence generated by FBi-PG to an optimal solution of the bilevel optimization problem (OP), we need to have the sequence of regularized parameters $\{\alpha_k\}_{k \in \mathbb{N}}$ to converge to 0 as $k \rightarrow \infty$. The numerical results presented in section 5 will further illustrate and corroborate this situation. The next results will be useful to overcome the alluded difficulties.

For convenience, we define for all $k \in \mathbb{N}$

$$(3.3) \quad d_k := t_{k-1}^2 - (t_k^2 - t_k) \text{ and } \eta_k := t_{k-1}^2 \alpha_{k-1} - (t_k^2 - t_k) \alpha_k.$$

We note that our choice of t_k satisfies $t_k^2 - t_k \leq t_{k-1}^2$ for all $k \in \mathbb{N}$. Indeed, with $a \geq 2$, we have that

$$t_{k-1}^2 - t_k^2 + t_k = ((a-1)^2 + (a-2)k) / a^2 > 0.$$

We note that since $t_k^2 - t_k \leq t_{k-1}^2$, then $d_k = t_{k-1}^2 - (t_k^2 - t_k) \geq 0$ for all $k \in \mathbb{N}$. Likewise, using the definition of α_k (see (2.4)) we have that $\alpha_k < \alpha_{k-1}$ and thus $\eta_k = t_{k-1}^2 \alpha_{k-1} - (t_k^2 - t_k) \alpha_k \geq t_{k-1}^2 (\alpha_{k-1} - \alpha_k) > 0$ for all $k \in \mathbb{N}$.

PROPOSITION 3.3. *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG. For all $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$, we have*

$$(3.4) \quad t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}')) \leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 - \sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}^s) - \omega(\mathbf{x}')).$$

Proof. From Proposition 3.2 (see (3.1)) we have for all $k \in \mathbb{N}$ and $\mathbf{x} \in \mathbb{R}^n$ that

$$(3.5) \quad \begin{aligned} t_k^2 (F_k(\mathbf{x}^{k+1}) - F_k(\mathbf{x})) + \frac{\beta}{2} \|\mathbf{z}^{k+1} - \mathbf{x}\|^2 &\leq (t_k^2 - t_k) (F_k(\mathbf{x}^k) - F_k(\mathbf{x})) \\ &\quad + \frac{\beta}{2} \|\mathbf{z}^k - \mathbf{x}\|^2. \end{aligned}$$

Now, using the definition of $F_k \equiv \varphi + \alpha_k \omega$, we get that

$$(3.6) \quad \begin{aligned} (t_k^2 - t_k) (F_k(\mathbf{x}^k) - F_k(\mathbf{x})) &= (t_k^2 - t_k) (\varphi(\mathbf{x}^k) - \varphi(\mathbf{x})) \\ &\quad + (t_k^2 - t_k) \alpha_k (\omega(\mathbf{x}^k) - \omega(\mathbf{x})) \\ &= t_{k-1}^2 (\varphi(\mathbf{x}^k) - \varphi(\mathbf{x})) \\ &\quad + (t_k^2 - t_k - t_{k-1}^2) (\varphi(\mathbf{x}^k) - \varphi(\mathbf{x})) \\ &\quad + (\alpha_{k-1} t_{k-1}^2 - \eta_k) (\omega(\mathbf{x}^k) - \omega(\mathbf{x})) \\ &= t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x})) - d_k (\varphi(\mathbf{x}^k) - \varphi(\mathbf{x})) \\ &\quad - \eta_k (\omega(\mathbf{x}^k) - \omega(\mathbf{x})), \end{aligned}$$

where the second equality follows from the fact that $(t_k^2 - t_k) \alpha_k = t_{k-1}^2 \alpha_{k-1} - \eta_k$ while the last equality follows from the definition of $F_{k-1} \equiv \varphi + \alpha_{k-1} \omega$ and the fact that $t_k^2 - t_k - t_{k-1}^2 = -d_k$. Combining it with (3.5) for $k = s$ yields that

$$(3.7) \quad \begin{aligned} t_s^2 (F_s(\mathbf{x}^{s+1}) - F_s(\mathbf{x})) + \frac{\beta}{2} \|\mathbf{z}^{s+1} - \mathbf{x}\|^2 &\leq t_{s-1}^2 (F_{s-1}(\mathbf{x}^s) - F_{s-1}(\mathbf{x})) + \frac{\beta}{2} \|\mathbf{z}^s - \mathbf{x}\|^2 \\ &\quad - d_s (\varphi(\mathbf{x}^s) - \varphi(\mathbf{x})) - \eta_s (\omega(\mathbf{x}^s) - \omega(\mathbf{x})). \end{aligned}$$

Summing (3.7) over $s = 0, 1, \dots, k-1$ results in

$$(3.8) \quad \begin{aligned} t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x})) + \frac{\beta}{2} \|\mathbf{z}^k - \mathbf{x}\|^2 &\leq t_{k-1}^2 (F_{-1}(\mathbf{x}^0) - F_{-1}(\mathbf{x})) + \frac{\beta}{2} \|\mathbf{z}^0 - \mathbf{x}\|^2 \\ &\quad - \sum_{s=0}^{k-1} d_s (\varphi(\mathbf{x}^s) - \varphi(\mathbf{x})) \\ &\quad - \sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}^s) - \omega(\mathbf{x})). \end{aligned}$$

Since φ is proper (see Assumption 2.1) and $\alpha_{-1} = 0$, it follows that $F_{-1} \equiv \varphi + \alpha_{-1}\omega = \varphi < \infty$, along with the facts that $t_{-1} = 0$ and $\mathbf{x}^0 = \mathbf{z}^0$, we obtain from (3.8) that

$$(3.9) \quad \begin{aligned} t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x})) + \frac{\beta}{2} \|\mathbf{z}^k - \mathbf{x}\|^2 &\leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}\|^2 - \sum_{s=0}^{k-1} d_s (\varphi(\mathbf{x}^s) - \varphi(\mathbf{x})) \\ &\quad - \sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}^s) - \omega(\mathbf{x})). \end{aligned}$$

Eliminating the nonnegative term $\beta\|\mathbf{z}^k - \mathbf{x}\|^2/2$ on the left-hand side, and plugging $\mathbf{x} = \mathbf{x}' \in X'$ in (3.9) it follows that

$$(3.10) \quad \begin{aligned} t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}')) &\leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 - \sum_{s=0}^{k-1} d_s (\varphi(\mathbf{x}^s) - \varphi(\mathbf{x}')) \\ &\quad - \sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}^s) - \omega(\mathbf{x}')). \end{aligned}$$

Since $\mathbf{x}' \in X'$ we have that $\varphi(\mathbf{x}^s) - \varphi(\mathbf{x}') \geq 0$ for all $s \in \mathbb{N}$, and since $d_s \geq 0$ for all $s \in \mathbb{N}$, the desired result follows. $\square$

We conclude this part with the following result.

PROPOSITION 3.4. *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG. For all $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$, we have*

$$(3.11) \quad t_{k-1}^2 (\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}')) \leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + (\omega(\mathbf{x}') - \omega^*) \sum_{s=0}^{k-1} \alpha_s t_s.$$

Proof. Let $k \in \mathbb{N}$. For the simplicity of the proof we define the following non-negative quantities (recall that $\omega^* = \inf_{\mathbf{x} \in \mathbb{R}^n} \omega(\mathbf{x})$) $\Omega_{\max} := \omega(\mathbf{x}') - \omega^*$ and $\Omega_k := \omega(\mathbf{x}^k) - \omega^*$, $k \in \mathbb{N}$. Using the definition of $F_{k-1} \equiv \varphi + \alpha_{k-1}\omega$ in Proposition 3.3 (see (3.4)) we obtain

$$t_{k-1}^2 (\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}')) + \alpha_{k-1} t_{k-1}^2 (\Omega_k - \Omega_{\max}) \leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 - \sum_{s=0}^{k-1} \eta_s (\Omega_s - \Omega_{\max}).$$

Now, since $\eta_k > 0$ and $\Omega_k \geq 0$ for all $k \in \mathbb{N}$, it follows from the above inequality that

$$(3.12) \quad t_{k-1}^2 (\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}')) - \alpha_{k-1} t_{k-1}^2 \Omega_{\max} \leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \Omega_{\max} \sum_{s=0}^{k-1} \eta_s.$$

To complete the proof, note that from the definition of $\eta_k = \alpha_{k-1}t_{k-1}^2 - (t_k^2 - t_k)\alpha_k$, we have that

$$\begin{aligned}\sum_{s=0}^{k-1} \eta_s &= \sum_{s=0}^{k-1} (\alpha_{s-1}t_{s-1}^2 - \alpha_s t_s^2) + \sum_{s=0}^{k-1} \alpha_s t_s = \alpha_{-1}t_{-1}^2 - \alpha_{k-1}t_{k-1}^2 + \sum_{s=0}^{k-1} \alpha_s t_s \\ &= -\alpha_{k-1}t_{k-1}^2 + \sum_{s=0}^{k-1} \alpha_s t_s,\end{aligned}$$

where the last equality follows from the fact that $\alpha_{-1} = t_{-1} = 0$. Plugging it in (3.12), the desired result follows. $\square$

In this section, we established several results for FBi-PG that will serve as the foundation for our main convergence analysis. As noted earlier, our algorithm can be interpreted as a dynamic version of FISTA, where the objective functions vary with each iteration. This dynamic nature introduces significant challenges in the analysis, which deviates notably from the classical FISTA framework. A closer look at the basic inequality in Proposition 3.2 reveals that (3.1) nearly exhibits a telescoping structure. However, the difficulty arises from the iteration-dependent objective function F_k . In Proposition 3.3, we address this issue, but at the cost of introducing an additional term in (3.4), which subsequently implies a more involved analysis and influences the strength of the results we can obtain.

3.2. Rates of convergence of FBi-PG. In this part, we will prove three types of rate of convergence results: on the inner objective function, on the outer objective function, and simultaneously on both the inner and outer objective functions. All the results under the standard Assumption 2.1.

We begin with two technical lemmas to simplify the proofs afterwards. The first one was proved in [17, Lemma 4.3].

LEMMA 3.5. *Let $n_1 \in \mathbb{N}$ and $n_2 \in \mathbb{N} \cup \{\infty\}$ be such that $n_1 \leq n_2$. Then, for any $0 < r < 1$, we have*

$$\sum_{n=n_1}^{n_2} n^{-r} \leq \frac{n_2^{1-r} - (n_1 - 1)^{1-r}}{1 - r}.$$

The result remains true for $r > 1$ when $n_1 \geq 2$.

The proof of the following lemma is provided in Appendix C.

LEMMA 3.6. *For all $k \geq 1$, the following inequality holds true:*

$$(3.13) \quad \sum_{s=0}^{k-1} \alpha_s t_s \leq \begin{cases} \left((k+a-1)^{2-\gamma} - 1 \right) (2-\gamma)^{-1}, & 0 < \gamma < 2, \\ \ln(k+a-1), & \gamma = 2, \\ \left(1 - (k+a-1)^{2-\gamma} \right) (\gamma-2)^{-1}, & \gamma > 2. \end{cases}$$

3.2.1. Inner rates of convergence. We next establish an accelerated convergence rate of order $O(1/k^2)$ for the inner objective when $\gamma > 2$. While this result mirrors the optimal rate in single-level settings and appears to be the first fast inner rate result, it is worth noting that, in the absence of a corresponding outer guarantee, its bilevel relevance is primarily illustrative of the algorithm's behavior in this regime. To the best of our knowledge, achieving simultaneous results in the $\gamma > 2$ regime is an open problem.

THEOREM 3.7 (fast inner rate for $\gamma > 2$). *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG with $\gamma > 2$. For all $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$, the following inequality holds true:*

$$(3.14) \quad \varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') \leq \frac{a^2}{2(k+1)^2} \left(\beta \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{2}{\gamma-2} (\omega(\mathbf{x}') - \omega^*) \right).$$

Proof. From Proposition 3.4 (see (3.11)) we have that

$$t_{k-1}^2 (\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}')) \leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + (\omega(\mathbf{x}') - \omega^*) \sum_{s=0}^{k-1} \alpha_s t_s.$$

Multiplying both sides by $t_{k-1}^{-2} > 0$ yields that

$$(3.15) \quad \varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') \leq t_{k-1}^{-2} \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + (\omega(\mathbf{x}') - \omega^*) t_{k-1}^{-2} \sum_{s=0}^{k-1} \alpha_s t_s.$$

Now, from Lemma 3.6 with $\gamma > 2$, we have that

$$t_{k-1}^{-2} \sum_{s=0}^{k-1} \alpha_s t_s \leq \frac{t_{k-1}^{-2}}{\gamma-2} \left(1 - (k+a-1)^{2-\gamma} \right) \leq \frac{t_{k-1}^{-2}}{\gamma-2},$$

which combined with (3.15) yields that (recall that $\omega(\mathbf{x}') - \omega^* \geq 0$ since $\omega^* = \inf_{\mathbf{x} \in \mathbb{R}^n} \omega(\mathbf{x})$)

$$\begin{aligned} \varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') &\leq t_{k-1}^{-2} \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{t_{k-1}^{-2}}{\gamma-2} (\omega(\mathbf{x}') - \omega^*) \\ &= \frac{t_{k-1}^{-2}}{2} \left(\beta \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{2}{\gamma-2} (\omega(\mathbf{x}') - \omega^*) \right). \end{aligned}$$

The desired result now follows from the fact that $t_{k-1}^{-1} = a/(k+a-1) \leq a/(k+1)$, where the inequality follows from the fact that $a \geq 2$. $\square$

While our primary goal was to achieve a fast rate of convergence, the results obtained above also enable us to derive the following rates of convergence concerning the inner problem. These rates are instrumental for deriving the additional results, particularly simultaneous rates pertaining to *both* the inner and outer optimization problems.

