On the convergence of proximal gradient methods for convex simple bilevel optimization*

Puya Latafat† Andreas Themelis‡ Silvia Villa§ Panagiotis Patrinos†

Abstract

This paper studies proximal gradient iterations for solving simple bilevel optimization problems where both the upper and the lower level cost functions are split as the sum of differentiable and (possibly nonsmooth) proximable functions. We develop a novel convergence recipe for iteration varying stepsizes that relies on Barzilai-Borwein type local estimates for the differentiable terms. Leveraging the convergence recipe, under global Lipschitz gradient continuity, we establish convergence for a nonadaptive stepsize sequence, without requiring any strong convexity or linesearch. In the locally Lipschitz differentiable setting, we develop an adaptive linesearch method that introduces a systematic adaptive scheme enabling large and nonmonotonic stepsize sequences while being insensitive to the choice of hyperparameters and initialization. Numerical simulations are provided showcasing favorable convergence speed of our methods.

Keywords. Convex optimization · bilevel programming · adaptive proximal gradient methods · locally Lipschitz gradient
AMS subject classifications. 65K05 · 90C06 · 90C25 · 90C30

1 Introduction

Bilevel programs consist of optimization problems with a hierarchical structure where the solution of one optimization problem is sought over the set of solutions of another one. Such problems originally emerged in the framework of game theory and have been studied extensively since the 1950s, see [11, 12] for an extensive overview. Recently, they have also found applications in various areas of machine learning such as hyperparameter optimization, meta learning, data poisoning attacks, and reinforcement learning [15, 33, 20, 17, 8]. Variational inequality variants have also been of much interest in recent years [14, 7, 23, 22, 31]. The standard approach for addressing bilevel programs consists of solving a series of approximate problems with better regularity properties; refer to [14, 2, 1] and the references therein. However, it is widely known that these techniques can suffer from many practical issues related to convergence speed and stability.

*This work was supported by: the Research Foundation Flanders (FWO) postdoctoral grant 12Y7622N and research projects G081222N, G033822N, G0A0920N; Research Council KU Leuven C1 project No. C14/18/068; European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 953348; Japan Society for the Promotion of Science (JSPS) KAKENHI grant JP21K17710. S. V. acknowledges the support of the European Commission (grant TraDE-OPT 861137), the US Air Force Office of Scientific Research (FA8655-22-1-7034), the Ministry of Education, University and Research (PRIN 202244A7YL project "Gradient Flows and Non-Smooth Geometric Structures with Applications to Optimization and Machine Learning"). The research by S. V. has been supported by the MIUR Excellence Department Project awarded to Dipartimento di Matematica, Università di Genova, CUP D33C23001110001. S. V. is a member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM). This work represents only the view of the authors. The European Commission and the other organizations are not responsible for any use that may be made of the information it contains.

†Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
‡Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University, 744 Motooka, Nishi-ku 819-0395, Fukuoka, Japan
§Dipartimento di Matematica, Università di Genova, Via Dodecaneso 35, 16146 Genova, Italy.
E-mails: {puya.latafat,panos.patrinos}@kuleuven.be, andreas.themelis@ees.kyushu-u.ac.jp, villa@dima.unige.it

In this work, we study *simple bilevel programs* which refers to problems where the lower level does not have a parametric dependence on the variables of the upper level problem. We split both the upper and the lower cost functions as the sum of differentiable and nonsmooth terms and study two explicit algorithms without the need to solve any inner minimizations. In particular, we consider structured simple bilevel programs of the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; \varphi^{(1)}(x) \coloneqq f^{(1)}(x) + g^{(1)}(x) \tag{1.1a}$$

$$\text{subject to} \;\; x \in \mathcal{X}^{(2)} \coloneqq \underset{w\in\mathbb{R}^n}{\arg\min} \left\{ \varphi^{(2)}(w) \coloneqq f^{(2)}(w) + g^{(2)}(w) \right\}, \tag{1.1b}$$

where functions $f^{(1)}, f^{(2)} : \mathbb{R}^n \to \mathbb{R}$ are convex and have (*locally*) Lipschitz continuous gradients, and $g^{(1)}, g^{(2)} : \mathbb{R}^n \to \overline{\mathbb{R}}$ are proper closed convex (potentially nonsmooth) functions. Some notable example applications include regularized problems in machine learning and signal processing, where the regularization can be captured by the upper level functions, e.g., $g^{(1)} = \|\cdot\|_p$, $p \geq 1$, corresponding to ℓ^p regularization, while the loss function can be captured by the lower level function $f^{(2)}$, and there may be additional constraints such as nonnegativity constraints captured by $g^{(2)}$. The above formulation also encompasses the class of convex nonlinear programs (NLPs) where nonsmooth proximable terms can be incorporated in the cost function, unlike typical NLP formulations (see Remark 2.2).

A fundamental approach for tackling bilevel programs relies on the so-called *diagonal* approach [2, 9, 37], which involves examining the scaled sum of the upper and lower cost functions

$$f_k \coloneqq \sigma_k f^{(1)} + f^{(2)}, \quad g_k \coloneqq \sigma_k g^{(1)} + g^{(2)}, \quad \text{and} \quad \varphi_k \coloneqq f_k + g_k = \sigma_k \varphi^{(1)} + \varphi^{(2)}, \tag{1.2}$$

parametrized with a scalar $\sigma_k > 0$. The method by Cabot [9] for solving simple bilevel programs, here dubbed *Cabot's proximal point algorithm* (CPPA), involves iterative proximal maps (see Section 1.3)

$$x_{k+1} = \operatorname{prox}_{\alpha_{k+1}\varphi_{k+1}}(x_k),$$

where the parameter σ_k is updated after each iteration. Convergence of CPPA was established under the *slow control* condition

$$\sigma_k \searrow 0 \quad \text{and} \quad \sum_{k\in\mathbb{N}} \sigma_k = \infty. \tag{1.3}$$

However, due to its implicit nature, in many applications CPPA leads to inner minimizations or matrix inversions. A notable advancement in this regard was achieved by Solodov in [37], who studies (1.1) when $g^{(1)} \equiv 0$ and $g^{(2)}$ is an indicator of a closed convex set D. The method, here dubbed *Solodov's explicit descent method* (SEDM), uses explicit oracles (gradients for $f^{(1)}, f^{(2)}$ and projections onto D) and updates σ_k after a single step of projected gradient method with Armijo linesearch, without the need to solve any inner minimizations. More specifically, given $\nu, \eta \in (0,1)$ and some $\widehat{\alpha}_0 > 0$, in each iteration an inverse penalty $0 < \sigma_{k+1} \leq \sigma_k$ is chosen and the variable is updated as

$$x_{k+1} = \Pi_D\left(x_k - \alpha_{k+1}\nabla f_{k+1}(x_k)\right), \tag{1.4a}$$

where $\alpha_{k+1} = \widehat{\alpha}_0 \eta^{m_k}$ and $m_k \in \mathbb{N}$ is the smallest such that

$$f_{k+1}(x_{k+1}) \leq f_{k+1}(x_k) + \nu\langle \nabla f_{k+1}(x_k), x_{k+1} - x_k\rangle. \tag{1.4b}$$

The challenge, inherent to the bilevel setting, lies in the fact that $\varphi^{(1)}$ may take values smaller than the optimal solution along the iterates, making the usual telescoping argument invalid. This was overcome in [9, 37] by an intricate analysis that, while not imposing any strong convexity, does require the set of solutions to be bounded. Moreover, the explicit setting of [37] relies crucially on the fact that $g^{(2)}$ is an indicator function, making the extension to the proximal case nontrivial.

In addition to [9, 37], several other methods have been proposed for solving simple bilevel problems. The *minimal norm gradient* method (MNG) [6] studies (1.1) when $g^{(1)} \equiv 0$, $g^{(2)}$ is the indicator of a closed convex set, and the upper level problem is strongly convex. MNG relies on a cutting plane approach which can lead to inner minimizations. The *bilevel gradient sequential averaging method* (BiGSAM) is an explicit method proposed in [35] based on a viscosity approximation approach [39, 28]. It considers problems with Lipschitz differentiable and strongly convex upper level cost

functions, and establishes an $O(1/K)$ worst-case convergence rate in terms of the lower cost function (see Section 4.1.2 for further details). Another related algorithm that we consider in the simulations is the *iterative regularization via dual diagonal descent* (iterative-3D) [16], which is designed for the iterative regularization of linear inverse problems. A remarkable property of iterative-3D is that it does not impose the slow control condition (see [16, Rem. 10]). The *Bi-Sub-Gradient method (version II)* (Bi-SG-II) detailed in Section 4.1.4 was proposed in [27] that allows for nonsmoothness on both levels. It extends BiGSAM by relaxing the strong convexity assumption. Like BiGSAM, it achieves a worst-case convergence rate of $O(1/K)$ in terms of the lower level cost function. Finally, the *diagonal gradient scheme* (DGS) was proposed in [32] for solving smooth simple bilevel programs. It however involves an implicit stepsize rule that is available in closed form only in certain scenarios, such as when a quadratic growth condition holds (see [32, Assumptions H1-H3 and §3.2]).

The above nonexhaustive literature review largely focused on *explicit* proximal gradient–based methods. There have been also many studies that have considered simple bilevel programs with nondifferentiable terms possibly on both levels. Most notably, *Solodov's bundle method* (SBM) proposed in [36] achieves this through explicit subgradient operations combined with minimization subroutines for constructing cutting-planes approximations. Another relevant work is [19] that offers a unified framework for the analysis of (sub)gradient-type iterations. Other recent contributions include [13] that does not require strong convexity or differentiability of the upper level problem, but involves inner subroutines. A cutting-planes strategy that employs conditional gradient–type updates is proposed in [21]. In [18] the authors propose a minimal like–norm gradient method under Hölderian-type assumptions.

1.1 Contributions

We show that proximal gradient iterates involving the family of scaled functions in (1.2) converge under global Lipschitz gradient continuity, without any additional strong convexity assumption. Notably, our proposed scheme, staBiM (Algorithm I), allows for nonsmoothness in the upper level cost, while only requiring proximal operations for individual nonsmooth terms $g^{(1)}, g^{(2)}$ (see Remark 2.1 for evaluating the proximal mapping).