PROPOSITION 3.8 (inner rate for $0 < \gamma \leq 2$). *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG. For all $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$, we have*

$$(3.16) \quad \varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') \leq \frac{a^2 \beta}{2(k+1)^2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{a^2 c_k}{(k+1)^2} (\omega(\mathbf{x}') - \omega^*),$$

where

$$(3.17) \quad c_k = \begin{cases} (2-\gamma)^{-1} (k+1)^{2-\gamma}, & 0 < \gamma < 2, \\ \ln(k+a-1), & \gamma = 2. \end{cases}$$

Proof. From (3.15), we obtain that

$$\begin{aligned}\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') &\leq t_{k-1}^{-2} \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + (\omega(\mathbf{x}') - \omega^*) t_{k-1}^{-2} \sum_{s=0}^{k-1} \alpha_s t_s \\ &\leq \frac{a^2 \beta}{2(k+1)^2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + (\omega(\mathbf{x}') - \omega^*) t_{k-1}^{-2} \sum_{s=0}^{k-1} \alpha_s t_s,\end{aligned}$$

where the last inequality follows from the fact that $t_{k-1}^{-1} = a/(k+a-1) \leq a/(k+1)$ since we used the fact that $a \geq 2$. To obtain the desired result, we split the proof to two cases. If $\gamma = 2$, we immediately get from Lemma 3.6 that

$$t_{k-1}^{-2} \sum_{s=0}^{k-1} \alpha_s t_s \leq t_{k-1}^{-2} \ln(k+a-1) = \frac{a^2 \ln(k+a-1)}{(k+a-1)^2} \leq \frac{a^2 \ln(k+a-1)}{(k+1)^2},$$

where the last inequality follows from the fact that $a \geq 2$. Moreover, when $0 < \gamma < 2$, we have from Lemma 3.6 that

$$t_{k-1}^{-2} \sum_{s=0}^{k-1} \alpha_s t_s \leq \frac{t_{k-1}^{-2} (k+a-1)^{2-\gamma} - t_{k-1}^{-2}}{2-\gamma} \leq \frac{a^2}{(2-\gamma)(k+a-1)^\gamma} \leq \frac{a^2}{(2-\gamma)(k+1)^\gamma},$$

where the second inequality follows from the definition of t_k (see (2.7)) after eliminating the term $t_{k-1}^2 \geq 0$ and the last inequality follows from the fact that $a \geq 2$. This concludes the proof. $\square$

3.2.2. Outer rate of convergence. Now, we can prove the upper rate of convergence in terms of the outer objective error.

THEOREM 3.9 (upper outer rate for $0 < \gamma < 2$). *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBI-PG with $\gamma \in (0, 2)$. For all $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$, the following inequality holds true:*

$$(3.18) \quad \min_{1 \leq s \leq k} \omega(\mathbf{x}^s) - \omega(\mathbf{x}') \leq \frac{a^2 \beta}{2(k+1)^{2-\gamma}} \|\mathbf{x}^0 - \mathbf{x}'\|^2.$$

Proof. For the simplicity of the proof, we denote $\bar{\eta}_k = 1/\sum_{s=1}^k \eta_s > 0$ (recall that $\eta_k > 0$ for all $k \in \mathbb{N}$; see (3.3)). We first prove the result under the assumption that $\bar{\eta}_{k-1} \sum_{s=1}^{k-1} \eta_s \omega(\mathbf{x}^s) - \omega(\mathbf{x}') \geq 0$. Indeed, in this case, we have that

$$0 \leq \bar{\eta}_{k-1} \sum_{s=1}^{k-1} \eta_s \omega(\mathbf{x}^s) - \omega(\mathbf{x}') = \bar{\eta}_{k-1} \sum_{s=1}^{k-1} \eta_s (\omega(\mathbf{x}^s) - \omega(\mathbf{x}')),$$

and therefore,

$$(3.19) \quad 0 \leq \sum_{s=1}^{k-1} \eta_s (\omega(\mathbf{x}^s) - \omega(\mathbf{x}')) = \sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}^s) - \omega(\mathbf{x}')),$$

where the equality follows from the fact that $\eta_0 = 0$. Using Proposition 3.3 (see (3.4)) combined with (3.19), we get that

$$t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}')) \leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 - \sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}^s) - \omega(\mathbf{x}')) \leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2.$$

From the definition of $F_{k-1} \equiv \varphi + \alpha_{k-1}\omega$ we get that

$$t_{k-1}^2 (\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}')) + \alpha_{k-1} t_{k-1}^2 (\omega(\mathbf{x}^k) - \omega(\mathbf{x}')) = t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}')) \\ \leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2,$$

and since $\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') \geq 0$ for all $k \in \mathbb{N}$, after dividing both sides by $t_{k-1}^2 \alpha_{k-1}$ we then obtain that

$$(3.20) \quad \omega(\mathbf{x}^k) - \omega(\mathbf{x}') \leq \alpha_{k-1}^{-1} t_{k-1}^{-2} \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 = \frac{a^2 \beta}{2(k+a-1)^{2-\gamma}} \|\mathbf{x}^0 - \mathbf{x}'\|^2,$$

where the last equality follows from the definitions of α_k and t_k (see (2.4) and (2.7), respectively). Recalling that $a \geq 2$ immediately proves (3.18) since obviously $\min_{1 \leq s \leq k} \omega(\mathbf{x}^s) \leq \omega(\mathbf{x}^k)$. Since the proof was done under the assumption that $\bar{\eta}_{k-1} \sum_{s=1}^{k-1} \eta_s \omega(\mathbf{x}^s) - \omega(\mathbf{x}') \geq 0$, we note that if

$$(3.21) \quad \bar{\eta}_{k-1} \sum_{s=1}^{k-1} \eta_s \omega(\mathbf{x}^s) - \omega(\mathbf{x}') \leq 0,$$

then we have that

$$\min_{1 \leq s \leq k} \omega(\mathbf{x}^s) - \omega(\mathbf{x}') \leq \min_{1 \leq s \leq k-1} \omega(\mathbf{x}^s) - \omega(\mathbf{x}') \leq \bar{\eta}_{k-1} \sum_{s=1}^{k-1} \eta_s \omega(\mathbf{x}^s) - \omega(\mathbf{x}') \leq 0,$$

and therefore, (3.18) is obviously true. $\square$

3.2.3. Simultaneous rates of convergence. Now, we prove a simultaneous ergodic rate of convergence for both the inner and outer optimization problems. Similarly to Theorem 3.9, for the outer optimization problem, the rate is an upper bound.

THEOREM 3.10 (simultaneous rates for $\gamma = 1$). *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG with $\gamma = 1$. For all $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$, the following inequalities hold true:*

$$(3.22) \quad \varphi(\bar{\mathbf{x}}^k) - \varphi(\mathbf{x}') \leq \frac{\pi^2 a^2 \beta}{12k} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{a^2 \ln(k+1)}{k} (\omega(\mathbf{x}') - \omega^*),$$

and

$$(3.23) \quad \omega(\bar{\mathbf{x}}^k) - \omega(\mathbf{x}') \leq \frac{a^2 \beta}{2(k+1)} \|\mathbf{x}^0 - \mathbf{x}'\|^2,$$

where $\bar{\mathbf{x}}^k = \operatorname{argmin} \{\omega(\mathbf{x}) : \mathbf{x} \in \{\mathbf{x}^k, \bar{\mathbf{x}}^k\}\}$ and $\bar{\mathbf{x}}^k = \sum_{s=1}^k \mathbf{x}^s / k$.

Proof. In order to prove (3.22), for all $1 \leq s \leq k$, we get from Proposition 3.8 (see (3.16)) with $k = s$ that (recall that $\gamma = 1$, and hence $c_s \equiv s + 1$)

$$(3.24) \quad \varphi(\mathbf{x}^s) - \varphi(\mathbf{x}') \leq \frac{a^2 \beta}{2(s+1)^2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{a^2}{s+1} (\omega(\mathbf{x}') - \omega^*).$$

Therefore, from the convexity of φ we obtain that

$$\begin{aligned}\varphi(\bar{\mathbf{x}}^k) - \varphi(\mathbf{x}') &\leq \frac{1}{k} \sum_{s=1}^k (\varphi(\mathbf{x}^s) - \varphi(\mathbf{x}')) \\ &\leq \frac{1}{k} \sum_{s=1}^k \left(\frac{a^2 \beta}{2(s+1)^2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{a^2}{s+1} (\omega(\mathbf{x}') - \omega^*) \right) \\ &= \frac{1}{k} \left(\frac{a^2 \beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 \sum_{s=2}^{k+1} \frac{1}{s^2} + a^2 (\omega(\mathbf{x}') - \omega^*) \sum_{s=2}^{k+1} \frac{1}{s} \right).\end{aligned}$$

Now, using classical results we have that $\sum_{s=2}^{k+1} s^{-2} < \pi^2/6$ and $\sum_{s=2}^{k+1} s^{-1} \leq \ln(k+1)$, which implies that

$$(3.25) \quad \varphi(\bar{\mathbf{x}}^k) - \varphi(\mathbf{x}') \leq \frac{\pi^2 a^2 \beta}{12k} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{a^2 \ln(k+1)}{k} (\omega(\mathbf{x}') - \omega^*).$$

Again from Proposition 3.8 (see (3.16)) and recall that $\gamma = 1$ (hence $c_k \equiv k+1$), we obtain that

$$\begin{aligned}\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') &\leq \frac{a^2 \beta}{2(k+1)^2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{a^2}{k+1} (\omega(\mathbf{x}') - \omega^*) \\ (3.26) \quad &\leq \frac{\pi^2 a^2 \beta}{12k} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{a^2 \ln(k+1)}{k} (\omega(\mathbf{x}') - \omega^*),\end{aligned}$$

where the last inequality follows from the fact that $1/2 < \pi^2/12$ and thanks to the convexity of the function $t \ln(t)$ for $t \geq 0$ we have that $k < (k+1) \ln(k+1)$. This concludes the proof of (3.22) due to the definition of $\tilde{\mathbf{x}}^k$ by combining (3.25) and (3.26).

Now, in order to prove (3.23), since $\gamma = 1$ and with $t_k = (k+a)/a$ (see (2.7)) we get from the definition of α_k (see (2.4)) that

$$(3.27) \quad \eta_k = \alpha_{k-1} t_{k-1}^2 - \alpha_k t_k^2 + \alpha_k t_k = \frac{k+a-1}{a^2} - \frac{k+a}{a^2} + \frac{1}{a} = \frac{a-1}{a^2}.$$

Hence, $\bar{\eta}_k := 1/\sum_{s=1}^k \eta_k = a^2/(k(a-1))$ for all $k \in \mathbb{N}$, and therefore, $\bar{\mathbf{x}}^k = \sum_{s=1}^k \mathbf{x}^s/k = \bar{\eta}_k \sum_{s=1}^k \eta_s \mathbf{x}^s$.

Following the proof of Theorem 3.9 (see (3.20) and (3.21)), we easily obtain that

$$(3.28) \quad \min \left\{ \omega(\mathbf{x}^k), \bar{\eta}_k \sum_{s=1}^k \eta_s \omega(\mathbf{x}^s) \right\} - \omega(\mathbf{x}') \leq \frac{a^2 \beta}{2(k+1)} \|\mathbf{x}^0 - \mathbf{x}'\|^2.$$

Moreover, from the convexity of ω , we get that

$$\omega(\bar{\mathbf{x}}^k) \leq \frac{1}{k} \sum_{s=1}^k \omega(\mathbf{x}^s) = \frac{a^2}{k(a-1)} \sum_{s=1}^k \eta_s \omega(\mathbf{x}^s) = \bar{\eta}_k \sum_{s=1}^k \eta_s \omega(\mathbf{x}^s),$$

where the first and last equalities follow from the facts that $\eta_k = (a-1)/a^2$ and $\bar{\eta}_k = a^2/(k(a-1))$, respectively. Integrating it into (3.28) yields that

$$\begin{aligned}\min \{ \omega(\mathbf{x}^k), \omega(\bar{\mathbf{x}}^k) \} - \omega(\mathbf{x}') &\leq \min \left\{ \omega(\mathbf{x}^k), \bar{\eta}_k \sum_{s=1}^k \eta_s \omega(\mathbf{x}^s) \right\} - \omega(\mathbf{x}') \\ &\leq \frac{a^2 \beta}{2(k+1)} \|\mathbf{x}^0 - \mathbf{x}'\|^2.\end{aligned}$$

The desired result now follows immediately from the definition of $\tilde{\mathbf{x}}^k$. $\square$

3.3. Improved simultaneous rates. This section provides simultaneous convergence rates that enjoy three benefits in comparison to the simultaneous rates provided in section 3.2 (see Theorem 3.10). First, the rates are on the last-iterate rather than on the best-ergodic iterate. Second, the order of convergence is improved in this section. Third, the new simultaneous rates include an outer lower convergence rate too (see Table 1). To obtain the improved simultaneous rates of convergence, we will impose a Hölderian type error bound assumption on the inner objective function φ . For more details and examples of functions satisfying such a property, we refer the interested readers to [6]. At this juncture³ we note that [1] and, very recently, [7] examined conditions weaker than the Hölderian error bound in their study of bilevel optimization problems in Hilbert spaces via continuous-time dynamics. In addition, [7] also investigated accelerated methods and their asymptotic convergence.

We recall now the definition of Hölderian type error bound.

DEFINITION 3.11 (Hölderian type EB). *A proper, lower semicontinuous, and convex function $h : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ satisfies an Hölderian error bound with a parameter $\tau > 0$ if for all $\mathbf{x} \in \mathbb{R}^n$*

$$(3.29) \quad \tau \cdot \text{dist}(\mathbf{x}, \text{argmin } h)^2 \leq h(\mathbf{x}) - \min_{\mathbf{z} \in \mathbb{R}^n} h(\mathbf{z}).$$

Throughout this section, we make the following additional assumption.

Assumption 3.12. The following items hold:

- (i) φ satisfies the Hölderian error bound with a parameter $\tau > 0$.
- (ii) $\text{ri}(\text{dom } \omega) \cap \text{ri } X' \neq \emptyset$.
- (iii) $\omega(\mathbf{x}') < \infty$ for $\mathbf{x}' \in X'$.

Item (iii) is a very mild and natural requirement. Because $X' \subseteq X^*$, the condition simply asserts that the outer objective remains finite on the set of feasible lower-level solutions. Any well-posed bi-level optimization model must already satisfy this; otherwise, the upper-level problem would be ill-defined.

Before establishing the promised improved simultaneous rates of convergence, we first need the following simple bound for η_k , $k \in \mathbb{N}$, which will be used throughout the section. The proof of this result is provided in Appendix C.

LEMMA 3.13. $1 < \gamma < 2$. *For all $k \in \mathbb{N}$, one has*

$$(3.30) \quad \eta_k < \frac{1}{2}(k+1)^{1-\gamma}.$$

In the next theorem, we establish the improved simultaneous convergence rates.

THEOREM 3.14 (simultaneous rates under Hölderian EB for $1 < \gamma < 2$). *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG with $1 < \gamma < 2$. For all $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$, the following inequalities hold true:*

- (i) $\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') \leq \frac{aC}{(k+1)^2}$,
- (ii) $\omega(\mathbf{x}') - \omega(\mathbf{x}^k) \leq \frac{C}{a^2(k+1)}$,
- (iii) $\omega(\mathbf{x}^k) - \omega(\mathbf{x}') \leq \frac{C}{(k+1)^{2-\gamma}}$,

where $C = a^2 \max\{\beta \|\mathbf{x}^0 - \mathbf{x}'\|^2 + (\omega(\mathbf{x}') - \omega^*), a^3 \rho^2 \tau^{-1} (\gamma - 1)^{-2}\}$ and $\rho = \|\xi\|$ for some $\xi \in \partial \omega(\mathbf{x}')$.

³We thank a referee for pointing out to us the work [1], and the very recently posted Arxiv manuscript [7].

Proof. For the simplicity of the proof, we denote $\Phi_k = \varphi(\mathbf{x}^k) - \varphi(\mathbf{x}')$ for all $k \in \mathbb{N}$. We will prove the first item using an induction argument on $k \in \mathbb{N}$. For $k = 1$, we obtain from Proposition 3.4 that (recall that $t_0 = 1$)

$$\Phi_1 = t_0^{-1} (\varphi(\mathbf{x}^1) - \varphi(\mathbf{x}')) \leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + (\omega(\mathbf{x}') - \omega^*) \alpha_0 t_0 \leq \frac{C}{2a^2} \leq \frac{aC}{4},$$

where the second inequality follows from the definition of C and the fact that $a_0 = a^{-\gamma} < 1/2$ since $\gamma > 1$ and $a \geq 2$, and the last inequality also follows from the fact that $a \geq 2$. This proves the induction base with $k = 1$.

Let $k \in \mathbb{N}$. We assume now that item (i) is true for all $1 \leq s \leq k-1$ and we will prove it for $s = k$. For all $1 \leq s \leq k-1$, we have that

$$(3.31) \quad \xi^T(\mathbf{x}' - \mathbf{x}^s) = \xi^T(\mathbf{x}' - P_{X^*}(\mathbf{x}^s)) + \xi^T(P_{X^*}(\mathbf{x}^s) - \mathbf{x}^s) \leq \xi^T(P_{X^*}(\mathbf{x}^s) - \mathbf{x}^s),$$

where P_{X^*} denotes the orthogonal projection onto the nonempty set X^* . It should be noted that the last inequality follows from the first-order optimality condition of the minimization of $\omega + \delta_{X^*}$ (which is convex, lsc, and proper, thanks to Assumption 3.12(ii) which implies that $\omega(\mathbf{x}') + \delta_{X^*}(\mathbf{x}') < \infty$), which reads (recall that $X' = \operatorname{argmin}\{\omega(\mathbf{x}) : \mathbf{x} \in X^*\}$):

$$\mathbf{0} \in \partial(\omega + \delta_{X^*})(\mathbf{x}') = \xi + \partial\delta_{X^*}(\mathbf{x}') = \xi + N_{X^*}(\mathbf{x}'),$$

where $\xi \in \partial\omega(\mathbf{x}')$, and the first equality above holds thanks to the standard constraint qualification (cf. Assumption B (ii)). Therefore, with the definition of the normal cone $N_{X^*}(\cdot)$ it follows that $\xi^T(\mathbf{x}' - \mathbf{u}) \leq 0$ for all $\mathbf{u} \in X^*$. Therefore, since $P_{X^*}(\mathbf{x}^s) \in X^*$, one has $\xi^T(\mathbf{x}' - P_{X^*}(\mathbf{x}^s)) \leq 0$. Now, using the Hölderian error bound of the function φ (see Assumption 3.12(i)) yields

$$(3.32) \quad \xi^T(P_{X^*}(\mathbf{x}^s) - \mathbf{x}^s) \leq \|\xi\| \cdot \|P_{X^*}(\mathbf{x}^s) - \mathbf{x}^s\| = \rho \cdot \operatorname{dist}(\mathbf{x}^s, X^*) \leq \frac{\rho}{\sqrt{\tau}} \sqrt{\Phi_s},$$

where the equality follows from the facts that $\|\xi\| = \rho$ and the definition of the distance function. Hence, using the gradient inequality on the convex function ω , we obtain that (recall that $\xi \in \partial\omega(\mathbf{x}')$)

$$(3.33) \quad \omega(\mathbf{x}') - \omega(\mathbf{x}^s) \leq \xi^T(\mathbf{x}' - \mathbf{x}^s) \leq \xi^T(P_{X^*}(\mathbf{x}^s) - \mathbf{x}^s) \leq \frac{\rho}{\sqrt{\tau}} \sqrt{\Phi_s},$$

where the second inequality follows from (3.31) and the last inequality follows from (3.32). Using the induction assumption on Φ_s for all $1 \leq s \leq k-1$, we have thus obtained

$$(3.34) \quad \omega(\mathbf{x}') - \omega(\mathbf{x}^s) \leq \frac{\rho}{\sqrt{\tau}} \sqrt{\Phi_s} \leq \frac{\rho\sqrt{a}\sqrt{C}}{\sqrt{\tau}} (s+1)^{-1}.$$

Therefore, using (3.30) and (3.34), we obtain that (recall that $\eta_0 = 0$)

$$\begin{aligned} \sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}') - \omega(\mathbf{x}^s)) &= \sum_{s=1}^{k-1} \eta_s (\omega(\mathbf{x}') - \omega(\mathbf{x}^s)) \leq \frac{\rho\sqrt{a}\sqrt{C}}{2\sqrt{\tau}} \sum_{s=1}^{k-1} (s+1)^{-\gamma} \\ &= \frac{\rho\sqrt{a}\sqrt{C}}{2\sqrt{\tau}} \sum_{s=2}^k s^{-\gamma}. \end{aligned}$$

Using Lemma 3.5 with $n_1 = 2$, $n_2 = k$, and $r = \gamma > 1$, we have

$$\begin{aligned} \sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}') - \omega(\mathbf{x}^s)) &\leq \frac{\rho\sqrt{a}\sqrt{C}}{2\sqrt{\tau}} \sum_{s=2}^k s^{-\gamma} = \frac{\rho\sqrt{a}\sqrt{C}}{2\sqrt{\tau}} \cdot \frac{1^{1-\gamma} - k^{1-\gamma}}{\gamma - 1} \\ (3.35) \qquad \qquad \qquad &\leq \frac{\rho\sqrt{a}\sqrt{C}}{2\sqrt{\tau}(\gamma - 1)}. \end{aligned}$$

From Proposition 3.3 and (3.35), we get that

$$\begin{aligned} t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}')) &\leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 - \sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}^s) - \omega(\mathbf{x}')) \\ &\leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \frac{\rho\sqrt{a}\sqrt{C}}{2\sqrt{\tau}(\gamma - 1)}. \end{aligned}$$

Hence, from the definition of C , we get

$$(3.36) \qquad t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}')) \leq \frac{C}{2a^2} + \frac{\sqrt{C}}{2} \cdot \frac{\sqrt{C}}{a^2} = \frac{C}{a^2}.$$

Moreover, using (3.33) with $s = k$, we have that

$$\omega(\mathbf{x}^k) - \omega(\mathbf{x}') \geq -\frac{\rho}{\sqrt{\tau}} \sqrt{\Phi_k}.$$

Using this, the definition of $F_{k-1} \equiv \varphi + \alpha_{k-1}\omega$ and (3.36) yields that

$$\begin{aligned} \Phi_k - \frac{\rho\alpha_{k-1}}{\sqrt{\tau}} \sqrt{\Phi_k} &\leq \Phi_k + \alpha_{k-1} (\omega(\mathbf{x}^k) - \omega(\mathbf{x}')) = F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}') \leq \frac{Ct_{k-1}^{-2}}{a^2} \\ (3.37) \qquad \qquad \qquad &\leq \frac{C}{(k+1)^2}, \end{aligned}$$

where the last inequality follows from the fact that $t_{k-1}^{-1} = a/(k+a-1) \leq a/(k+1)$ since $a \geq 2$. Now, we will prove the induction step by splitting the proof into two cases. First, we assume that

$$\frac{\Phi_k}{a} \leq \Phi_k - \frac{\rho\alpha_{k-1}}{\sqrt{\tau}} \sqrt{\Phi_k}.$$

Using (3.37) yields that

$$\frac{\Phi_k}{a} \leq \Phi_k - \frac{\rho\alpha_{k-1}}{\sqrt{\tau}} \sqrt{\Phi_k} \leq \frac{C}{(k+1)^2},$$

and therefore, we obviously have that $\Phi_k \leq aC/(k+1)^2$, as required. Second, we assume the converse, that is

$$\Phi_k - \frac{\rho\alpha_{k-1}}{\sqrt{\tau}} \sqrt{\Phi_k} < \frac{\Phi_k}{a}.$$

Therefore, by rearranging the inequality we get that $\Phi_k(a-1)/a < (\rho\alpha_{k-1}/\sqrt{\tau}) \sqrt{\Phi_k}$. Dividing both sides by $(a-1) \sqrt{\Phi_k}/a$ and then squaring both sides yields

$$\Phi_k < \frac{a^2 \rho^2}{\tau(a-1)^2} \alpha_{k-1}^2 \leq \frac{a^2 \rho^2}{\tau(a-1)^2} (k+1)^{-2} \leq \frac{a^2 \rho^2}{\tau(\gamma-1)^2} (k+1)^{-2} \leq \frac{C}{a^3 (k+1)^2},$$

where the second inequality follows from the definition of $\alpha_{k-1} = (k+a-1)^{-\gamma} < (k+1)^{-1}$ since $\gamma > 1$ and $a \geq 2$, the third inequality follows from the fact that $0 < \gamma-1 < a-1$ since $a \geq 2$ and $1 < \gamma < 2$ while the last inequality follows again from the definition of C . Thus, we obtain that $\Phi_k \leq aC/(k+1)^2$ since $a \geq 2$, which proves the desired result.