The convergence is actually established for a more general framework that identifies three main properties of the stepsize sequence that, combined with the slow control condition (1.3) are shown to suffice. Our analysis crucially relies on utilization of certain Barzilai-Borwein type estimates for the differentiable terms (see (2.4)). As a result, diverging from [37, 9] that establish a quasi-descent inequality in terms of the distance from the solution, we instead show that such a property holds for a more intricate quantity that depends on a combination of cost function, distance from solution(s) and fixed point residual (see Lemma 3.3).

Thanks to this more general perspective, a new linesearch method, adaBiM (Algorithm II), is proposed, that similar to SEDM [37] can cope with problems involving merely *locally* Lipschitz continuous gradients, while extending the aforementioned work to the proximal setting. AdaBiM provides a dynamic update of the initial stepsize $\widehat{\alpha}_k$, see (2.5), (as opposed to fixing a predefined stepsize initialization hyperparameter) which is refined over the iterations based on the local Barzilai-Borwein–type estimates, yielding larger stepsizes with considerably fewer backtrackings (see the top row of Figure 1).

1.2 Organization

After listing some preliminary material, the next section starts by presenting our assumptions along with commentary on their generality, and concludes with the two proposed algorithms. The proofs of the methods are deferred to Appendix A and rely on a unifying convergence recipe that is presented in detail in Section 3. After some preliminary lemmas related to the adaptive strategy, we derive a quasi-descent inequality in Section 3.1; this is used in Section 3.2 for developing the aforementioned convergence recipe. Numerical simulations are carried out in Section 4, and the paper concludes with some final comments in Section 5.

1.3 Preliminaries

The sets of natural, real, and extended-real numbers are $\mathbb{N}$, $\mathbb{R} := (-\infty, \infty)$ and $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$, respectively, while the positive and strictly positive reals are $\mathbb{R}_+ := [0, \infty)$ and $\mathbb{R}_{++} := (0, \infty)$. We adopt the conventions that $0 \in \mathbb{N}$ and $1/0 = \infty$. Given $x \in \overline{\mathbb{R}}$, its positive part is indicated as $[x]_+ := \max\{0, x\}$. We denote by $\langle \cdot\,, \cdot \rangle$ and $\|\cdot\|$ the standard Euclidean inner product and the induced norm, and with id the identity function defined on a suitable space. The closed Euclidean ball of radius $r > 0$ and centered at $\bar{x} \in \mathbb{R}^n$ is denoted as $\overline{\mathrm{B}}(\bar{x}; r) := \{x \in \mathbb{R}^n \mid \|x - \bar{x}\| \leq r\}$. Given two nonempty sets $U, V \subseteq \mathbb{R}^n$, with $U + V := \{u + v \mid u \in U, v \in V\}$ we indicate their Minkowski sum, while $\operatorname{conv} V$ is used to denote the convex hull of V.

The *domain* and *epigraph* of an extended real–valued function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ are, respectively, the sets $\operatorname{dom} h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$ and $\operatorname{epi} h := \{(x, c) \in \mathbb{R}^n \times \mathbb{R} \mid h(x) \leq c\}$. Function h is said to be *proper* if $\operatorname{dom} h \neq \emptyset$, and *lower semicontinuous (lsc)* if $\operatorname{epi} h$ is a closed subset of $\mathbb{R}^{n+1}$. We say that h is *level bounded* if its c-sublevel set $\operatorname{lev}_{\leq c} h := \{x \in \mathbb{R}^n \mid h(x) \leq c\}$ is bounded for all $c \in \mathbb{R}$.

The *indicator function* of a set $E \subseteq \mathbb{R}^n$ is denoted by δ_E, namely $\delta_E(x) = 0$ if $x \in E$ and ∞ otherwise. The projection onto and the distance from E are respectively denoted by

$$\Pi_E(x) := \operatorname*{arg\,min}_{z \in E} \|z - x\| \quad \text{and} \quad \operatorname{dist}(x, E) := \inf_{z \in E} \|z - x\|.$$

The (convex) *subdifferential* of a proper lsc convex function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ at a point $\bar{x}$ is the set $\partial h(\bar{x}) := \{v \in \mathbb{R}^n \mid h(x) \geq h(\bar{x}) + \langle v, x - \bar{x}\rangle \ \forall x \in \mathbb{R}^n\}$. The *proximal mapping* of h is $\operatorname{prox}_h : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$\operatorname{prox}_h(x) = \operatorname*{arg\,min}_{w \in \mathbb{R}^n} \big\{h(w) + \tfrac{1}{2}\|w - x\|^2\big\},$$

and is characterized by the implicit subdifferential inclusion [4, Eq. (24.2)]

$$x - \operatorname{prox}_h(x) \in \partial h\big(\operatorname{prox}_h(x)\big). \tag{1.5}$$

2 Problem setup and proposed algorithms

We will pattern the frameworks of [37, 9], while considering general convex functions $g^{(1)}$ and $g^{(2)}$ (as opposed to indicator functions), and differentiable functions $f^{(1)}, f^{(2)}$ with locally Lipschitz-continuous gradients. Our main assumptions are as follows.

Assumption I (basic requirements)**.** *The following hold in problem (1.1):*

A1 *$f^{(1)}, f^{(2)} : \mathbb{R}^n \to \mathbb{R}$ are convex and have locally Lipschitz-continuous gradients;*

A2 *$g^{(1)}, g^{(2)} : \mathbb{R}^n \to \overline{\mathbb{R}}$ are proper lsc convex functions with easy to compute proximal mappings;*

A3 *the upper level problem restricted to $\operatorname{dom} \varphi^{(2)}$ is lower bounded:*

$$\bar{\phi}^{(1)} := \inf_{x \in \operatorname{dom} \varphi^{(2)}} \Big\{f^{(1)}(x) + g^{(1)}(x)\Big\} > -\infty; \tag{2.1a}$$

A4 *the set of solutions $\mathcal{X}_\star := \operatorname{arg\,min}\big\{\varphi^{(1)}(x) \mid x \in \operatorname{arg\,min} \varphi^{(2)}\big\}$ is nonempty and bounded; in particular,*

$$\bar{\phi}^{(2)} := \inf_{x \in \mathbb{R}^n} \Big\{f^{(2)}(x) + g^{(2)}(x)\Big\} > -\infty. \tag{2.1b}$$

The boundedness of solution set in Assumption I.A4 is a standard assumption in the generality of our setting, see [37]. It is in particular implied by conditions such as coercivity of the upper cost function $\varphi^{(1)}$.

The iterations of our proposed method amount to selecting a stepsize $\alpha_{k+1} > 0$ together with an (inverse) penalty parameter $\sigma_{k+1} \leq \sigma_k$, followed by one proximal gradient step on the inversely penalized cost function $f_{k+1} + g_{k+1}$:

$$x_{k+1} = \operatorname{prox}_{\alpha_{k+1} g_{k+1}}(x_k - \alpha_{k+1} \nabla f_{k+1}(x_k)). \tag{2.2}$$

Remark 2.1 (prox-friendliness of g_k)**.** For the sake of "explicitness", we assume throughout that the proximal mapping of g_k can be evaluated efficiently. This can be done without losing generality over Assumption I.A2, possibly up to lifting of the problem.

(i) Lifted reformulation. Although in general prox-friendliness is not preserved by the sum, by exploiting the idea presented in [13] the equivalent lifted problem

$$\underset{x=(z,z')\in\mathbb{R}^{2n}}{\text{minimize}} \quad f^{(1)}(z) + g^{(1)}(z) \tag{2.3a}$$

$$\text{subject to } x \in \mathcal{X}^{(2)} \coloneqq \underset{(w,w')\in\mathbb{R}^{2n}}{\arg\min} \left\{ f^{(2)}(w') + \tfrac{1}{2}\|w - w'\|^2 + g^{(2)}(w') \right\} \tag{2.3b}$$

may instead be considered, which by including the quadratic term $\frac{1}{2}\|w - w'\|^2$ into the smooth component of the lower level results in proximal gradient updates being carried out in parallel (cf. step I.2 or step II.2):

$$\begin{aligned} z_{k+1} &= \operatorname{prox}_{\alpha_{k+1}\sigma_{k+1}g^{(1)}}\Big(z_k - \alpha_{k+1}\big(z_k - z'_k + \sigma_{k+1}\nabla f^{(1)}(z_k)\big)\Big), \\ z'_{k+1} &= \operatorname{prox}_{\alpha_{k+1}g^{(2)}}\Big(z'_k - \alpha_{k+1}\big(z'_k - z_k + \nabla f^{(2)}(z'_k)\big)\Big). \end{aligned}$$

(This reformulation still complies with Assumption I provided that $f^{(1)} + g^{(1)}$ is lower bounded on $\mathbb{R}^n$, as opposed to merely on $\operatorname{dom} g^{(2)}$.) This procedure can be generalized to address $g^{(1)}$ of the form $g^{(1)}(x) = g_1(x) + \cdots + g_m(x)$ with each g_i (individually) being proximable, up to suitably adding slack variables and modifying the smooth term $f^{(2)}$ in the lower level.