Now, we will prove the second item. Using (3.34) with $s = k$, yields

$$\omega(\mathbf{x}') - \omega(\mathbf{x}^k) \leq \frac{\rho\sqrt{a}\sqrt{C}}{\sqrt{\tau}}(k+1)^{-1} \leq \frac{\rho\sqrt{a}\sqrt{C}}{\sqrt{\tau}(\gamma-1)}(k+1)^{-1} \leq \frac{C}{a^2}(k+1)^{-1},$$

where the second inequality follows from the fact that $1 < (\gamma-1)^{-1}$ since $1 < \gamma < 2$ and the last inequality follows from the definition of C . For proving the last item, using (3.36), we have that

$$\alpha_{k-1}(\omega(\mathbf{x}^k) - \omega(\mathbf{x}')) \leq \Phi_k + \alpha_{k-1}(\omega(\mathbf{x}^k) - \omega(\mathbf{x}')) = F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}') \leq \frac{Ct_{k-1}^{-2}}{a^2},$$

where the first inequality follows from the fact that $\Phi_k \geq 0$ and the equality follows from the definition of $F_{k-1} \equiv \varphi + \alpha_{k-1}\omega$. Hence, immediately from the definitions of α_k and t_k (see (2.4) and (2.7), respectively) we obtain that

$$\omega(\mathbf{x}^k) - \omega(\mathbf{x}') \leq \frac{Ct_{k-1}^{-2}}{a^2\alpha_{k-1}} = \frac{C}{(k+a-1)^{2-\gamma}} \leq \frac{C}{(k+1)^{2-\gamma}},$$

where the last inequality follows from the fact that $a-1 \geq 1$ since $a \geq 2$. This proves the desired result. $\square$

Before concluding this section, we note that, as explained in Remark 2.3, applying FBI-PG to the lifted problem avoids the need to compute the proximal mapping of the sum of two functions. In the following remark, we record the effect on the convergence rates using the lifted problem.

Remark 3.15. Following Remark 2.3, applying any of our convergence rate results to the lifted inner objective $\bar{\varphi}$ yields the bound $\bar{\varphi}(\mathbf{w}^k) - \bar{\varphi}(\mathbf{w}') \leq \mathcal{R}_k^I$, where $\mathcal{R}_k^I$ denotes any of the upper bound (which depends of the choice of γ) as derived in Theorems 3.7 and 3.10 and Proposition 3.8. As a consequence, recalling that $\bar{\varphi}(\mathbf{w}') = \varphi(\mathbf{x}')$, it is immediate to deduce that

$$\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') \leq \mathcal{R}_k^I \quad \text{and} \quad \|\mathbf{x}^k - \mathbf{z}^k\| \leq \sqrt{\mathcal{R}_k^I}.$$

This demonstrates that the lifting approach enables us to completely circumvent the need for computing the proximal mapping of the composite term $\sigma + \psi$, while still retaining the original convergence rates for the inner problem. The situation for the outer objective is more subtle. While at optimality we have $\bar{\omega}(\mathbf{w}') = \omega(\mathbf{x}')$ (since $\mathbf{x}' = \mathbf{z}'$), this equality does not hold, in general. This can be circumvented by assuming that ψ is L_ψ -Lipschitz continuous. Applying our convergence rate results in the lifted objective $\bar{\omega}$ yielding the bound $\bar{\omega}(\mathbf{w}^k) - \bar{\omega}(\mathbf{w}') \leq \mathcal{R}_k^O$, where $\mathcal{R}_k^O$ depends on γ as established in Theorems 3.9 and 3.10. It is then easy to show that the outer rate for the original outer objective function satisfies

$$\omega(\mathbf{x}^k) - \omega(\mathbf{x}') \leq \mathcal{R}_k^O + L_\psi \|\mathbf{x}^k - \mathbf{z}^k\| \leq \mathcal{R}_k^O + L_\psi \sqrt{\mathcal{R}_k^I}.$$

In summary, by adopting the lifting approach, we eliminate the computational bottleneck associated with evaluating the proximal mapping of a composite term, while preserving the inner convergence rates derived in our analysis. For the outer problem, we must pay a price, which is expressed in terms of an additional Lipschitz continuity assumption, and a potentially weaker rate, due to the added Lipschitz error term.

4. Pointwise convergence analysis. This section is devoted for proving that the sequence $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ generated by FBi-PG converges to an optimal solution of the bilevel optimization problem (OP). This result is achieved under Assumption 3.12, meaning in the setting where the inner objective function $\varphi(\cdot)$ satisfies the Hölderian error bound property. Moreover, throughout this section we assume that the sequences $\{\alpha_k\}_{k \in \mathbb{N}}$ and $\{t_k\}_{k \in \mathbb{N}}$ are defined with $1 < \gamma < 2$ and $a > 2$.

As already mentioned, each iteration of FBi-PG can be seen as an iteration of FISTA applied to the regularized function F_k . Unfortunately, due to the dynamic nature of the regularizing sequence $\{\alpha_k\}_{k \in \mathbb{N}}$, much like we could not apply the convergence rate results of FISTA to get corresponding fast rate for FBi-PG, we also cannot directly apply the existing pointwise convergence results of FISTA established in [9] to derive the corresponding convergence of the sequence $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ generated by FBi-PG to an optimal solution of the bilevel optimization problem (OP). Nevertheless, inspired by the proof techniques developed in [9], below we show that within a similar line of proof, we can adapt and extend their result for handling the bilevel setting.

Throughout the rest of this section, we will define for all $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$ the following two quantities:

$$(4.1) \quad \delta_k = \frac{1}{2} \|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 \quad \text{and} \quad \mu_k = \frac{1}{2} \|\mathbf{x}^k - \mathbf{x}'\|^2.$$

The main tasks of the forthcoming analysis will be to establish that both sequences $\{\delta_k\}_{k \in \mathbb{N}}$ and $\{\mu_k\}_{k \in \mathbb{N}}$ converge, from which the culminating main convergence result will follow by standard arguments.

We begin with a technical lemma (see Appendix B for the proof), which will be useful in proving the main result.

LEMMA 4.1. *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG with $1 < \gamma < 2$ and $a > 2$. Then, there exists $M > 0$ such that $\sum_{s=0}^{k-1} t_{s-1} \delta_s < M$ for all $k \in \mathbb{N}$.*

For all $s \in \mathbb{N}$, we define $\pi_{s,k} = 1$ for $k < s$ and otherwise we define

$$(4.2) \quad \pi_{s,k} := \prod_{j=s}^k \lambda_j,$$

where $\lambda_k := t_k^{-1}(t_{k-1} - 1)$ for all $k \in \mathbb{N}$.

The following bound is based on a result obtained in [9] with a minor modification to our setting. Its proof is provided in Appendix C.

LEMMA 4.2. *For all $s \geq 1$ we have that*

$$\sum_{k=0}^{\infty} \pi_{s,k} \leq \frac{5a}{2} t_{s-1}.$$

The next theorem proves that the FBi-PG's sequence, $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$, is convergent to an optimal solution of the bilevel problem (OP).

THEOREM 4.3 (pointwise convergence under Hölderian EB for $1 < \gamma < 2$). *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG with $1 < \gamma < 2$ and $a > 2$. Then, the sequence $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ converges to some $\mathbf{x}' \in X'$.*

Proof. Let $\mathbf{x}' \in X'$. Since $X' \subset X^*$, it follows that $\mathbf{0} \in \partial\varphi(\mathbf{x}') = \nabla f(\mathbf{x}') + \partial g(\mathbf{x}')$. Therefore, $-\nabla f(\mathbf{x}') \in \partial g(\mathbf{x}')$. Taking $\mathbf{v}' \in \partial\psi(\mathbf{x}')$, it follows that

$$(4.3) \quad -\nabla f(\mathbf{x}') + \alpha_k \mathbf{v}' \in \partial g(\mathbf{x}') + \alpha_k \partial\psi(\mathbf{x}') \subset \partial(g + \alpha_k \psi)(\mathbf{x}') = \partial g_k(\mathbf{x}'),$$

where the inclusion follows from standard convex subdifferential calculus [3] and the last equality follows from the definition of $g_k \equiv g + \alpha_k \psi$. Writing the optimality condition of the proximal step (2.6) yields that

$$\beta (\mathbf{y}^k - \mathbf{x}^{k+1}) - \nabla f_k (\mathbf{y}^k) \in \partial g_k (\mathbf{x}^{k+1}).$$

Now, since the function g_k is convex, it follows that its subdifferential ∂g_k is a monotone mapping [3]. Therefore, we have that

$$\begin{aligned} 0 &\leq \langle \beta (\mathbf{y}^k - \mathbf{x}^{k+1}) - \nabla f_k (\mathbf{y}^k) - (-\nabla f (\mathbf{x}') + \alpha_k \mathbf{v}'), \mathbf{x}^{k+1} - \mathbf{x}' \rangle \\ &= \langle \beta (\mathbf{y}^k - \mathbf{x}^{k+1}) - \nabla f_k (\mathbf{y}^k) + \nabla f_k (\mathbf{x}') - \alpha_k (\nabla \sigma (\mathbf{x}') + \mathbf{v}'), \mathbf{x}^{k+1} - \mathbf{x}' \rangle \\ &= \beta \langle \mathbf{y}^k - \mathbf{x}^{k+1}, \mathbf{x}^{k+1} - \mathbf{x}' \rangle + \langle \nabla f_k (\mathbf{x}') - \nabla f_k (\mathbf{y}^k), \mathbf{x}^{k+1} - \mathbf{x}' \rangle \\ (4.4) \quad &+ \alpha_k \langle \nabla \sigma (\mathbf{x}') + \mathbf{v}', \mathbf{x}' - \mathbf{x}^{k+1} \rangle, \end{aligned}$$

where the first equality follows from the fact that $f_k \equiv f + \alpha_k \sigma$. Since f_k is convex and has a Lipschitz continuous gradient with a parameter β (see Remark 2.2), the latter is also $(1/\beta)$ -co-coercive mapping and hence we have

$$\begin{aligned} \langle \nabla f_k (\mathbf{x}') - \nabla f_k (\mathbf{y}^k), \mathbf{x}^{k+1} - \mathbf{x}' \rangle &= \langle \nabla f_k (\mathbf{x}') - \nabla f_k (\mathbf{y}^k), \mathbf{x}^{k+1} - \mathbf{y}^k + \mathbf{y}^k - \mathbf{x}' \rangle \\ &\leq \langle \nabla f_k (\mathbf{x}') - \nabla f_k (\mathbf{y}^k), \mathbf{x}^{k+1} - \mathbf{y}^k \rangle \\ &\quad - \frac{1}{\beta} \|\nabla f_k (\mathbf{x}') - \nabla f_k (\mathbf{y}^k)\|^2 \\ &\leq \frac{1}{\beta} \|\nabla f_k (\mathbf{x}') - \nabla f_k (\mathbf{y}^k)\|^2 + \frac{\beta}{4} \|\mathbf{x}^{k+1} - \mathbf{y}^k\|^2 \\ &\quad - \frac{1}{\beta} \|\nabla f_k (\mathbf{x}') - \nabla f_k (\mathbf{y}^k)\|^2 \\ (4.5) \quad &= \frac{\beta}{4} \|\mathbf{x}^{k+1} - \mathbf{y}^k\|^2, \end{aligned}$$

where the last inequality follows from the fact that $\mathbf{u}^T \mathbf{w} \leq \|\mathbf{u}\|^2 / \beta + \beta \|\mathbf{w}\|^2 / 4$ for all $\mathbf{u}, \mathbf{w} \in \mathbb{R}^n$ and $\beta > 0$. Combining (4.4) and (4.5) yields after dividing both sides by β that

$$(4.6) \quad 0 \leq \langle \mathbf{y}^k - \mathbf{x}^{k+1}, \mathbf{x}^{k+1} - \mathbf{x}' \rangle + \frac{1}{4} \|\mathbf{x}^{k+1} - \mathbf{y}^k\|^2 + \xi^{k+1},$$

where $\xi^k \equiv \alpha_{k-1} \langle \nabla \sigma (\mathbf{x}') + \mathbf{v}', \mathbf{x}' - \mathbf{x}^k \rangle / \beta$ for all $k \in \mathbb{N}$. Moreover, using the fundamental Pythagoras identity, we obtain that

$$\begin{aligned} \langle \mathbf{y}^k - \mathbf{x}^{k+1}, \mathbf{x}^{k+1} - \mathbf{x}' \rangle &= \langle \mathbf{y}^k - \mathbf{x}^k, \mathbf{x}^{k+1} - \mathbf{x}' \rangle + \langle \mathbf{x}^k - \mathbf{x}^{k+1}, \mathbf{x}^{k+1} - \mathbf{x}' \rangle \\ &= \langle \mathbf{y}^k - \mathbf{x}^k, \mathbf{x}^{k+1} - \mathbf{x}' \rangle \\ &\quad + \frac{1}{2} \left(\|\mathbf{x}^k - \mathbf{x}'\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}'\|^2 - \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \right) \\ (4.7) \quad &= \langle \mathbf{y}^k - \mathbf{x}^k, \mathbf{x}^{k+1} - \mathbf{x}' \rangle + \mu_k - \mu_{k+1} - \delta_{k+1}, \end{aligned}$$

where the third equality follows from the definitions of δ_k and μ_k (see (4.1)). Since $\lambda_k = t_k^{-1}(t_{k-1} - 1)$ we can write from (2.5) that $\mathbf{y}^k - \mathbf{x}^k = \lambda_k (\mathbf{x}^k - \mathbf{x}^{k-1})$. Using again the fundamental Pythagoras identity we obtain that

$$\begin{aligned} \langle \mathbf{y}^k - \mathbf{x}^k, \mathbf{x}^{k+1} - \mathbf{x}' \rangle &= \langle \mathbf{y}^k - \mathbf{x}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \langle \mathbf{y}^k - \mathbf{x}^k, \mathbf{x}^k - \mathbf{x}' \rangle \\ &= \langle \mathbf{y}^k - \mathbf{x}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \lambda_k \langle \mathbf{x}^k - \mathbf{x}^{k-1}, \mathbf{x}^k - \mathbf{x}' \rangle \\ (4.8) \quad &= \langle \mathbf{y}^k - \mathbf{x}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \lambda_k (\mu_k - \mu_{k-1} + \delta_k). \end{aligned}$$