(ii) Nonlifted option. In many instances of practical interest there is no need to resort to a lifting, as the proximal map of g_k can be evaluated based on that of $g^{(1)}$ and/or $g^{(2)}$. This is trivially the case if either $g^{(1)}$ or $g^{(2)}$ is zero. Other typical instances involve the case in which either one is the ℓ^1 norm, or the indicator of a simple set such as a box or an ℓ^2-ball, for which the proximal mapping of the sum is available in closed form; refer to [4, §24] and [5, §6] for examples and further details of proximable functions. □

Remark 2.2 (Nonlinear programs (NLPs))**.** A notable example is the class of convex NLPs

$$\begin{aligned} &\underset{x\in\mathbb{R}^n}{\text{minimize}} \quad f^{(1)}(x) + g^{(1)}(x) \\ &\text{subject to } x \in D, \quad Ax = b, \quad h(x) \leq 0, \end{aligned}$$

where $h = (h_1, \ldots, h_m)$ and $f^{(1)}, h_i : \mathbb{R}^n \to \mathbb{R}$ are convex and with locally Lipschitz continuous gradients, $i = 1, \ldots, m$, $g^{(1)} : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a proper closed convex (possibly nonsmooth) proximable function, $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$, and D is a nonempty closed convex set easy to project onto. As observed in [37], this problem can be formulated in the form of (1.1) by setting $g^{(2)} = \delta_D$ and $f^{(2)}(x) = \|Ax - b\|^2 + \|\max\{0, h(x)\}\|^2$, where δ_D denotes the indicator function of the set D. (Assumption I holds provided that the set of solutions is bounded and that $\inf_D f^{(1)} + g^{(1)} > -\infty$.) As commented in Remark 2.1*(i)*, as long as $f^{(1)} + g^{(1)}$ is lower bounded the projection onto D and the proximal mapping of $g^{(1)}$ can be decoupled by suitably lifting; further lifting in fact allows for any finite sum structure of proximable terms. □

2.1 The globally Lipschitz case: StaBiM

We begin with a linesearch-free proximal gradient method involving the family of scaled functions in (1.2), under the assumption that $f^{(1)}$ and $f^{(2)}$ have globally Lipschitz continuous gradients. The full generality of Assumption I.A1 will be addressed in the next subsection through introducing a novel linesearch procedure with an adaptive stepsize initialization. The linesearch-free method, StaBiM (Algorithm I), is *static* in that it uses nonadaptive (and nevertheless increasing) stepsizes (the naming convention will become clear in the sequel). It can be viewed as a direct generalization of the iterations considered in [9, 37] to the full splitting setting of (1.1). It extends [9] by allowing explicit gradient oracles, while it extends [37] in that it is not limited to constrained problems.

Algorithm I Static **Bi**level **M**ethod (staBiM) when $f^{(i)}$ are globally $L_{f^{(i)}}$-Lipschitz smooth, $i = 1, 2$

REQUIRE starting point $x_0 \in \mathbb{R}^n$, (inverse) penalty $\sigma_0 > 0$, and $\nu \in (0, 1)$ (e.g., $\nu = 0.99$)

REPEAT FOR $k = 0, 1, \ldots$ until convergence

I.1: Choose $\sigma_{k+1} \in \left[\frac{3}{4}\sigma_k, \sigma_k\right]$

I.2: $\alpha_{k+1} = \frac{\nu}{\sigma_{k+1} L_{f^{(1)}} + L_{f^{(2)}}}$

I.3: $x_{k+1} = \operatorname{prox}_{\alpha_{k+1} g_{k+1}}\left(x_k - \alpha_{k+1} \nabla f_{k+1}(x_k)\right)$ % see Remark 2.1 for the proximal evaluation of g_{k+1} based on that of $g^{(1)}$ and $g^{(2)}$

RETURN x_{k+1}

Although staBiM uses global estimates and nonadaptive stepsizes, its proof is *only* made possible through the study of a more complex *adaptive stepsize* sequence that allows us to establish a *quasi*-descent behavior on a combination of distance to solutions, cost, and fixed-point residual, see Lemma 3.3. The main convergence results for staBiM is presented next. Similarly to [37, 9], it is shown that the distance from the set of solutions converges to zero. Because of nonemptiness and boundedness of the optimal set $\mathcal{X}_\star$ prescribed by Assumption I.A4, this condition is equivalent to existence and optimality of the cluster points.

Theorem 2.3 (convergence of staBiM)**.** *Additionally to Assumption I, suppose that $\nabla f^{(i)}$ are globally $L_{f^{(i)}}$-Lipschitz continuous, $i = 1, 2$, and that $(\sigma_k)_{k\in\mathbb{N}}$ complies with (1.3). Then $((x_k)_{k\in\mathbb{N}}$ is bounded and) $(\operatorname{dist}(x_k, \mathcal{X}_\star))_{k\in\mathbb{N}}$ converges to zero. Moreover, the following sublinear rate holds*

$$\min_{k\le K} \|x_{k+1} - x_k\|^2 \le \tfrac{\nu\bar{\varphi}_0(x_0)}{(1-\nu)L_{f^{(2)}}(1+K)} \quad \text{and} \quad \min_{k\le K} \operatorname{dist}^2(0, \partial\varphi_k(x_{k+1})) \le \left(1 + \sigma_0 \tfrac{L_{f^{(1)}}}{L_{f^{(2)}}}\right) \tfrac{M_{\max}^2 \bar{\varphi}_0(x_0)}{(1-\nu)(1+K)},$$

where $M_{\max} = 1 + \nu + \nu\sigma_0 L_{f^{(1)}}/L_{f^{(2)}}$ and $\bar{\varphi}_0 := \sigma_0(\varphi^{(1)} - \bar{\phi}^{(1)}) + \varphi^{(2)} - \bar{\phi}^{(2)}$ ($\bar{\phi}^{(i)}$ are as in (2.1)).

The above worst-case rate result is in line with existing ones in the bilevel setting in terms of the lower cost function, see e.g., [35]. Whether convergence rates can be obtained in terms of the upper level cost under additional assumptions is an open problem for future work.

2.2 The locally Lipschitz case: AdaBiM

The key idea of our adaptive scheme is based on recent works [26, 25] that study the proximal gradient method, and implicitly enforces a descent inequality bypassing the need for a linesearch. While in the bilevel setting a linesearch is still necessary, this analysis provides a systematic adaptive approach for initializing the linesearch. Specifically, we use local Lipschitz estimates of the differentiable functions $f_k = \sigma_k f^{(1)} + f^{(2)}$ at the previous iterates x_k, x_{k-1} as

$$\ell_k := \frac{\langle \nabla f_k(x_{k-1}) - \nabla f_k(x_k), x_{k-1} - x_k\rangle}{\|x_{k-1} - x_k\|^2}, \quad L_k := \frac{\|\nabla f_k(x_{k-1}) - \nabla f_k(x_k)\|^2}{\|x_{k-1} - x_k\|^2}, \tag{2.4a}$$

and for each $f^{(i)}$ with

$$\ell_k^{(i)} := \frac{\langle \nabla f^{(i)}(x_{k-1}) - \nabla f^{(i)}(x_k), x_{k-1} - x_k\rangle}{\|x_{k-1} - x_k\|^2}, \quad i = 1, 2, \tag{2.4b}$$

so that $\ell_k = \sigma_k \ell_k^{(1)} + \ell_k^{(2)}$. Noting that the denominator of ℓ_k, L_k, or $\ell_k^{(i)}$ is zero iff $\nabla f(x^k) - \nabla f(x^{k-1}) = 0$, we use the convention $0/0 = 0$ so that both ℓ_k and L_k are (well-defined, positive) real numbers. We also remark that these quantities are reminiscent of widely popular Barzilai-Borwein stepsize choices [3] commonly used as heuristics in various minimization settings. The proposed adaBiM uses these quantities in order to establish a quasi-descent inequality (see Lemma 3.3) and can be viewed as an extension of [25, adaPGM (Alg. 1)]; if $\sigma_k = \sigma$ for all $k \ge 0$ and the linesearch is eliminated, then the algorithm reduces to adaPGM applied to the problem of minimizing $f_k + g_k$.

The following theorem establishes the convergence results for adaBiM in the full generality of Assumption I. Its proof relies on the convergence recipe provided in Theorem 3.5 and is provided in

Algorithm II Adaptive **Bi**level **M**ethod (adaBiM) using local Lipschitz estimates (2.4)

REQUIRE starting point $x_{-1} \in \mathbb{R}^n$, stepsize $\alpha_0 > 0$, (inverse) penalty $\sigma_{-1} = \sigma_0 > 0$
backtracking linesearch parameters $\eta, \nu \in (0,1)$ (e.g., $\nu = 0.99$, $\eta = 1/2$)

INITIALIZE $x_0 = \operatorname{prox}_{\alpha_0 g_0}\big(x_{-1} - \alpha_0 \nabla f_0(x_{-1})\big)$, $\alpha_{\max} \gg \frac{1}{\ell_0}$, and $\alpha_{-1} = \alpha_0 \cdot \begin{cases} 1 & \text{if } \alpha_0\ell_0 \geq 1/2 \\ \frac{(\alpha_0\ell_0)^2}{1-(\alpha_0\ell_0)^2} & \text{otherwise} \end{cases}$

REPEAT FOR $k = 0, 1, \ldots$ until convergence

II.1: Choose $\sigma_{k+1} \in \big[\frac{3}{4}\sigma_k, \sigma_k\big]$, and denoting $\rho_k := \frac{\sigma_k \alpha_k}{\sigma_{k-1}\alpha_{k-1}}$ initialize the next stepsize as[1,2]

$$\widehat{\alpha}_{k+1} = \tfrac{\sigma_k}{\sigma_{k+1}}\alpha_k \min\left\{ \sqrt{\tfrac{\sigma_k}{\sigma_{k-1}}\big(1+\rho_k\big)},\ \frac{\sqrt{1-4\Big(1-\frac{\sigma_k}{\sigma_{k-1}}\Big)\alpha_k \ell_k^{(2)}}}{2\sqrt{[\alpha_k^2 L_k^2 - \alpha_k \ell_k]_+}} \right\} \tag{2.5}$$

II.2: $x_{k+1} = \operatorname{prox}_{\alpha_{k+1} g_{k+1}}\big(x_k - \alpha_{k+1}\nabla f_{k+1}(x_k)\big)$, with α_{k+1} the largest in $\big\{\eta^i \min(\alpha_{\max}, \widehat{\alpha}_{k+1}) \mid i \in \mathbb{N}\big\}$ such that

% see Remark 2.1 for the proximal evaluation of g_{k+1} based on that of $g^{(1)}$ and $g^{(2)}$

$$\alpha_{k+1}\ell_{k+1} \leq \nu \tag{2.6}$$

RETURN x_{k+1}

Appendix A. The shown lower bound on the stepsize involves a local Lipschitz modulus for $\sigma_0 \nabla f^{(1)} + \nabla f^{(2)}$ over a compact and convex set $\mathcal{V}$ that, in addition to containing all the iterates x^k, also includes some of those which were discarded during the linesearch (if any).