Another application of the Pythagoras identity yields that

$$\begin{aligned} \langle \mathbf{y}^k - \mathbf{x}^k, \mathbf{x}^{k+1} - \mathbf{x}^k \rangle &= \frac{1}{2} \left(\|\mathbf{y}^k - \mathbf{x}^k\|^2 - \|\mathbf{x}^{k+1} - \mathbf{y}^k\|^2 + \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 \right) \\ (4.9) \qquad \qquad \qquad &= \lambda_k^2 \delta_k - \frac{1}{2} \|\mathbf{x}^{k+1} - \mathbf{y}^k\|^2 + \delta_{k+1}, \end{aligned}$$

where the last equality uses the fact that $\mathbf{y}^k - \mathbf{x}^k = \lambda_k (\mathbf{x}^k - \mathbf{x}^{k-1})$ and the definition of δ_k . Combining (4.6), (4.7), (4.8), and (4.9) yields after eliminating $-\|\mathbf{x}^{k+1} - \mathbf{y}^k\|^2/4 \leq 0$ that

$$\begin{aligned} \mu_{k+1} - \mu_k &\leq \lambda_k (\mu_k - \mu_{k-1} + \delta_k) + \lambda_k^2 \delta_k + \xi^{k+1} \\ (4.10) \qquad \qquad &= \lambda_k (\mu_k - \mu_{k-1}) + (\lambda_k^2 + \lambda_k) \delta_k + \xi^{k+1}. \end{aligned}$$

Using the fact that $\lambda_k + \lambda_k^2 \leq 2\lambda_k$ since $\lambda_k^2 \leq \lambda_k$ (recall that $\lambda_k \leq 1$ for all $k \in \mathbb{N}$), we obtain

$$(4.11) \qquad \qquad \mu_{k+1} - \mu_k \leq \lambda_k (\mu_k - \mu_{k-1}) + 2\lambda_k \delta_k + \xi^{k+1}.$$

By defining $\theta_k = [\mu_k - \mu_{k-1}]_+ := \max\{0, \mu_k - \mu_{k-1}\}$ for all $k \geq 1$, it follows from (4.11) that

$$\begin{aligned} \theta_{k+1} &= [\mu_{k+1} - \mu_k]_+ \\ &\leq [\lambda_k (\mu_k - \mu_{k-1}) + 2\lambda_k \delta_k + \xi^{k+1}]_+ \\ &\leq \lambda_k \theta_k + 2\lambda_k [\delta_k]_+ + [\xi^{k+1}]_+ \\ (4.12) \qquad \qquad &= \lambda_k (\theta_k + 2\delta_k) + [\xi^{k+1}]_+, \end{aligned}$$

where the first inequality follows from the fact that $[x]_+ \leq [y]_+$ when $x \leq y$, the second inequality follows from the fact that $[x + y]_+ \leq [x]_+ + [y]_+$, and the equality follows from the fact that $\lambda_k, \delta_k \geq 0$. Thus, applying (4.12) successively yields, for all $k \geq 2$, that (we use the convention that $\prod_{j=k+1}^k \lambda_j = 1$)

$$\begin{aligned} \theta_{k+1} &\leq \lambda_k \theta_k + 2\lambda_k \delta_k + [\xi^{k+1}]_+ \\ &\leq \lambda_k \left(\lambda_{k-1} (\theta_{k-1} + 2\delta_{k-1}) + [\xi^k]_+ \right) + 2\lambda_k \delta_k + [\xi^{k+1}]_+ \\ &= \lambda_k \lambda_{k-1} \theta_{k-1} + 2\lambda_k \lambda_{k-1} \delta_{k-1} + 2\lambda_k \delta_k + \lambda_k [\xi^k]_+ + [\xi^{k+1}]_+ \\ &= \prod_{j=2}^k \lambda_j \theta_2 + 2 \sum_{s=2}^k \prod_{j=s}^k \lambda_j \delta_s + \sum_{s=3}^{k+1} \prod_{j=s}^k \lambda_j [\xi^s]_+ \\ &\leq 2 \sum_{s=2}^k \prod_{j=s}^k \lambda_j \delta_s + \sum_{s=2}^{k+1} \prod_{j=s}^k \lambda_j [\xi^s]_+ \\ (4.13) \qquad \qquad &= 2 \sum_{s=2}^k \pi_{s,k} \delta_s + \sum_{s=2}^{k+1} \pi_{s,k} [\xi^s]_+, \end{aligned}$$

where the last inequality follows from (4.12) with $k = 1$ which implies that $\theta_2 \leq [\xi^2]_+$ since $\lambda_1 = 0$ (recall $\lambda_1 = t_1^{-1}(t_0 - 1) = 0$) and the last equality follows from the definition of $\pi_{s,k}$ (see (4.2)). Note that the inequality (4.13) also holds true for $k = 1$ under the convention that $\sum_{s=2}^1 \pi_{s,k} \delta_s = 0$ (recall that $\pi_{2,1} = 1$). Therefore, summing (4.13) from $k = 0$ to ∞ , we have

$$\begin{aligned}
(4.14) \quad \sum_{k=0}^{\infty} \theta_{k+1} &\leq 2 \sum_{k=0}^{\infty} \sum_{s=2}^k \pi_{s,k} \delta_s + \sum_{k=0}^{\infty} \sum_{s=2}^{k+1} \pi_{s,k} [\xi^s]_+ \\
&\leq 2 \sum_{k=0}^{\infty} \sum_{s=2}^{\infty} \pi_{s,k} \delta_s + \sum_{k=0}^{\infty} \sum_{s=2}^{\infty} \pi_{s,k} [\xi^s]_+ \\
&= 2 \sum_{s=2}^{\infty} \left(\sum_{k=0}^{\infty} \pi_{s,k} \right) \delta_s + \sum_{s=2}^{\infty} \left(\sum_{k=0}^{\infty} \pi_{s,k} \right) [\xi^s]_+ \\
&\leq 5a \sum_{s=2}^{\infty} t_{s-1} \delta_s + \frac{5a}{2} \sum_{s=2}^{\infty} t_{s-1} [\xi^s]_+,
\end{aligned}$$

where the last inequality follows from Lemma 4.2. Next, we show that $\{\theta_k\}_{k \in \mathbb{N}}$ is summable. This goal will be achieved by showing that each of the two terms on the right-hand side of (4.14) is finite. From Lemma 4.1 there exists $M_1 > 0$ such that

$$\sum_{s=2}^{\infty} t_{s-1} \delta_s \leq \sum_{s=0}^{\infty} t_{s-1} \delta_s \leq M_1,$$

proving the finiteness of the first term. Now, from the definition of ξ^k and since $\nabla \sigma(\mathbf{x}') + \mathbf{v}' \in \partial \omega(\mathbf{x}')$ (recall that $\omega \equiv \sigma + \psi$ and $\mathbf{v}' \in \partial \psi(\mathbf{x}')$), we have that

$$\xi^k = \frac{\alpha_{k-1}}{\beta} \langle \nabla \sigma(\mathbf{x}') + \mathbf{v}', \mathbf{x}' - \mathbf{x}^k \rangle = \frac{\alpha_{k-1}}{\beta} \langle \xi, \mathbf{x}' - \mathbf{x}^k \rangle,$$

with $\xi \in \partial \omega(\mathbf{x}')$, and hence we obtain that

$$\begin{aligned}
(4.15) \quad \xi^k &= \frac{\alpha_{k-1}}{\beta} \langle \xi, \mathbf{x}' - \mathbf{x}^k \rangle \leq \frac{\rho \alpha_{k-1}}{\beta \sqrt{\tau}} \sqrt{\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}')} \leq \frac{\rho \alpha_{k-1} \sqrt{aC}}{\beta \sqrt{\tau} (k+1)} \\
&\leq \frac{\rho \sqrt{aC}}{\beta \sqrt{\tau} (k+1)^{1+\gamma}},
\end{aligned}$$

where the first inequality follows from (3.31) combined with (3.32), the second inequality follows from Theorem 3.14(i), and the last inequality follows from the definition of $\alpha_k := (k+a)^{-\gamma}$ (see (2.3)) and the fact that $(k+a-1)^{-\gamma} < (k+1)^{-\gamma}$ since $a > 2$. Therefore, from (4.15), we deduce that

$$\begin{aligned}
(4.16) \quad \sum_{s=1}^{\infty} t_{s-1} [\xi^s]_+ &\leq \frac{\rho \sqrt{aC}}{\beta \sqrt{\tau}} \sum_{s=1}^{\infty} \frac{t_{s-1}}{(s+1)^{1+\gamma}} \leq \frac{\rho \sqrt{aC}}{2\beta \sqrt{\tau}} \sum_{s=1}^{\infty} \frac{1}{(s+1)^{\gamma}} = \frac{\rho \sqrt{aC}}{2\beta \sqrt{\tau}} \sum_{s=2}^{\infty} \frac{1}{s^{\gamma}} \\
&= \frac{\rho \sqrt{aC}}{2\beta \sqrt{\tau}} \cdot \frac{1}{\gamma-1},
\end{aligned}$$

where the second inequality follows from the fact that $t_{s-1} = (s-1)/a+1 \leq (s-1)/2+1 = (s+1)/2$ (recall that $a > 2$) and the last inequality follows from Lemma 3.5 with $n_1 = 2$, $n_2 = \infty$, and $r = \gamma > 1$. This proves the finiteness of the second term in (4.14) and therefore completes the proof that $\{\theta_k\}_{k \in \mathbb{N}}$ is summable.

Now, we define $a_k = \mu_k - \sum_{s=1}^k \theta_s$ and since $\mu_k \geq 0$ for all $k \in \mathbb{N}$ and $\{\theta_k\}_{k \in \mathbb{N}}$ is summable, we see that $\{a_k\}_{k \in \mathbb{N}}$ is bounded from below. On the other hand, since $\theta_k = [\mu_k - \mu_{k-1}]_+$, we get that

$$a_{k+1} = \mu_{k+1} - \theta_{k+1} - \sum_{s=1}^k \theta_s \leq \mu_{k+1} - \mu_{k+1} + \mu_k - \sum_{s=1}^k \theta_s = a_k,$$

and hence $\{a_k\}_{k \in \mathbb{N}}$ is also a nonincreasing sequence and thus is convergent. Since $\mu_k = a_k + \sum_{s=1}^k \theta_s$ and both sequences converge, it follows that $\{\mu_k\}_{k \in \mathbb{N}}$ converges. To complete the proof, we will prove that the limit is zero. Since $\{\mu_k\}_{k \in \mathbb{N}}$ converges, it is also bounded, and therefore, from the definition of μ_k , $k \in \mathbb{N}$, we also obtain that $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ is bounded. Hence, it has at least one converging subsequence, which we denote by $\{\mathbf{x}^{j_k}\}_{k \in \mathbb{N}}$ and its limit point with $\tilde{\mathbf{x}}$. From the fact that φ is lower semicontinuous (see Assumption 2.1)) and using Theorem 3.14(i), it follows that

$$\varphi(\tilde{\mathbf{x}}) \leq \liminf_{k \rightarrow \infty} \varphi(\mathbf{x}^{j_k}) = \varphi(\mathbf{x}').$$

Similarly, we obtain that $\omega(\tilde{\mathbf{x}}) \leq \omega(\mathbf{x}')$. Since $\mathbf{x}' \in X'$, these two facts yield that $\tilde{\mathbf{x}} \in X'$. On the other hand, as we proved above, the sequence $\{\|\mathbf{x}^k - \mathbf{x}\|\}_{k \in \mathbb{N}}$ converges for all $\mathbf{x} \in X'$ and, in particular, for $\mathbf{x} = \tilde{\mathbf{x}}$. Thus,

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^k - \tilde{\mathbf{x}}\| = \lim_{k \rightarrow \infty} \|\mathbf{x}^{j_k} - \tilde{\mathbf{x}}\| = 0,$$

which concludes the proof. $\square$

5. Numerical results. This section is devoted to evaluating FBi-PG numerically and comparing its performance with existing algorithms that are proposed to solve convex bilevel optimization problems. For this purpose, we will follow the numerical experimental set-up of [17]. Here is the description of the numerical experiments. We conduct two experiments on real-world problems,

- Songs release year prediction using linear regression.
- Fake news classification using logistic regression.

In each experiment, we run FBi-PG with three parameters $\gamma = 1.3$, $\gamma = 1.5$, and $\gamma = 3$ (all with $a = 2$), which corresponds to three possible inner rates of convergence (see Theorem 3.7 and Proposition 3.8), and compare it to the following algorithms:

- (i) Bi-SG of [17] with $\alpha = 0.95$ and $c = 1$.
- (ii) FISTA [4] applied on the regularized problem (2.1) with a fixed regularization parameter $\alpha_K = 1/K$ when the number of iterations is $K = 10^5$.
- (iii) BIG-SAM [19], where we applied the algorithm on the Moreau envelope of ω (instead of ω itself), since it is a nonsmooth function. Therefore, it is worth noting that BiG-SAM in this case tackles the smoothed bilevel problem rather than the original nonsmooth bilevel problem. We use $\delta = 1$.
- (iv) aPG-APG [10], where we used the same values for the parameters as the ones used in the numerical experiments in [10]. These are $\gamma = 2^{-5}$, $v = 20$, $\eta = 10$, and $\epsilon = 10^{-6}$.