Theorem 2.4 (convergence of adaBiM)**.** *Suppose that Assumption I holds and that $(\sigma_k)_{k\in\mathbb{N}}$ complies with (1.3). Then, the following holds for the iterates generated by adaBiM (Algorithm II):*

(i) $\alpha_k \geq \alpha_{\min} := \frac{1}{2L_{f_0,\mathcal{V}}} \min\big\{\sqrt{1-\nu}, \sqrt{3}\eta\nu\big\}$ for all $k \geq 1$. Here, $L_{f_0,\mathcal{V}}$ is a Lipschitz modulus for $\nabla f_0 = \sigma_0 \nabla f^{(1)} + \nabla f^{(2)}$ on the bounded set $\mathcal{V} := \operatorname{conv}\big\{x^k \mid k \in \mathbb{N}\big\} + \mathrm{B}(0; \frac{1-\eta}{\eta} r)$, where $r := \max_{k\in\mathbb{N}} \|x^{k+1} - x^k\|$.

(ii) The following worst-case rates hold

$$\min_{k\leq K} \|x_{k+1} - x_k\|^2 \leq \tfrac{\alpha_{\max}\bar{\varphi}_0(x_0)}{(1-\nu)(1+K)} \quad \text{and} \quad \min_{k\leq K} \operatorname{dist}^2(0, \partial\varphi_k(x_{k+1})) \leq \tfrac{\alpha_{\max}}{\alpha_{\min}} \tfrac{M_{\max}^2 \bar{\varphi}_0(x_0)}{(1-\nu)(1+K)},$$

where $M_{\max} = 1 + \alpha_{\max} L_{f_0,\mathcal{V}}$ and $\bar{\varphi}_0$ is as in Theorem 2.3.

(iii) $(x_k)_{k\in\mathbb{N}}$ is bounded and $(\operatorname{dist}(x_k, \mathcal{X}_\star))_{k\in\mathbb{N}}$ converges to zero.

We remark that $\alpha_{\max}$ in step II.2 is required for theoretical reasons only, while in practice it can be set to a large quantity. The only other parameter involved, the initial stepsize α_0, can be set equal to the inverse of a Lipschitz estimate of f_k. In referring to [25, §2] for such practical details, we emphasize that the choice of α_0 plays a marginal role, since the update (2.5) in combination with the suggested expression of α_{-1} ensures that any inappropriate initialization is immediately corrected. We also remark that our convergence results in Theorem 2.4 are in fact independent of the value of α_{-1}, up to replacing $\alpha_{\min} \leftarrow \min\{\alpha_{\min}, \alpha_0\}$ in the statement.

2.2.1 Observations about the stepsizes

We make some observations about the stepsize sequence of adaBiM. In the initialization at step II.1, when $\alpha_k^2 L_k^2 - \alpha_k \ell_k \leq 0$, i.e.,

$$c_k := \frac{L_k^2}{\ell_k} = \frac{\|\nabla f_k(x_{k-1}) - \nabla f_k(x_k)\|^2}{\langle \nabla f_k(x_{k-1}) - \nabla f_k(x_k), x_{k-1} - x_k\rangle} \leq \frac{1}{\alpha_k}, \tag{2.7}$$

[1] The argument of the square root in the numerator of the second term in (2.5) is larger than $1 - \nu > 0$ owing to the conditions $\sigma_{k+1}/\sigma_k \geq 3/4$ and (2.6) enforced during the preceding iteration (see (A.1)).

[2] By the convention $\frac{1}{0} = \infty$, when $\alpha_k^2 L_k^2 - \alpha_k \ell_k \leq 0$ the definition of $\widehat{\alpha}_{k+1}$ reduces to the first term in the minimum.

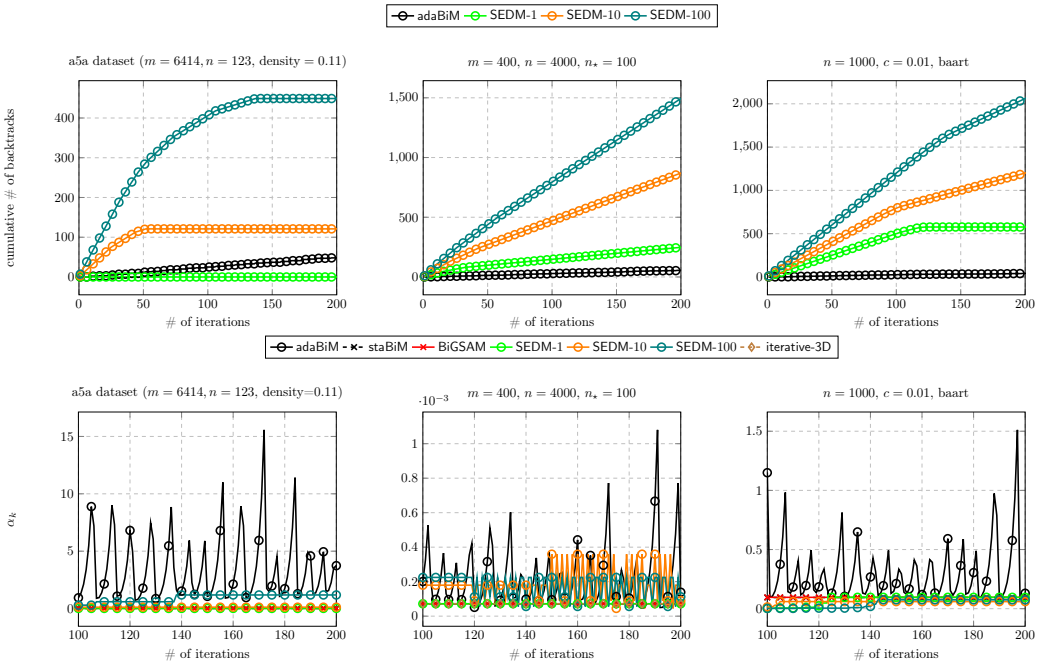


Figure 1: A representative plot showing cumulative number of backtracks needed by adaBiM and SEDM (top row) and stepsize magnitudes in a window of 100 iterations (bottom row) in sample simulations from Section 4. The numerical suffixes -1, -10, and -100 in SEDM indicate different choices for the value of $\widehat{\alpha}_0$ as defined in Section 4.1.1. Left: logistic regression (a5a dataset); center: linear inverse problem; right: solution of integral equations.

the second term in (2.5) reduces to $1/0 = \infty$ and the stepsize initialization simplifies as $\widehat{\alpha}_{k+1} = \sqrt{\frac{\sigma_k}{\sigma_{k-1}}(1+\rho_k)}\frac{\sigma_k}{\sigma_{k+1}}\alpha_k$. This is a critical feature since it allows $\widehat{\alpha}_{k+1}$ to strictly increase compared to the stepsize α_k; for instance, under the standard choice $\sigma_k = 1/k+1$, the first term is always larger than α_k. There is an apparent trade-off between the first and the second term: a large first term allows for faster recovery from small stepsizes at the expense of a smaller second one which affects the global lower bound on α_k. While this interplay can be tweaked by introducing additional algorithmic parameters as done in [24], for clarity of exposition we limit the discussion to this simpler setting.

AdaBiM also retains the linesearch nature of SEDM, but compared to the fixed value $\widehat{\alpha}_0$ of (1.4) it provides a dynamic update of the initial stepsize $\widehat{\alpha}_k$, cf. (2.5), which is refined over the iterations based on local estimates in (2.4), yielding much larger stepsizes with considerably fewer backtrackings. This online self-correcting feature renders the proposed algorithm insensitive to parameters chosen at initalization. As the numerical evidence in Section 4 well highlights, the overhead caused by the backtracking is negligible compared to the speedup that the dynamic update of adaBiM yields, and even under global Lipschitzian assumption this method exhibits superior performance compared to the *static* counterpart staBiM presented in the previous subsection.

The top row of Figure 1 illustrates the cumulative number of backtrackings per iteration required by adaBiM compared to SEDM with different choices of $\widehat{\alpha}_0$; high values of $\widehat{\alpha}_0$ allow for large stepsizes to be tested and potentially accepted, thereby favoring convergence speed in terms of number of iterations, but may lead to more backtrackings and function evaluations in the linesearch (1.4b). Conversely, small values of $\widehat{\alpha}_0$ reduce the complexity of each iteration by reducing the number of backtrackings at the expense of smaller stepsizes and consequently slower convergence. These plots correspond to the respective simulations of Section 4.2 where comparisons in terms of the total number of gradient evaluations are also presented. We also remark that the oscillatory behavior of the stepsizes of adaBiM reported in the bottom plots of Figure 1 is a key element enabling its fast convergence and has been observed also in the setting of minimization problems, cf. [25, §4.3].

3 Convergence analysis

In this section we examine the convergence properties of adaBiM (resp. staBiM) for solving problem (1.1) under local (resp. global) Lipschitz continuity of the gradients of $f^{(1)}$ and $f^{(2)}$. Crucially, regardless of the stepsize selection strategy, our analysis relies on a quasi-descent inequality for the proximal gradient updates (2.2). This allows for a unified analysis provided in Theorem 3.5 that is based on the identification of a set of properties that the stepsizes should satisfy in order to guarantee convergence. We begin by elaborating on some of the notational conventions; a full list is synopsized in Table 1 for convenience.

Remark 3.1 (Bar notation for minima and shifted costs)**.** Let us define

$$\bar{\varphi}^{(i)} \coloneqq \varphi^{(i)} - \bar{\phi}^{(i)},\ i = 1, 2, \quad \text{and} \quad \bar{\varphi}_k \coloneqq \sigma_k \bar{\varphi}^{(1)} + \bar{\varphi}^{(2)},$$

where $\bar{\phi}^{(i)} = \inf_{\operatorname{dom} \varphi^{(2)}} \varphi^{(i)}$ as in Assumption I (restricting to $\operatorname{dom} \varphi^{(2)}$ is superfluous for $i = 2$), and let

$$\phi_\star \coloneqq \inf_{\mathcal{X}^{(2)}} \varphi^{(1)} \tag{3.1}$$

be the optimal cost of problem (1.1). Then,

(i) $\bar{\varphi}^{(1)}(x) \geq 0$ for any $x \in \operatorname{dom} \varphi^{(2)}$ and $\bar{\varphi}^{(2)}(x) \geq 0$ for any $x \in \mathbb{R}^n$;

(ii) $\frac{1}{\sigma_k}\big(\varphi_k(x_\star) - \bar{\phi}^{(2)}\big) = \phi_\star$ for any $x_\star \in \mathcal{X}_\star$ and $k \in \mathbb{N}$. □

A key step in our convergence analysis based on adaptive stepsizes relies on the introduction of the quantities in (2.4). We define the shorhand notation for the forward operator $\mathrm{H}_k = \mathrm{id} - \alpha_k \nabla f_k$, and note that by optimality conditions of the prox-grad update (2.2)

$$\tfrac{1}{\alpha_k}(\mathrm{H}_k(x_{k-1}) - \mathrm{H}_k(x_k)) \in \partial \varphi_k(x_k). \tag{3.2}$$

As we state in the next lemma, not only do the quantities in (2.4) provide an exact description of the local Lipschitz modulus of ∇f_k, but also that of the forward operator H_k.

Fact 3.2 ([25, Lem. 2.1])**.** *Suppose that Assumption I.A1 holds, and for $x_{k-1}, x_k \in \mathbb{R}^n$ and $\alpha_k > 0$ let ℓ_k and L_k be as in (2.4) and $\mathrm{H}_k \coloneqq \mathrm{id} - \alpha_k \nabla f_k$. Then, the following hold:*

(i) $\|\mathrm{H}_k(x_{k-1}) - \mathrm{H}_k(x_k)\| = M_k \|x_{k-1} - x_k\|$, *where*

$$|1 - \alpha_k L_k| \leq M_k \coloneqq \sqrt{1 + \alpha_k^2 L_k^2 - 2\alpha_k \ell_k} \leq 1 + \alpha_k L_k.$$

(ii) $\ell_k \leq L_k \leq \sigma_k L_{f^{(1)},\mathcal{V}} + L_{f^{(2)},\mathcal{V}} \leq \sigma_0 L_{f^{(1)},\mathcal{V}} + L_{f^{(2)},\mathcal{V}}$, *where $L_{f^{(i)},\mathcal{V}}$ is a Lipschitz modulus for $\nabla f^{(i)}$ on a convex set $\mathcal{V}$ containing x_{k-1} and x_k, $i = 1, 2$.*[3]

Notational conventions

As done above and throughout, we use subscripts for iteration counters, typically k, and bracketed superscripts to indicate the level (either 1 or 2). Other symbols that will be introduced for the sake of the convergence proofs adhere to the same conventions and are collected in Table 1, inclusive of references to the respective definitions (those which are local to the scope of individual proofs are omitted from the list). In particular, the uppercase F_k, G_k and Φ_k will be useful for the convergence analysis, and correspond to the respective lowercase symbols scaled by $\frac{1}{\sigma_k}$. Keeping in mind that σ_k is an *inverse* penalty parameter, in the sense that it is driven to 0, we refer to $\Phi_k = \varphi^{(1)} + \frac{1}{\sigma_k}\varphi^{(2)}$ as the penalized cost, and to $\varphi_k = \sigma_k \varphi^{(1)} + \varphi^{(2)}$ as the *inversely* penalized cost of the single-level subproblems.

[3] The first inequality owes to Cauchy-Schwarz; the last one to convexity of $f^{(1)}$ together with the fact that $\sigma_k \leq \sigma_0$.

upper level $\varphi^{(1)} = f^{(1)} + g^{(1)}$

$f^{(1)}$	smooth part	(1.1a)
$g^{(1)}$	proximable part	
$\bar{\phi}^{(1)}$	$\inf_{\operatorname{dom}\varphi^{(2)}} \varphi^{(1)}$	(2.1a)
$\bar{\varphi}^{(1)}$	$\varphi^{(1)} - \bar{\phi}^{(1)}$ (≥ 0 on $\operatorname{dom}\varphi^{(2)}$)	Remark 3.1
$\mathcal{X}_\star$	$\arg\min_{\mathcal{X}^{(2)}} \varphi^{(1)}$ (optimal set)	Assumption I.A4
$\phi_\star$	$\min_{\mathcal{X}^{(2)}} \varphi^{(1)}$ (optimal cost)	(3.1)

lower level $\varphi^{(2)} = f^{(2)} + g^{(2)}$

$f^{(2)}$	smooth part	(1.1b)
$g^{(2)}$	proximable part	
$\bar{\phi}^{(2)}$	$\inf \varphi^{(2)}$	(2.1b)
$\bar{\varphi}^{(2)}$	$\varphi^{(2)} - \bar{\phi}^{(2)}$ (≥ 0)	Remark 3.1
$\mathcal{X}^{(2)}$	$\arg\min \varphi^{(2)}$ (feasible set)	(1.1b)

single-level inverse-penalty reformulation

f_k	$\sigma_k f^{(1)} + f^{(2)}$ smooth part	(1.2)
g_k	$\sigma_k g^{(1)} + g^{(2)}$ proximable part	
φ_k	$f_k + g_k = \sigma_k \varphi^{(1)} + \varphi^{(2)}$	
$\bar{\varphi}_k$	$\sigma_k \bar{\varphi}^{(1)} + \bar{\varphi}^{(2)}$ (≥ 0 on $\operatorname{dom}\varphi^{(2)}$)	Remark 3.1

single-level penalty reformulation

F_k	$f^{(1)} + \frac{1}{\sigma_k} f^{(2)}$ smooth part	(3.3)
G_k	$g^{(1)} + \frac{1}{\sigma_k} g^{(2)}$ proximable part	
Φ_k	$F_k + G_k = \varphi^{(1)} + \frac{1}{\sigma_k}\varphi^{(2)}$	
Δ_{1/σ_k}	$1/\sigma_k - 1/\sigma_{k-1}$	Lemma 3.3

algorithmic parameters

α_{k+1}	stepsize	Algorithm II
σ_{k+1}	(inverse) penalty	
ρ_{k+1}	$\sigma_{k+1}\alpha_{k+1}/\sigma_k\alpha_k$	

adaptive estimates

ℓ_k	Lipschitz estimates of ∇f_k at x_k	(2.4)
L_k		
$\ell_k^{(i)}$	a Lipschitz estimate of $\nabla f^{(i)}$ at x_k	

Table 1: Schematics of the notation adopted in the paper with references to their definitions.

3.1 A quasi-descent inequality

Before delving into the convergence analysis, we will present a series of preliminary results that can be regarded as an extension of the adaptive mechanism of adaPGM proposed in [25, Alg. 1]. This adaptive scheme not only significantly improves the computational efficiency of our approach, but it also allows us to consider nonsmooth terms on both levels.

We proceed to investigate the progress of a single proximal gradient step as described in (2.2) using an arbitrary stepsize $\alpha_{k+1} > 0$. Departing from [37, 9], the key of our convergence analysis, captured in the following lemma, is the adoption of penalized (as opposed to inversely penalized) costs. As already mentioned in the preview of Table 1, we adopt an uppercase notation for the penalized cost

$$\Phi_k = F_k + G_k \quad \text{with} \quad F_k = \tfrac{1}{\sigma_k} f_k = f^{(1)} + \tfrac{1}{\sigma_k} f^{(2)} \quad \text{and} \quad G_k = \tfrac{1}{\sigma_k} g_k = g^{(1)} + \tfrac{1}{\sigma_k} g^{(2)}, \tag{3.3}$$

noticing that $\Phi_k = \varphi^{(1)} + \frac{1}{\sigma_k}\varphi^{(2)} = \frac{1}{\sigma_k}\varphi_k$. Doing so allows us to express the difference

$$\Phi_{k+1} - \Phi_k = \Big(\tfrac{1}{\sigma_{k+1}} - \tfrac{1}{\sigma_k}\Big)\varphi^{(2)} \tag{3.4}$$

as a multiple of the lower-level cost $\varphi^{(2)}$, rather than of the upper-level cost $\varphi^{(1)}$.

Lemma 3.3 (quasi-descent inequality)**.** *Suppose that Assumption I holds and consider iterations (2.2) with $0 < \sigma_{k+1} \leq \sigma_k$. Let $\bar{\varphi}^{(2)} := \varphi^{(2)} - \bar{\phi}^{(2)} \geq 0$ be as in Remark 3.1, and define $\Delta_{1/\sigma_k} := \frac{1}{\sigma_k} - \frac{1}{\sigma_{k-1}}$, $\rho_{k+1} := \frac{\sigma_{k+1}\alpha_{k+1}}{\sigma_k\alpha_k}$, and*

$$\mathcal{L}_k(x_\star) := \tfrac{1}{2}\|x_k - x_\star\|^2 + W_k$$

with

$$\begin{aligned} W_k := {} & \tfrac{1 - 4\alpha_k\sigma_k\ell_k^{(2)}\Delta_{1/\sigma_k}}{4}\|x_k - x_{k-1}\|^2 + \tfrac{\sigma_{k+1}}{\sigma_k}\alpha_{k+1}\rho_{k+1}\bar{\varphi}^{(2)}(x_{k-1}) \\ & + \sigma_k\alpha_k\Delta_{1/\sigma_k}\bar{\varphi}^{(2)}(x_k) + \sigma_k\alpha_k(1+\rho_k)\big(\varphi^{(1)}(x_{k-1}) - \phi_\star\big). \end{aligned}$$

Then, for every $k \in \mathbb{N}$ and $x_\star \in \mathcal{X}_\star$ it holds that

$$\begin{aligned} \mathcal{L}_{k+1}(x_\star) - \mathcal{L}_k(x_\star) \leq {} & -\Big(\tfrac{1}{4} + \rho_{k+1}^2(\alpha_k\ell_k - \alpha_k^2 L_k^2) - \sigma_k\alpha_k\ell_k^{(2)}\Delta_{1/\sigma_k}\Big)\|x_{k-1} - x_k\|^2 \\ & - \sigma_k\alpha_k\big(1 + \rho_k - \rho_{k+1}^2\big)\big(\varphi^{(1)}(x_{k-1}) - \phi_\star\big) \\ & - \sigma_{k+1}\alpha_{k+1}\Big(\tfrac{\Delta_{1/\sigma_k}}{\rho_{k+1}} + \tfrac{1}{\sigma_k}\Big(1 + \rho_{k+1} - \rho_{k+2}^2\tfrac{\sigma_k}{\sigma_{k+1}}\Big)\Big)\bar{\varphi}^{(2)}(x_k). \end{aligned} \tag{3.5}$$