All experiments were run on 11th Gen Intel Core i7-1195G7 @ 2.90 GHz–2.92 GHz with a total RAM memory of 16 GB and four physical cores. We run each algorithm from the same starting point $\mathbf{x}^0 = \mathbf{0}$ and provide details about both the inner and outer optimization problems. In terms of the inner optimization problem, for each method $1 \leq l \leq 5$, we record the following measure over iterations:

$$\Delta^l \varphi(k) := \varphi(\mathbf{x}_{(l)}^k) - \varphi^*,$$

where $\mathbf{x}_{(l)}^k$ is the point computed by the method l after k iterations, and φ^* is the optimal value of φ , which we compute using an off-the-shelf solver.⁴

As for the outer optimization problem, we present a plot that records the progress of the values of ω versus $-\log(\Delta^l \varphi(k))$, namely the number of accuracy digits, for

⁴https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression

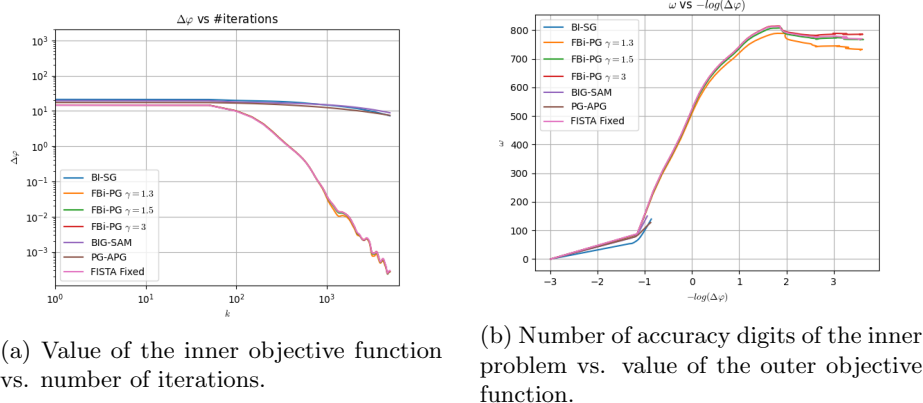


FIG. 1. The progress of the algorithms. In the right-hand side plot, the blue line is “short” due to the low level of accuracy achieved in solving the inner problem, as can be seen from the left-hand side plot.

an algorithm l at iteration k . This plot aims at understanding the behavior of the algorithm in solving both the inner and outer optimization problems. On the x -axis, we plot the values of $-\log(\Delta^l\varphi(k))$, which is an indicator of how accurately the inner problem has been solved. On the y -axis, we plot the values of the outer objective function ω as we get more accurate in solving the inner problem. Since the x -axis represents the negative of (the log of) the optimality gap of the inner-level problem, algorithms that were not able to achieve a good accuracy in solving the inner problem appear as a “short” line in this plot. Among the algorithms that achieve a good accuracy in the inner optimality gap, we seek the algorithm that obtains the minimal value of ω .

5.1. Songs release year prediction using linear regression. In this experiment, we used the YearPredictionMSD dataset⁵ to train a linear regression model to predict song release years. We considered a sample of the dataset consisting of 2,000 uniformly i.i.d. randomly selected songs (without replacement). For each sample, we use A to denote the feature matrix induced from it. Therefore, the (i, j) entry of A is the value of the j th attribute of the i th song in the sample. In addition, we use $\mathbf{b}$ to denote the release year of the songs in the sample. That is, the i th entry in $\mathbf{b}$ is the release year of the i th song in the sample. Our goal is to train a predictor by finding a weights vector $\mathbf{x}$ that minimizes the following objective function:

$$\varphi(\mathbf{x}) = \frac{1}{2N} \|A\mathbf{x} - \mathbf{b}\|^2,$$

where N is the number of songs in the sample. In this case, we consider the outer objective function $\omega \equiv \|\cdot\|_1$. It should be noted that this inner objective function indeed satisfies the Hölderian error bound property [6].

In this experiment, as can be seen on Figure 1, we observe that the fixed regularization parameter we utilized performs comparably to the FBi-PG algorithm with all values of γ . However, as discussed in this paper, the major drawback of this approach is the inability to determine the “right” regularization parameter in advance when solving a problem. This will be illustrated numerically in the forthcoming additional example.

⁵<https://archive.ics.uci.edu/ml/datasets/yearpredictionmsd>

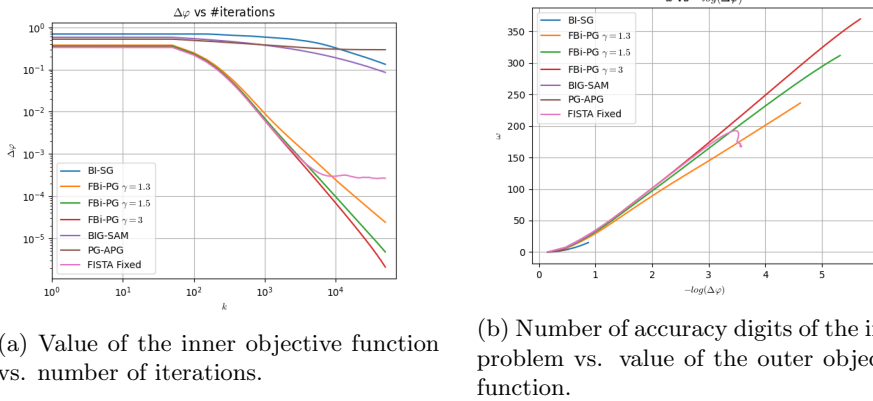


FIG. 2. The progress of the algorithms. In the left-hand side plot, the purple line, which is FISTA Fixed, “gets stuck” after 10^4 iterations due to the fixed regularization parameter.

5.2. Fake news classification using logistic regression. Following the problem description in [17], we present the optimization model immediately. We are given a matrix $A \in \mathbb{R}^{N \times m}$ and a vector $\mathbf{z} \in \mathbb{R}^N$, which contains the labels of the articles. The corresponding logistic loss function is defined by

$$\varphi(\mathbf{x}) \equiv \frac{1}{N} \sum_{i=1}^N (\log(w(\mathbf{x}^T \mathbf{a}_i)) \mathbf{z}_i + \log(1 - w(\mathbf{x}^T \mathbf{a}_i)) (1 - \mathbf{z}_i)),$$

where $w: \mathbb{R} \rightarrow \mathbb{R}_{++}$ is the sigmoid function defined by $w(t) = \exp^t / (1 + \exp^t)$ and $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N)^T$ for which $\mathbf{z}_i \in \{0, 1\}$. We note that the logistic loss function satisfies the Hölderian error bound property [12] and therefore our improved simultaneous rate and convergence guarantees proven in Theorems 3.14 and 4.3 are valid, respectively.

In this example, the bilevel approach can be used here to reduce the number of features in the vector $\mathbf{x}$ obtained from minimizing the logistic loss $\varphi(\cdot)$. In other words, we wish to find the sparsest vector in between all the minimizers of the logistic regression function $\varphi(\cdot)$. Therefore, we take the following classical outer objective function:

$$\omega(\mathbf{x}) = \|\mathbf{x}\|_1.$$

In our experiment, as can be seen on Figure 2, we seek to differentiate between fake and genuine newspaper articles using a dataset from a Kaggle competition⁶ comprising 10,413 real and 10,387 fake news articles. Each article is accompanied by textual content and its authenticity status. We randomly uniformly select 500 fake and real newspapers, employing the preprocessing steps from [17], and represent each article with 250 features as detailed in [17].

In this case, we observe that the fixed regularization parameter we choose doesn’t solve the bilevel optimization problem; it “gets stuck” when solving the inner problem. This means that for our choice of regularizing parameter, optimal solutions of the regularized problem don’t coincide with optimal solutions of the bilevel optimization problem. However, we see in the left-hand side plot that all versions of FBi-PG solve the inner problem. Moreover, as our theory shows, the inner rate of FBi-PG is getting better as γ increases.

⁶Refer to <https://link.springer.com/article/10.1007/s10107-016-1091-6>.

Appendix A. A generic rate for a fixed regularized problem. We show a generic rate result which is independent of the structure of the bilevel problem and of any single-level optimization algorithm when using a fixed regularizing parameter. We describe this formally as follows.

- Let $F : \mathbb{R}^n \rightarrow (-\infty, +\infty]$ be a lower semicontinuous, proper, and convex function that is bounded from below and satisfies some other hypothesis, which we quantify by writing $F \in \mathcal{H}$.
- Assume that a *fixed* regularizing parameter is known in advance for the regularized bilevel problem, namely, the solution of the regularized problem coincides with the solution of the bilevel problem.
- Any optimization algorithm $\mathcal{A}$ which generates iteratively a sequence $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ for minimizing $F \in \mathcal{H}$ enjoys a proven convergence rate, in the sense that, for any $k \in \mathbb{N}$, we have

$$F(\mathbf{x}^k) - F(\mathbf{x}^*) \leq \mathcal{R}_k,$$

where $\mathcal{R}_k$ is the rate of convergence and $\mathbf{x}^*$ is a minimizer of F .

We will use this single-level algorithm to solve bilevel optimization problems. Let φ and ω be the inner and outer objective functions of the bilevel problem, and for a predefined iteration index $K \in \mathbb{N}$, we consider the regularized objective $F_K := \varphi + \alpha_K \omega$ where $\alpha_K > 0$ is the fixed regularizing parameter known in advance, and such that $F_K \in \mathcal{H}$. Under this scenario, we have the following generic rate result.

THEOREM A.1. *Let $\mathbf{x}^K$ be the output after running K iterations of algorithm $\mathcal{A}$ for minimizing the function F_K . Then,*

$$\varphi(\mathbf{x}^K) - \varphi(\mathbf{x}') \leq \mathcal{R}_K + \alpha_K (\omega(\mathbf{x}') - \omega^*) \quad \text{and} \quad \omega(\mathbf{x}^K) - \omega(\mathbf{x}') \leq \alpha_K^{-1} \mathcal{R}_K,$$

where $\mathbf{x}'$ is an optimal solution of the bilevel optimization problem and $\omega^* = \inf_{\mathbf{x} \in \mathbb{R}^n} \omega(\mathbf{x})$.

Proof. Applying K iterations of the algorithm $\mathcal{A}$ to minimize F_K we obviously get that

$$(A.1) \quad F_K(\mathbf{x}^K) - F_K(\mathbf{x}_K^*) \leq \mathcal{R}_K,$$

where $\mathbf{x}_K^*$ is a minimizer of F_K which is also a minimizer of the bilevel problem since α_K is a fixed regularization parameter that was chosen that way. Using the definition of F_K , and since $\varphi(\mathbf{x}^K) - \varphi(\mathbf{x}') \geq 0$, we immediately deduce from (A.1) the following (recall that $\mathbf{x}_K^* = \mathbf{x}'$):

$$\begin{aligned} \alpha_K (\omega(\mathbf{x}^K) - \omega(\mathbf{x}')) &\leq \varphi(\mathbf{x}^K) - \varphi(\mathbf{x}') + \alpha_K (\omega(\mathbf{x}^K) - \omega(\mathbf{x}')) \\ &= F_K(\mathbf{x}^K) - F_K(\mathbf{x}_K^*) \leq \mathcal{R}_K, \end{aligned}$$

which proves the rate of convergence for the outer problem. For the inner problem, using again (A.1), we get

$$\varphi(\mathbf{x}^K) - \varphi(\mathbf{x}') = F_K(\mathbf{x}^K) - F_K(\mathbf{x}') + \alpha_K (\omega(\mathbf{x}') - \omega(\mathbf{x}^K)) \leq \mathcal{R}_K + \alpha_K (\omega(\mathbf{x}') - \omega^*),$$

which completes the proof. $\square$

Appendix B. Proof of Lemma 4.1. We begin with two technical lemmas that will be useful in proving the required result. Before doing so, we recall the definitions of the sequences $\{d_k\}_{k \in \mathbb{N}}$ and $\{\eta_k\}_{k \in \mathbb{N}}$, which will be used below, and are, respectively, defined by

$$d_k := t_{k-1}^2 - (t_k^2 - t_k) \quad \text{and} \quad \eta_k := t_{k-1}^2 \alpha_{k-1} - (t_k^2 - t_k) \alpha_k.$$

We begin with a result that will be used in what follows.