Proof. Let

$$p_k \coloneqq \varphi_k(x_k) - \varphi_k(x_\star) = \sigma_k\big(\varphi^{(1)}(x_k) - \phi_\star\big) + \bar{\varphi}^{(2)}(x_k) \tag{3.6}$$

(which is not necessarily positive). We will prove the claim using the uppercase notation of (3.3) and, consistently, set $P_k \coloneqq \frac{1}{\sigma_k} p_k = \Phi_k(x_k) - \Phi_k(x_\star)$. Being a simple matter of multiplicative constants, note that proximal gradient iterations (2.2) can equivalently be expressed in the penalized cost reformulation $\Phi_{k+1} = F_{k+1} + G_{k+1}$ up to suitable scaling of the stepsize, namely,

$$x_{k+1} = \operatorname{prox}_{\sigma_{k+1}\alpha_{k+1}G_{k+1}}(x_k - \sigma_{k+1}\alpha_{k+1}\nabla F_{k+1}(x_k)).$$

The subgradient characterization of the proximal mapping as in (3.2) yields

$$\tfrac{\mathrm{H}_k(x_{k-1}) - x_k}{\sigma_k\alpha_k} = \tfrac{x_{k-1} - x_k}{\sigma_k\alpha_k} - \nabla F_k(x_{k-1}) \in \partial G_k(x_k), \tag{3.7}$$

where we remind that $\mathrm{H}_k = \mathrm{id} - \alpha_k \nabla f_k$, cf. (3.2). Hence, since $\partial\Phi_k = \nabla F_k + \partial G_k$,

$$\begin{aligned}
0 &\le \Phi_k(x_{k-1}) - \Phi_k(x_k) - \tfrac{1}{\sigma_k\alpha_k}\langle x_{k-1} - x_k, x_{k-1} - x_k\rangle + \langle \nabla F_k(x_{k-1}) - \nabla F_k(x_k), x_{k-1} - x_k\rangle \\
&= \Phi_k(x_{k-1}) - \Phi_k(x_k) - \tfrac{1}{\sigma_k\alpha_k}\|x_{k-1} - x_k\|^2 + \tfrac{1}{\sigma_k}\langle \nabla f_k(x_{k-1}) - \nabla f_k(x_k), x_{k-1} - x_k\rangle \\
&= \Phi_k(x_{k-1}) - \Phi_k(x_k) - \tfrac{1-\alpha_k\ell_k}{\sigma_k\alpha_k}\|x_{k-1} - x_k\|^2 \qquad (3.8a) \\
&= P_{k-1} - P_k + (\tfrac{1}{\sigma_k} - \tfrac{1}{\sigma_{k-1}})\bar{\varphi}^{(2)}(x_{k-1}) - \tfrac{1-\alpha_k\ell_k}{\sigma_k\alpha_k}\|x_{k-1} - x_k\|^2. \qquad (3.8b)
\end{aligned}$$

Again from (3.7), this time with $k \leftarrow k+1$, we have

$$\begin{aligned}
0 &\le G_{k+1}(x_\star) - G_{k+1}(x_{k+1}) + \langle \nabla F_{k+1}(x_k), x_\star - x_{k+1}\rangle - \tfrac{1}{\sigma_{k+1}\alpha_{k+1}}\langle x_k - x_{k+1}, x_\star - x_{k+1}\rangle \\
&= G_{k+1}(x_\star) - G_{k+1}(x_{k+1}) + \langle \nabla F_{k+1}(x_k), x_\star - x_{k+1}\rangle + \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_k - x_\star\|^2 - \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_\star - x_{k+1}\|^2 \\
&\quad - \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_k - x_{k+1}\|^2 \\
&= G_{k+1}(x_\star) - G_{k+1}(x_{k+1}) + \langle \nabla F_{k+1}(x_k), x_\star - x_k\rangle + \langle \nabla F_{k+1}(x_k), x_k - x_{k+1}\rangle + \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_k - x_\star\|^2 \\
&\quad - \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_\star - x_{k+1}\|^2 - \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_k - x_{k+1}\|^2 \\
&\le G_{k+1}(x_\star) - G_{k+1}(x_{k+1}) + F_{k+1}(x_\star) - F_{k+1}(x_k) + \underbrace{\langle \nabla F_{k+1}(x_k), x_k - x_{k+1}\rangle}_{\text{(A)}} + \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_k - x_\star\|^2 \\
&\quad - \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_\star - x_{k+1}\|^2 - \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_k - x_{k+1}\|^2, \qquad (3.9a)
\end{aligned}$$

where the last inequality uses convexity of F_{k+1}. As to term (A), we have

$$\begin{aligned}
\text{(A)} &= \tfrac{1}{\sigma_k\alpha_k}\langle \mathrm{H}_k(x_{k-1}) - x_k, x_{k+1} - x_k\rangle + \tfrac{1}{\sigma_k\alpha_k}\langle \mathrm{H}_k(x_{k-1}) - x_k + \sigma_k\alpha_k\nabla F_{k+1}(x_k), x_k - x_{k+1}\rangle \\
&\overset{(3.7)}{\le} G_k(x_{k+1}) - G_k(x_k) + \langle \nabla F_{k+1}(x_k) - \nabla F_k(x_k), x_k - x_{k+1}\rangle \\
&\quad + \underbrace{\tfrac{1}{\sigma_k\alpha_k}\langle \mathrm{H}_k(x_{k-1}) - \mathrm{H}_k(x_k), x_k - x_{k+1}\rangle}_{\text{(B)}}. \qquad (3.9b)
\end{aligned}$$

Next, we bound the term (B) by ε_{k+1}-Young's inequality as

$$\begin{aligned}
\text{(B)} &\le \tfrac{\varepsilon_{k+1}}{2\sigma_k\alpha_k}\|x_k - x_{k+1}\|^2 + \tfrac{1}{2\varepsilon_{k+1}\sigma_k\alpha_k}\|\mathrm{H}_k(x_{k-1}) - \mathrm{H}_k(x_k)\|^2 \\
&\overset{3.2(i)}{=} \tfrac{\varepsilon_{k+1}}{2\sigma_k\alpha_k}\|x_k - x_{k+1}\|^2 + \tfrac{M_k^2}{2\varepsilon_{k+1}\sigma_k\alpha_k}\|x_{k-1} - x_k\|^2. \qquad (3.9c)
\end{aligned}$$

Combining the three inequalities (3.9) yields

$$\begin{aligned}
0 &\le \Phi_{k+1}(x_\star) - G_{k+1}(x_{k+1}) - F_{k+1}(x_k) + G_k(x_{k+1}) - G_k(x_k) + (\tfrac{1}{\sigma_{k+1}} - \tfrac{1}{\sigma_k})\langle \nabla f^{(2)}(x_k), x_k - x_{k+1}\rangle \\
&\quad + \Bigg\{ \tfrac{M_k^2}{2\varepsilon_{k+1}\sigma_k\alpha_k}\|x_{k-1} - x_k\|^2 - \Big(\tfrac{1}{2\sigma_{k+1}\alpha_{k+1}} - \tfrac{\varepsilon_{k+1}}{2\sigma_k\alpha_k}\Big)\|x_k - x_{k+1}\|^2 - \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_\star - x_{k+1}\|^2 \\
&\qquad + \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_k - x_\star\|^2 \Bigg\}
\end{aligned}$$

$$
\begin{aligned}
&\overset{3.1(ii)}{=} \phi_\star + \tfrac{1}{\sigma_{k+1}}\bar\phi^{(2)} - F_k(x_k) - G_k(x_k) + \big(\tfrac{1}{\sigma_{k+1}} - \tfrac{1}{\sigma_k}\big)\Big(\langle \nabla f^{(2)}(x_k), x_k - x_{k+1}\rangle - f^{(2)}(x_k) - g^{(2)}(x_{k+1})\Big) \\
&\quad + \{\cdots\} \\
&= -P_k + \overbrace{\Big(\tfrac{1}{\sigma_{k+1}} - \tfrac{1}{\sigma_k}\Big)}^{\geq 0}\overbrace{\Big(\langle \nabla f^{(2)}(x_k), x_k - x_{k+1}\rangle - f^{(2)}(x_k) - g^{(2)}(x_{k+1}) + \bar\phi^{(2)}\Big)}^{\text{(D)}} + \Big\{\cdots\Big\}.
\end{aligned}
$$

By using convexity of $f^{(2)}$ we can bound the term (D) as

$$
\begin{aligned}
\text{(D)} &= \langle \nabla f^{(2)}(x_{k+1}), x_k - x_{k+1}\rangle - f^{(2)}(x_k) - g^{(2)}(x_{k+1}) + \langle \nabla f^{(2)}(x_{k+1}) - \nabla f^{(2)}(x_k), x_{k+1} - x_k\rangle + \bar\phi^{(2)} \\
&\leq \qquad\qquad - f^{(2)}(x_{k+1}) \qquad\qquad - g^{(2)}(x_{k+1}) + \qquad\qquad \ell^{(2)}_{k+1}\|x_{k+1} - x_k\|^2 \qquad\qquad + \bar\phi^{(2)},
\end{aligned}
$$

which plugged in the previous inequality results in

$$
\begin{aligned}
0 \leq &- P_k - \big(\tfrac{1}{\sigma_{k+1}} - \tfrac{1}{\sigma_k}\big)\bar\varphi^{(2)}(x_{k+1}) + \big(\tfrac{1}{\sigma_{k+1}} - \tfrac{1}{\sigma_k}\big)\ell^{(2)}_{k+1}\|x_{k+1} - x_k\|^2 + \tfrac{M_k^2}{2\varepsilon_{k+1}\sigma_k\alpha_k}\|x_{k-1} - x_k\|^2 \\
&- \Big(\tfrac{1}{2\sigma_{k+1}\alpha_{k+1}} - \tfrac{\varepsilon_{k+1}}{2\sigma_k\alpha_k}\Big)\|x_k - x_{k+1}\|^2 - \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_\star - x_{k+1}\|^2 + \tfrac{1}{2\sigma_{k+1}\alpha_{k+1}}\|x_k - x_\star\|^2. \qquad (3.10)
\end{aligned}
$$