LEMMA B.1. *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG with $1 < \gamma < 2$ and $a > 2$. Suppose that $\{b_k\}_{k \in \mathbb{N}} \subseteq \mathbb{R}_+$ is a sequence satisfying $b_k \leq b\eta_k$ for all $k \in \mathbb{N}$ and some $b > 0$. Then, there exists $M > 0$ such that $\sum_{s=0}^{k-1} b_s(\omega(\mathbf{x}') - \omega(\mathbf{x}^s)) \leq M$ for all $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$.*

Proof. Let $k \in \mathbb{N}$. From Theorem 3.14(ii), we have for all $k \in \mathbb{N}$ that

$$(B.1) \quad \sum_{s=0}^{k-1} b_s(\omega(\mathbf{x}') - \omega(\mathbf{x}^s)) \leq \sum_{s=0}^{k-1} b_s[\omega(\mathbf{x}') - \omega(\mathbf{x}^s)]_+ \leq \frac{C}{a^2} \sum_{s=0}^{k-1} b_s(s+1)^{-1},$$

where $[x]_+ \equiv \max\{0, x\}$. Using (3.30), we have that $b_s \leq b\eta_s < b(s+1)^{1-\gamma}/2$ for all $s \in \mathbb{N}$. Thus, we obtain that $b_s(s+1)^{-1} \leq b(s+1)^{-\gamma}/2$. Now, from Lemma 3.5 with $n_1 = 2$, $n_2 = k$, and $r = \gamma > 1$, we get that

$$(B.2) \quad \sum_{s=0}^{k-1} (s+1)^{-\gamma} = \sum_{s=1}^k s^{-\gamma} = 1 + \sum_{s=2}^k s^{-\gamma} \leq 1 + \frac{1^{1-\gamma} - k^{1-\gamma}}{\gamma - 1} \leq 1 + \frac{1}{\gamma - 1} = \frac{\gamma}{\gamma - 1}.$$

Combining (B.1) with (B.2) yields the desired result. $\square$

Before proving Lemma 4.1 about the summability of the sequence $\{t_{k-1}\delta_{k-1}\}_{k \in \mathbb{N}}$, we prove the summability of the sequence $\{d_k(\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}'))\}_{k \in \mathbb{N}}$.

LEMMA B.2. *Let $\{\mathbf{x}^k\}_{k \in \mathbb{N}}$ be a sequence generated by FBi-PG with $1 < \gamma < 2$ and $a > 2$. Then, there exists $M > 0$ such that $\sum_{s=0}^{k-1} d_s(\varphi(\mathbf{x}^s) - \varphi(\mathbf{x}')) < M$ for all $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$.*

Proof. Let $k \in \mathbb{N}$ and $\mathbf{x}' \in X'$. From the proof of Proposition 3.3 (cf. (3.10)), we have that

$$\begin{aligned} t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}')) &\leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 - \sum_{s=0}^{k-1} d_s(\varphi(\mathbf{x}^s) - \varphi(\mathbf{x}')) \\ &\quad - \sum_{s=0}^{k-1} \eta_s(\omega(\mathbf{x}^s) - \omega(\mathbf{x}')). \end{aligned}$$

Rearranging this inequality yields that

$$(B.3) \quad \begin{aligned} t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}')) + \sum_{s=0}^{k-1} d_s(\varphi(\mathbf{x}^s) - \varphi(\mathbf{x}')) &\leq \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 \\ &\quad + \sum_{s=0}^{k-1} \eta_s(\omega(\mathbf{x}') - \omega(\mathbf{x}^s)). \end{aligned}$$

Since $\mathbf{x}' \in X'$, it follows that $\varphi(\mathbf{x}^k) - \varphi(\mathbf{x}') \geq 0$. Thus, combining with the fact that $F_{k-1} \equiv \varphi + \alpha_{k-1}\omega$, it follows from Theorem 3.14(ii) that

$$t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}')) \geq \alpha_{k-1} t_{k-1}^2 (\omega(\mathbf{x}^k) - \omega(\mathbf{x}')) \geq -\alpha_{k-1} t_{k-1}^2 \frac{C}{a^2(k+1)}.$$

Using now the definitions α_k and t_k (see (2.4) and (2.7), respectively) yields that

$$\begin{aligned} t_{k-1}^2 (F_{k-1}(\mathbf{x}^k) - F_{k-1}(\mathbf{x}')) &\geq -\frac{C(k+a-1)^{2-\gamma}}{a^4(k+1)} \geq -\frac{C(k+a-1)}{a^4(k+1)} \geq -\frac{C(ak+a)}{a^4(k+1)} \\ &= -\frac{C}{a^3}, \end{aligned}$$

where the second inequality from the fact that $(k+a-1)^{2-\gamma} \leq k+a-1$ since $2-\gamma < 1$ (recall that $\gamma > 1$) and $a-1 > 0$ (recall that $a > 2$), and the last inequality follows from the fact that $k+a-1 < k+a < ak+a$ since $a > 1$. By plugging it into (B.3) and adding C/a^3 to both sides, we get that

$$(B.4) \quad \sum_{s=0}^{k-1} d_s (\varphi(\mathbf{x}^s) - \varphi(\mathbf{x}')) \leq \frac{C}{a^3} + \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 + \sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}') - \omega(\mathbf{x}^s)).$$

Applying Lemma B.1 with $b_k \equiv \eta_k$ and $b = 1$ it follows that there exists $\tilde{M} > 0$ such that $\sum_{s=0}^{k-1} \eta_s (\omega(\mathbf{x}') - \omega(\mathbf{x}^s)) \leq \tilde{M}$. Hence, using this bound in (B.4) yields the desired result. $\square$

Now, we are ready to prove Lemma 4.1.

Proof of Lemma 4.1. Let $k \in \mathbb{N}$. From Proposition 3.2 with $\mathbf{x} = \mathbf{x}^k$, we obtain

$$(B.5) \quad \frac{\beta}{2} \|\mathbf{z}^{k+1} - \mathbf{x}^k\|^2 - \frac{\beta}{2} \|\mathbf{z}^k - \mathbf{x}^k\|^2 \leq t_k^2 (F_k(\mathbf{x}^k) - F_k(\mathbf{x}^{k+1})).$$

From (2.5) and the definition of the sequence $\mathbf{z}^k$ we have that $\mathbf{z}^{k+1} - \mathbf{x}^k = t_k (\mathbf{x}^{k+1} - \mathbf{x}^k)$ and $\mathbf{z}^k - \mathbf{x}^k = (t_{k-1} - 1) (\mathbf{x}^k - \mathbf{x}^{k-1})$. Therefore, using the definition of δ_k (see (4.1)), we obtain that

$$\begin{aligned} \frac{\beta}{2} \|\mathbf{z}^{k+1} - \mathbf{x}^k\|^2 - \frac{\beta}{2} \|\mathbf{z}^k - \mathbf{x}^k\|^2 &= t_k^2 \frac{\beta}{2} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2 - (t_{k-1} - 1)^2 \frac{\beta}{2} \|\mathbf{x}^k - \mathbf{x}^{k-1}\|^2 \\ &= \beta \left(t_k^2 \delta_k - (t_{k-1} - 1)^2 \delta_{k-1} \right) \\ &= \beta (t_k^2 \delta_k - t_{k-1}^2 \delta_{k-1}) + \beta (2t_{k-1} - 1) \delta_{k-1} \\ &\geq \beta (t_k^2 \delta_k - t_{k-1}^2 \delta_{k-1}) + \beta t_{k-1} \delta_{k-1}, \end{aligned}$$

where the last inequality follows from the facts that $t_k \geq 1$ and $\delta_k \geq 0$ for all $k \in \mathbb{N}$. Plugging this into (B.5) with $k = s$ yields that

$$(B.6) \quad \beta (t_s^2 \delta_s - t_{s-1}^2 \delta_{s-1}) + \beta t_{s-1} \delta_{s-1} \leq t_s^2 (F_s(\mathbf{x}^s) - F_s(\mathbf{x}^{s+1})).$$

Summing (B.6) for $s = 0, 1, \dots, k-1$, we get that

$$\begin{aligned} \beta \sum_{s=0}^{k-1} t_{s-1} \delta_{s-1} + \beta (t_{k-1}^2 \delta_{k-1} - t_{-1}^2 \delta_{-1}) &= \beta \sum_{s=0}^{k-1} (t_s^2 \delta_s - t_{s-1}^2 \delta_{s-1} + t_{s-1} \delta_{s-1}) \\ &\leq \sum_{s=0}^{k-1} t_s^2 (F_s(\mathbf{x}^s) - F_s(\mathbf{x}^{s+1})). \end{aligned}$$

Since $t_{-1} = 0$ and $t_{k-1}^2 \delta_{k-1} \geq 0$ we obtain that

$$\beta \sum_{s=0}^{k-1} t_{s-1} \delta_{s-1} \leq \sum_{s=0}^{k-1} t_s^2 (F_s(\mathbf{x}^s) - F_s(\mathbf{x}^{s+1})).$$

To complete the proof, we will prove that the sum on the right-hand side is bounded by some $M > 0$ for any $k \in \mathbb{N}$. By defining $\Phi_k = \varphi(\mathbf{x}^k) - \varphi(\mathbf{x}')$ and $\Omega_k = \omega(\mathbf{x}') - \omega(\mathbf{x}^k)$ for all $k \in \mathbb{N}$, we obtain from the definition of $F_k \equiv \varphi + \alpha_k \omega$ that

$$\begin{aligned} F_k(\mathbf{x}^k) - F_k(\mathbf{x}^{k+1}) &= F_k(\mathbf{x}^k) - F_k(\mathbf{x}') + F_k(\mathbf{x}') - F_k(\mathbf{x}^{k+1}) \\ &= \Phi_k - \Phi_{k+1} - \alpha_k(\Omega_k - \Omega_{k+1}), \end{aligned}$$

and hence

$$\begin{aligned} \sum_{s=0}^{k-1} t_s^2 (F_s(\mathbf{x}^s) - F_s(\mathbf{x}^{s+1})) &= \sum_{s=0}^{k-1} t_s^2 (\Phi_s - \Phi_{s+1} - \alpha_s(\Omega_s - \Omega_{s+1})) \\ &= t_0^2 \Phi_0 - t_{k-1}^2 \Phi_k + \sum_{s=1}^{k-1} (t_s^2 - t_{s-1}^2) \Phi_s - \alpha_0 t_0^2 \Omega_0 + \alpha_{k-1} t_{k-1}^2 \Omega_k \\ &\quad - \sum_{s=1}^{k-1} (\alpha_s t_s^2 - \alpha_{s-1} t_{s-1}^2) \Omega_s \\ &\leq t_0^2 \Phi_0 - \alpha_0 t_0^2 \Omega_0 + \frac{a}{a-2} \sum_{s=1}^{k-1} d_s \Phi_s + \alpha_{k-1} t_{k-1}^2 \Omega_k \\ &\quad + \sum_{s=1}^{k-1} (\alpha_{s-1} t_{s-1}^2 - \alpha_s t_s^2) \Omega_s, \end{aligned}$$

where the last inequality follows from the fact that $-t_{k-1}^2 \Phi_k \leq 0$ (since $\Phi_k \geq 0$) and that

$$(B.7) \quad t_s^2 - t_{s-1}^2 \leq t_s = \frac{s+a}{a} = \frac{a^2 d_s - 1}{a(a-2)} \leq \frac{a d_s}{a-2},$$

where the second equality follows from the definition of d_k since

$$d_s = t_{s-1}^2 - (t_s^2 - t_s) = a^{-2} \left((s+a-1)^2 - (s+a)s \right) = a^{-2} ((s+a)(a-2) + 1).$$

Now, we will show that there exists $M > 0$ such that for all $k \in \mathbb{N}$ we have that

$$(B.8) \quad \frac{a}{a-2} \sum_{s=1}^{k-1} d_s \Phi_s + \alpha_{k-1} t_{k-1}^2 \Omega_k + \sum_{s=1}^{k-1} (\alpha_{s-1} t_{s-1}^2 - \alpha_s t_s^2) \Omega_s \leq M.$$

We will prove that each of the three terms above is bounded from above. First, from Lemma B.2 it follows that there exists $M_1 > 0$ such that $\sum_{s=0}^{k-1} d_s \Phi_s \leq M_1$. Second, using Theorem 3.14(ii) and the definition of Ω_k , we obtain that

$$\begin{aligned} \alpha_{k-1} t_{k-1}^2 \Omega_k &= \alpha_{k-1} t_{k-1}^2 (\omega(\mathbf{x}') - \omega(\mathbf{x}^k)) \leq \frac{\alpha_{k-1} t_{k-1}^2 C}{a^2 (k+1)} = \frac{(k+a-1)^{2-\gamma} C}{a^4 (k+1)} \\ (B.9) \quad &\leq \frac{(k+a)C}{a^4 (k+1)} \leq \frac{C}{a^3}, \end{aligned}$$

where the second equality from the definitions of α_k and t_k (see (2.4) and (2.7), respectively), the second inequality follows from the fact that $2-\gamma < 1$ since $\gamma > 1$, and the last inequality from the fact that $k+a < ak+a = a(k+1)$ since $a > 2$. For the third term in (B.8), we first split it as follows:

$$(B.10) \quad \sum_{s=1}^{k-1} \left(\alpha_{s-1} t_{s-1}^2 - \alpha_s t_s^2 + \frac{a}{a-2} \eta_s \right) \Omega_s - \frac{a}{a-2} \sum_{s=1}^{k-1} \eta_s \Omega_s.$$