Summing (3.10)+β_{k+1}(3.8b), multiplying by $\sigma_{k+1}\alpha_{k+1}$, and rearranging yields that for every $\beta_{k+1} \geq 0$ and $\varepsilon_{k+1} > 0$ the following hold.

$$
\begin{aligned}
&\tfrac{1}{2}\|x_{k+1} - x_\star\|^2 + \tfrac{\sigma_{k+1}}{\sigma_k}\alpha_{k+1}(1+\beta_{k+1})p_k + \alpha_{k+1}\sigma_{k+1}\Delta_{1/\sigma_{k+1}}\bar\varphi^{(2)}(x_{k+1}) \\
&+\Big(\tfrac{1-\rho_{k+1}\varepsilon_{k+1}}{2} - \alpha_{k+1}\sigma_{k+1}\Delta_{1/\sigma_{k+1}}\ell^{(2)}_{k+1}\Big)\|x_k - x_{k+1}\|^2 \leq \tfrac{1}{2}\|x_k - x_\star\|^2 + \tfrac{\sigma_{k+1}}{\sigma_{k-1}}\beta_{k+1}\alpha_{k+1}p_{k-1} \\
&-\rho_{k+1}\Big(\beta_{k+1}(1 - \alpha_k\ell_k) - \tfrac{M_k^2}{2\varepsilon_{k+1}}\Big)\|x_{k-1} - x_k\|^2 + \alpha_{k+1}\sigma_{k+1}\beta_{k+1}\Delta_{1/\sigma_k}\bar\varphi^{(2)}(x_{k-1}). \qquad (3.11)
\end{aligned}
$$

By selecting $\beta_k = \rho_k$ and $\varepsilon_k := 1/2\rho_k$ the inequality reduces to the claimed quasi-descent in terms of $\mathcal{L}_k(x^\star)$. □

By looking at the update rule (2.5) for $\widehat{\alpha}_{k+1}$, it is apparent that the choice of stepsizes in Algorithm II is designed so as to ensure that all the multiplying coefficients on the right-hand side of (3.5) are positive. Even so, it should be noted that the inequality does not, in general, imply a monotonic decrease of $\mathcal{L}_k(x_\star)$ along the iterates, the reason being that the term $\varphi^{(1)}(x_{k-1}) - \phi_\star$ therein is not necessarily positive (by the same argument, $\mathcal{L}_k(x_\star)$ is not guaranteed to be positive). For this reason we talk in terms of *quasi*-descent when referring to inequality (3.5), a complication that, similarly to the analysis in [37, 9], is the culprit of a nonstraightforward derivation of convergence results. Nevertheless, regardless of the sign of $\varphi^{(1)}(x_{k-1}) - \phi_\star$, the combination of Assumption I.A3, boundedness of $(\alpha_k)_{k\in\mathbb{N}}$ and $(\rho_k)_{k\in\mathbb{N}}$ (to be established later), the fact that $x_k \in \operatorname{dom}\varphi^{(2)}$ holds for all k, and the slow control condition ensures that $(\mathcal{L}_k(x_\star))_{k\in\mathbb{N}}$ is lower bounded (in fact, $\liminf_{k\to\infty}\mathcal{L}_k(x_\star) \geq 0$) for any $x_\star \in \mathcal{X}_\star$.

3.2 Convergence recipe for proximal gradient iterations

The convergence of the two proposed algorithms hinges on the behavior of proximal gradient iterations when some implicit conditions are met. This is materialized through a convergence recipe relying on the following properties of the generated stepsize sequence.

Properties of stepsizes α_k and inverse penalties σ_k

There exist $\alpha_{\max} \geq \alpha_{\min} > 0$ and $\nu \in (0,1)$ such that, for every $k \in \mathbb{N}$,

$\mathsf{p}_\alpha 1$ $\alpha_{k+1} \leq \min\{\widehat{\alpha}_{k+1}, \alpha_{\max}\}$ with $\widehat{\alpha}_{k+1}$ as in (2.5)	$\mathsf{p}_\sigma 1$ $0 < \sigma_{k+1} \leq \sigma_k$
$\mathsf{p}_\alpha 2$ $\alpha_{k+1}\ell_{k+1} \leq \nu$	$\mathsf{p}_\sigma 2$ $\sigma_k \to 0$ and $\sum_{k\in\mathbb{N}}\sigma_k = \infty$
$\mathsf{p}_\alpha 3$ $\alpha_{k+1} \geq \alpha_{\min}$	

Before presenting the unifying convergence recipe, we establish intermediate but crucial results such as boundedness of the sequence without imposing a uniform lower bound on the stepsize as in Property $\mathsf{p}_\alpha 3$. Optimality of the limit points, however, will ultimately hinge on this final assumption and will be presented in Theorem 3.5.

Lemma 3.4. *Suppose that Assumption I holds, and consider proximal gradient iterations (2.2) with $(\alpha_k)_{k\in\mathbb{N}}$ and $(\sigma_k)_{k\in\mathbb{N}}$ complying with Properties $\mathsf{p}_\alpha 1$ and $\mathsf{p}_\alpha 2$ and Property $\mathsf{p}_\sigma 1$. Then, the following hold:*

(i) $\rho_k := \frac{\sigma_k\alpha_k}{\sigma_{k-1}\alpha_{k-1}} \le \rho_{\max} := \max\left\{\frac{\alpha_0}{\alpha_{-1}}, \frac{1+\sqrt{5}}{2}\right\}$ *for every* $k\in\mathbb{N}$.

(ii) For $\mathcal{L}_k$ *as in Lemma 3.3 it holds that* $\mathcal{L}_{k+1}(x_\star) \le \mathcal{L}_k(x_\star) - \sigma_k\alpha_k(1+\rho_k-\rho_{k+1}^2)\big(\varphi^{(1)}(x_{k-1})-\phi_\star\big)$ *for all* $k\in\mathbb{N}$ *and* $x_\star\in\mathcal{X}_\star$.

(iii) $\bar\varphi_{k+1}(x_{k+1}) \le \bar\varphi_k(x_k) - \frac{1-\nu}{\alpha_{k+1}}\|x_{k+1}-x_k\|^2$ *holds for every* $k\in\mathbb{N}$. *In particular,* $\|x_k-x_{k-1}\|\to 0$ *as* $k\to\infty$, $(\bar\varphi_k(x_k))_{k\in\mathbb{N}}$ *is convergent, the following worst-case rate holds*

$$\min_{k\le K}\|x_{k+1}-x_k\|^2 \le \tfrac{\alpha_{\max}\bar\varphi_0(x_0)}{(1-\nu)(1+K)},$$

and both $(\sigma_k\varphi^{(1)}(x_k))_{k\in\mathbb{N}}$ *and* $(\varphi^{(2)}(x_k))_{k\in\mathbb{N}}$ *are bounded.*

(iv) The sequence $(x^k)_{k\in\mathbb{N}}$ *is bounded.*

Proof.

♠ 3.4*(i)* Observing that $\rho_{k+1} \le \sqrt{\frac{\sigma_k}{\sigma_{k-1}}(1+\rho_k)} \le \sqrt{1+\rho_k}$ by Property $\mathsf{p}_\alpha 1$, the assertion follows from a trivial induction argument.

♠ 3.4*(ii)* Since $\alpha_k \le \widehat{\alpha}_k$, from the definition of $\widehat{\alpha}_k$ in (2.5) it follows that the multiplying coefficients in (3.5) are positive:

$$\begin{aligned}\mathcal{L}_{k+1}(x_\star) - \mathcal{L}_k(x_\star) \le\ & -\sigma_{k+1}\alpha_{k+1}\Big(\overbrace{\tfrac{\Delta_{1/\sigma_k}}{\rho_{k+1}}}^{\ge 0} + \tfrac{1}{\sigma_k}\overbrace{\big(1+\rho_{k+1}-\rho_{k+2}^2\tfrac{\sigma_k}{\sigma_{k+1}}\big)}^{\ge 0}\Big)\bar\varphi^{(2)}(x_k)\\ & -\sigma_k\alpha_k\overbrace{\big(1+\rho_k-\rho_{k+1}^2\big)}^{\ge 0}\big(\varphi^{(1)}(x_{k-1})-\phi_\star\big)\\ & -\Big(\underbrace{\tfrac14+\rho_{k+1}^2(\alpha_k\ell_k-\alpha_k^2L_k^2)-\alpha_k\sigma_k\ell_k^{(2)}\Delta_{1/\sigma_k}}_{\ge 0}\Big)\|x_{k-1}-x_k\|^2.\end{aligned}$$

The claim then readily follows from the fact that $\bar\varphi^{(2)}\ge 0$, cf. Remark 3.1*(i)*.

♠ 3.4*(iii)* We have

$$\begin{aligned}\bar\varphi_{k+1}(x_{k+1}) - \bar\varphi_k(x_k) &\le \bar\varphi_{k+1}(x_{k+1}) - \bar\varphi_{k+1}(x_k)\\ &= \varphi_{k+1}(x_{k+1}) - \varphi_{k+1}(x_k)\\ &= \sigma_{k+1}(\Phi_{k+1}(x_{k+1}) - \Phi_{k+1}(x_k)) \overset{(3.8a)}{\le} -\tfrac{1-\nu}{\alpha_{k+1}}\|x_{k+1}-x_k\|^2.\end{aligned} \tag{3.12}$$

Here, the first inequality uses the fact that $\sigma_k\ge\sigma_{k+1}$, and therefore $\bar\varphi_k = \sigma_k\bar\varphi^{(1)}+\bar\varphi^{(2)}$ is smaller than $\bar\varphi_{k+1} = \sigma_{k+1}\bar\varphi^{(1)}+\bar\varphi^{(2)}$ on $\operatorname{dom}\varphi^{(2)}$, cf. Remark 3.1*(i)*, and $x_k\in\operatorname{dom} g_k = \operatorname{dom}\varphi^{(1)}\cap\operatorname{dom}\varphi^{(2)}\subseteq\operatorname{dom}\varphi^{(2)}$; the second inequality follows from the fact that $\alpha_{k+1}\ell_{k+1}\le\nu$. This shows the sought inequality. In turn, since $\nu<1$, the sequence $(\bar\varphi_k(x_k))_{k\in\mathbb{N}}$ is decreasing; since $x_k\in\operatorname{dom}\varphi^{(2)}$, it follows from Remark 3.1*(i)* that $\bar\varphi_k(x_k) = \sigma_k\bar\varphi^{(1)}(x^k)+\bar\varphi^{(2)}(x^k)\ge 0$, hence that it is convergent and that the positive-valued sequences $(\sigma_k\bar\varphi^{(1)}(x^k))_{k\in\mathbb{N}}$ and $(\bar\varphi^{(2)}(x^k))_{k\in\mathbb{N}}$ are bounded.