Moreover, from (B.7) and the definitions of d_s and η_s , we obtain that

$$\begin{aligned}\alpha_s t_s &\leq \frac{a}{a-2} \cdot \alpha_s d_s = \frac{a}{a-2} (\alpha_s t_{s-1}^2 - \alpha_s (t_s^2 - t_s)) \leq \frac{a}{a-2} (\alpha_{s-1} t_{s-1}^2 - \alpha_s (t_s^2 - t_s)) \\ &= \frac{a}{a-2} \eta_s,\end{aligned}$$

where the last inequality follows from the fact that $\alpha_s < \alpha_{s-1}$. Using it and the definition of η_s we obtain that

$$\alpha_{s-1} t_{s-1}^2 - \alpha_s t_s^2 + \frac{a}{a-2} \eta_s \geq \alpha_{s-1} t_{s-1}^2 - \alpha_s t_s^2 + \alpha_s t_s = \eta_s > 0,$$

and

$$\alpha_{s-1} t_{s-1}^2 - \alpha_s t_s^2 + \frac{a}{a-2} \eta_s = \eta_s - \alpha_s t_s + \frac{a}{a-2} \eta_s \leq \eta_s + \frac{a}{a-2} \eta_s = \frac{2a-2}{a-2} \eta_s,$$

where the inequality follows from the fact that $\alpha_s t_s \geq 0$. Therefore, by applying Lemma B.1 with $b_k \equiv \alpha_{k-1} t_{k-1}^2 - \alpha_k t_k^2 + a\eta_k / (a-2) < b\eta_k$ with $b = (2a-2)/(a-2)$ we obtain that there exists $M_3 > 0$ such that (recall that $t_{-1} = \eta_0 = 0$ and $t_0 = 1$)

$$\begin{aligned}\sum_{s=1}^{k-1} \left(\alpha_{s-1} t_{s-1}^2 - \alpha_s t_s^2 + \frac{a}{a-2} \eta_s \right) \Omega_s &= \sum_{s=0}^{k-1} \left(\alpha_{s-1} t_{s-1}^2 - \alpha_s t_s^2 + \frac{a}{a-2} \eta_s \right) \Omega_s + \alpha_0 \omega_{(0)} \\ &\leq M_3 + \alpha_0 \omega_{(0)}.\end{aligned}$$

To complete the proof, we also need to bound the second term in (B.10). Indeed, from Proposition 3.3 after rearranging and using the definitions of Ω_k and F_{k-1} we have (recall that $\eta_0 = 0$)

$$\begin{aligned}-\sum_{s=1}^{k-1} \eta_s \Omega_s &= -\sum_{s=0}^{k-1} \eta_s \Omega_s \leq t_{k-1}^2 (F_{k-1}(\mathbf{x}') - F_{k-1}(\mathbf{x}^k)) + \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 \\ &= t_{k-1}^2 (\varphi(\mathbf{x}') - \varphi(\mathbf{x}^k) + \alpha_{k-1} (\omega(\mathbf{x}') - \omega(\mathbf{x}^k))) + \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 \\ &\leq t_{k-1}^2 \alpha_{k-1} \Omega_k + \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2 \leq \frac{C}{a^3} + \frac{\beta}{2} \|\mathbf{x}^0 - \mathbf{x}'\|^2,\end{aligned}$$

where the second inequality follows from the fact that $\varphi(\mathbf{x}') \leq \varphi(\mathbf{x}^k)$ and last inequality follows from (B.9). This completes the proof of the boundedness of (B.8).

Appendix C. Proofs of Lemmas 3.6, 3.13, and 4.2.

Proof of Lemma 3.6. Let $k \geq 1$. Using the definitions of α_k and t_k (see (2.4) and (2.7), respectively), we obtain that

$$(C.1) \quad \sum_{s=0}^{k-1} \alpha_s t_s = \frac{1}{a} \sum_{s=0}^{k-1} (s+a)^{1-\gamma} = \frac{1}{a} \sum_{s=a}^{k+a-1} s^{1-\gamma} \leq \frac{1}{a} \sum_{s=2}^{k+a-1} s^{1-\gamma},$$

where the last inequality follows from the fact that $a \geq 2$ since we sum (potentially) more elements. Now, we split to three cases. First, if $\gamma = 2$, we get from (C.1) that

$$\sum_{s=0}^{k-1} \alpha_s t_s \leq \frac{1}{a} \sum_{s=2}^{k+a-1} s^{-1} \leq \frac{1}{a} \ln(k+a-1).$$

For all $0 < \gamma \neq 2$, we first obtain from (C.1) and Lemma 3.5 with $n_1 = 2$, $n_2 = k + a - 1$, and $r = \gamma - 1$ that

$$(C.2) \quad \sum_{s=2}^{k+a-1} s^{1-\gamma} \leq \frac{(k+a-1)^{2-\gamma} - 1^{2-\gamma}}{2-\gamma} = \frac{(k+a-1)^{2-\gamma} - 1}{2-\gamma}.$$

If $0 < \gamma < 2$, the desired result follows immediately from (C.2). If $\gamma > 2$, we obtain from (C.2) that

$$\sum_{s=0}^{k-1} \alpha_s t_s \leq \frac{1}{a} \cdot \frac{(k+a-1)^{2-\gamma} - 1}{2-\gamma} = \frac{1}{a} \cdot \frac{1 - (k+a-1)^{2-\gamma}}{\gamma-2}.$$

This concludes the proof of the desired result. $\square$

Proof of Lemma 3.13. Using the definitions of $\alpha_k = (k+a)^{-\gamma}$ and $t_k = (k+a)/a$ for all $k \in \mathbb{N}$, it follows that

$$\begin{aligned} \eta_k &= \alpha_{k-1} t_{k-1}^2 - \alpha_k (t_k^2 - t_k) = \frac{1}{a^2} (k+a-1)^{2-\gamma} - \frac{1}{a^2} (k+a)^{2-\gamma} + \frac{1}{a} (k+a)^{1-\gamma} \\ &< \frac{1}{2} (k+1)^{1-\gamma}, \end{aligned}$$

where the last inequality follows from the facts that $(k+a-1)^{2-\gamma} < (k+a)^{2-\gamma}$ and $(k+a)^{1-\gamma} < (k+1)^{1-\gamma}$ since $1 < \gamma < 2$ and $a \geq 2 > 1$. $\square$

Proof of Lemma 4.2. Let $s \geq 1$. Since $\pi_{s,k} = 1$ for all $k < s$ we obtain that

$$\sum_{k=0}^{\infty} \pi_{s,k} = \sum_{k=0}^{s-1} \pi_{s,k} + \sum_{k=s}^{\infty} \pi_{s,k} = \sum_{k=0}^{s-1} 1 + \sum_{k=s}^{\infty} \pi_{s,k} = s + \sum_{k=s}^{\infty} \pi_{s,k} \leq s + \frac{s+5}{2} = \frac{3s+5}{2},$$

where the inequality follows from [9]. Now, since $a-1 > 1$ we have that $3s+5 < 5(s+a-1) = 5at_{s-1}$ where the last equality follows from the definition of t_s (see (2.7)). This completes the proof. $\square$

REFERENCES

- [1] H. ATTOUCH AND M.-O. CZARNECKI, *Asymptotic behavior of coupled dynamical systems with multiscale aspects*, J. Differential Equations, 248 (2010), pp. 1315–1344, <https://doi.org/10.1016/j.jde.2009.06.014>.
- [2] A. AUSLENDER AND M. TEBoulLE, *Interior gradient and proximal methods for convex and conic optimization*, SIAM J. Optim., 16 (2006), pp. 697–725, <https://doi.org/10.1137/S1052623403427823>.
- [3] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books Math./Ouvrages Math. SMC, Springer, New York, 2011.
- [4] A. BECK AND M. TEBoulLE, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202, <https://doi.org/10.1137/080716542>.
- [5] A. BECK AND M. TEBoulLE, *Gradient-based algorithms with applications to signal-recovery problems*, in Convex Optimization in Signal Processing and Communications, Cambridge University Press, Cambridge, 2010, pp. 42–88.
- [6] J. BOLTE, T. P. NGUYEN, J. PEYPOUQUET, AND B. SUTER, *From error bounds to the complexity of first-order descent methods for convex functions*, Math. Program., 165 (2017), pp. 471–507, <https://doi.org/10.1007/s10107-016-1091-6>.
- [7] R. I. BOT, E. CHENCHENE, E. R. CSETNEK, AND D. A. HULETT, *Accelerating Diagonal Methods for Bilevel Optimization: Unified Convergence via Continuous-time Dynamics*, preprint, <https://arxiv.org/html/2505.14389v1>, 2025.

- [8] J. CAO, R. JIANG, E. H. YAZDANDOOST, AND A. MOKHTARI, *An accelerated gradient method for convex smooth simple bilevel optimization*, in Advances in Neural Information Processing Systems, Vol. 37, 2024, pp. 45126–45154.
- [9] A. CHAMBOLLE AND C. DOSSAL, *On the convergence of the iterates of the “fast iterative shrinkage/thresholding algorithm”*, J. Optim. Theory Appl., 166 (2015), pp. 968–982, <https://doi.org/10.1007/s10957-015-0746-4>.
- [10] P. CHEN, X. SHI, R. JIANG, AND J. WANG, *Penalty-based methods for simple bilevel optimization under h lderian error bounds*, in Advances in Neural Information Processing Systems, Vol. 37, 2024, pp. 140731–140765.
- [11] S. DEMPE AND A. ZEMKHO, *Bilevel optimization: Theory, algorithms, applications and a bibliography*, in Optimization—Advances and Next Challenges, Springer Optim. Appl. 161, Springer, Cham, 2020, pp. 581–672.
- [12] L. DORON AND S. SHTERN, *Methodology and first-order algorithms for solving nonsmooth and non-strongly convex bilevel optimization problems*, Math. Program., 201 (2023), pp. 521–558, <https://doi.org/10.1007/s10107-022-01914-4>.
- [13] M. P. FRIEDLANDER AND P. TSENG, *Exact regularization of convex programs*, SIAM J. Optim., 18 (2008), pp. 1326–1350, <https://doi.org/10.1137/060675320>.
- [14] R. JIANG, N. ABOLFAZLI, A. MOKHTARI, AND E. Y. HAMEDANI, *A conditional gradient-based method for simple bilevel optimization with convex lower-level problem*, in Proceedings of the 26th International Conference on Artificial Intelligence and Statistics, Vol. 206, PMLR, 2023, pp. 10305–10323.
- [15] H. D. KAUSHIK AND F. YOUSEFIAN, *A method with convergence rates for optimization problems with variational inequality constraints*, SIAM J. Optim., 31 (2021), pp. 2171–2198, <https://doi.org/10.1137/20M1357378>.
- [16] L. LAMPARIELLO, S. SAGRATELLA, AND V. G. SASSO, *Addressing hierarchical jointly convex generalized Nash equilibrium problems with nonsmooth payoffs*, SIAM J. Optim., 35 (2025), pp. 445–475, <https://doi.org/10.1137/23M1574026>.
- [17] R. MERCHAV AND S. SABACH, *Convex bi-level optimization problems with nonsmooth outer objective function*, SIAM J. Optim., 33 (2023), pp. 3114–3142, <https://doi.org/10.1137/22M1533608>.
- [18] Y. NESTEROV, *A method for solving the convex programming problem with convergence rate $o(1/k^2)$* , Doklady Akademii Nauk SSSR, 269 (1983), pp. 543–547.
- [19] S. SABACH AND S. SHTERN, *A first order method for solving convex bilevel optimization problems*, SIAM J. Optim., 27 (2017), pp. 640–660, <https://doi.org/10.1137/16M105592X>.
- [20] S. SAMADI, D. BURBANO, AND F. YOUSEFIAN, *Achieving optimal complexity guarantees for a class of bilevel convex optimization problems*, in Proceedings of the 2024 American Control Conference (ACC), 2024, pp. 2206–2211.
- [21] S. SAMADI AND F. YOUSEFIAN, *Improved guarantees for optimal Nash equilibrium seeking and bilevel variational inequalities*, SIAM J. Optim., 35 (2025), pp. 369–399, <https://doi.org/10.1137/23M1589402>.
- [22] M. SOLODOV, *An explicit descent method for bilevel convex optimization*, J. Convex Anal., 14 (2007), pp. 227–237.
- [23] M. TEBoulLE, *A simplified view of first order methods for optimization*, Math. Program., 170 (2018), pp. 67–96, <https://doi.org/10.1007/s10107-018-1284-2>.
- [24] A. TEUSH, *Proximal Gradient Methods for Finding Minimal Norm-like Solutions in Convex Optimization Problems*, M.Sc. Thesis, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel, 2021, <https://tau.primo.exlibrisgroup.com/permalink/972TAU-INST/quev9q/alma9933339452504146>.
- [25] A. N. TIKHONOV AND V. Y. ARSEININ, *Solutions of Ill-posed Problems*, Scripta Series in Mathematics, V. H. Winston & Sons, Washington, DC, John Wiley & Sons, New York, Toronto, London, 1977, translated from the Russian, preface by translation editor Fritz John.
- [26] J. WANG, X. SHI, AND R. JIANG, *Near-optimal convex simple bilevel optimization with a bisection method*, in Proceedings of the 27th International Conference on Artificial Intelligence and Statistics, Vol. 238, Proceedings of Machine Learning Research, PMLR, 2024, pp. 2008–2016.
- [27] H. ZHANG, L. CHEN, J. XU, AND J. ZHANG, *Functionally constrained algorithm solves convex simple bilevel problem*, in Advances in Neural Information Processing Systems, Vol. 37, 2024, pp. 57591–57618.