The convergence rate result follows immediately by telescoping (3.12)

$$\tfrac{(1-\nu)(K+1)}{\alpha_{\max}}\min_{k\le K}\|x_{k+1}-x_k\|^2 \le \sum_{k=0}^{K}\tfrac{1-\nu}{\alpha_{k+1}}\|x_{k+1}-x_k\|^2 \le \bar\varphi_0(x_0), \tag{3.13}$$

where Property $\mathsf{p}_\alpha 1$ was used to bound $\alpha_{k+1}\le\alpha_{\max}$.

♠ 3.4*(iv)* We pattern the proof structure of [37, Thm. 3.2], thereby considering two mutually exclusive cases.

♢ *Case 1: $\varphi^{(1)}(x_k) \geq \phi_\star$ holds for k large enough.*

Recall the definition of W_k and $\mathcal{L}_k$ in Lemma 3.3. Observing that $W_k \geq 0$ and $\frac{1}{2}\|x_k - x_\star\|^2 \leq \mathcal{L}_k(x_\star)$, Lemma 3.4*(ii)* implies that $(\mathcal{L}_k(x_\star))_{k\in\mathbb{N}}$ converges and that consequently $(x_k)_{k\in\mathbb{N}}$ is bounded.

♢ *Case 2: $\varphi^{(1)}(x_k) < \phi_\star$ holds infinitely often.*

In this case, for every k large enough the index

$$i_k := \max\left\{i \leq k \mid \varphi^{(1)}(x_i) < \phi_\star\right\} \tag{3.14}$$

is well defined. We proceed by intermediate claims.

Claim 3.4.1: the sequences $(x_{i_k})_{k\in\mathbb{N}}$ and $(x_{i_k+1})_{k\in\mathbb{N}}$ are bounded.
The optimal set

$$\mathcal{X}_\star = \left\{x \in \mathcal{X}^{(2)} \mid \varphi^{(1)}(x) \leq \phi_\star\right\} = \left\{x \mid \varphi^{(1)}(x) \leq \phi_\star,\ \bar{\varphi}^{(2)}(x) \leq 0\right\}$$

coincides with a sublevel set of the convex function $h := \max\left\{\varphi^{(1)} - \phi_\star, \bar{\varphi}^{(2)}\right\}$. Since it is nonempty and bounded by Assumption I.A4, h is level bounded; see, e.g., [4, Prop. 11.13] or [38, Lem. 1]. Note that Lemma 3.4*(iii)* implies that $(\bar{\varphi}^{(2)}(x_k))_{k\in\mathbb{N}}$ is (upper) bounded, which combined with the fact that $\varphi^{(1)}(x_{i_k}) < \phi_\star$ implies that $(x_{i_k})_{k\in\mathbb{N}}$ lies in a sublevel set of h, and is therefore bounded. In turn, Lemma 3.4*(iii)* implies that so is $(x_{i_k+1})_{k\in\mathbb{N}}$.

Claim 3.4.2: the whole sequence $(x_k)_{k\in\mathbb{N}}$ is bounded.
To this end, it remains to show that $(x_{k'})_{k'\in K'}$ is bounded, where

$$K' := \{k' \in \mathbb{N} \mid k' \geq i_{k'} + 2\}. \tag{3.15}$$

With $\rho_{\max}$ as in Lemma 3.4*(i)*, for every $k \in \mathbb{N}$

$$\mathcal{L}_{i_k+1}(x_\star) \leq \overbrace{\tfrac{1}{2}\|x_{i_k+1} - x_\star\|^2}^{\text{bounded by Claim 3.4.1}} + \overbrace{\tfrac{1}{4}\|x_{i_k+1} - x_{i_k}\|^2}^{\to 0 \text{ by Lemma 3.4(iii)}} + \overbrace{\sigma_{i_k+1}\alpha_{i_k+1}(1+\rho_{i_k+1})(\varphi^{(1)}(x_{i_k}) - \phi_\star)}^{<0}$$
$$\underbrace{+ \alpha_{\max}\rho_{\max}^2\bar{\varphi}^{(2)}(x_{i_k}) + \alpha_{\max}\left(1 - \tfrac{\sigma_{i_k+1}}{\sigma_{i_k}}\right)\bar{\varphi}^{(2)}(x_{i_k+1})}_{\text{bounded by Lemma 3.4(iii)}}.$$

In particular, we have

$$\sup_{k\in\mathbb{N}} \mathcal{L}_{i_k+1}(x_\star) < \infty. \tag{3.16}$$

Let now $k' \in K'$. Since $\varphi^{(1)}(x_j) - \phi_\star \geq 0$ for $j = i_{k'} + 1, \ldots, k'$, observe that

$$\tfrac{1}{2}\|x_{k'} - x_\star\|^2 \leq \mathcal{L}_{k'}(x_\star) \quad \forall k' \in K', \tag{3.17}$$

and Lemma 3.4*(ii)* yields that

$$\begin{aligned}\mathcal{L}_{k'}(x_\star) &\leq \mathcal{L}_{k'-1}(x_\star) \leq \cdots \leq \mathcal{L}_{i_{k'}+2}(x_\star) \\ &\leq \mathcal{L}_{i_{k'}+1}(x_\star) - \sigma_{i_{k'}+1}\alpha_{i_{k'}+1}(1 + \rho_{i_{k'}+1} - \rho_{i_{k'}+2}^2)\big(\varphi^{(1)}(x_{i_{k'}}) - \phi_\star\big) \\ &\leq \mathcal{L}_{i_{k'}+1}(x_\star) + \underbrace{\tfrac{\sigma_{i_{k'}+1}}{\sigma_{i_{k'}}}}_{\leq 1}\underbrace{\alpha_{\max}(1 + \rho_{i_{k'}+1} - \rho_{i_{k'}+2}^2)}_{\text{bounded by Lemma 3.4(i)}}\underbrace{\sigma_{i_{k'}}\big|\varphi^{(1)}(x_{i_{k'}}) - \phi_\star\big|}_{\text{bounded by Lemma 3.4(iii)}} < \infty \qquad \forall k' \in K',\end{aligned} \tag{3.18}$$

where (3.16) was used in the last inequality. Here, boundedness of the under-bracketed term follows from boundedness of $(x_{i_{k'}})_{k'\in K'}$ and lower semicontinuity of $\varphi^{(1)}$. Then, (3.17) implies that the sequence $(x_{k'})_{k'\in K'}$ is bounded. Combined with Claim 3.4.1 and the fact that the index set K' is the complement of the indices therein, the claim follows. □

Theorem 3.5 (convergence recipe for proximal gradient iterations)**.** *Suppose that Assumption I holds, and consider the proximal gradient iterations (2.2) with $(\sigma_k)_{k\in\mathbb{N}}$ and $(\alpha_k)_{k\in\mathbb{N}}$ complying with all Properties $\mathsf{p}_\alpha 1$ to $\mathsf{p}_\alpha 3$ and Properties $\mathsf{p}_\sigma 1$ and $\mathsf{p}_\sigma 2$. Then,*

(i) $(x_k)_{k\in\mathbb{N}}$ is bounded and) $(\operatorname{dist}(x_k,\mathcal{X}_\star))_{k\in\mathbb{N}}$ converges to zero.

(ii) both $\bar\varphi_k(x_k)$ and W_k as in Lemma 3.3 converge to 0 as $k\to\infty$, and the following worst-case rate holds

$$\min_{k\le K}\operatorname{dist}^2(0,\partial\varphi_{k+1}(x_{k+1}))\le\tfrac{(1+\alpha_{\max}L_{f_0,\mathcal{V}})^2}{\alpha_{\min}(1-\nu)(K+1)}\bar\varphi_0(x_0),$$

where $L_{f_0,\mathcal{V}}$ is a Lipschitz modulus for $\sigma_0\nabla f^{(1)}+\nabla f^{(2)}$ on $\mathcal{V}\coloneqq\operatorname{conv}\{x^k\mid k\in\mathbb{N}\}$.

Proof. We begin by remarking that boundedness of the sequence $(x_k)_{k\in\mathbb{N}}$ is ensured by Properties $\mathsf{p}_\alpha 1$ and $\mathsf{p}_\alpha 2$, as shown in Lemma 3.4*(iv)*. We next prove each claim individually.

♠ 3.5*(ii)* Consider a convergent subsequence $x_{k_j}\to x_\infty$, so that $x_{k_j+1}\to x_\infty$ by Lemma 3.4*(iii)*. Up to further extracting if necessary we have that $\alpha_{k_j+1}\to\alpha_\infty\ge\alpha_{\min}>0$ and $\sigma_{k_j}\to 0$. Observe that

$$x_{k+1}=\operatorname{arg\,min}h(\,\cdot\,;x_k,\alpha_{k+1},\sigma_{k+1}),$$

where

$$h(w;x,\alpha,\sigma)\coloneqq(\sigma g^{(1)}+g^{(2)})(w)+\tfrac{1}{2\alpha}\big\|w-x+\alpha\nabla\big(\sigma f^{(1)}+f^{(2)}\big)(x)\big\|^2$$

is level bounded in w locally uniformly in (x,α,σ), as a function from $\mathbb{R}^n\times\big(\mathbb{R}^n\times[\alpha_{\min},\infty)\times[0,\infty)\big)$ to $\overline{\mathbb{R}}$. Since h is continuous in (x,α,σ), it follows from [34, Thm. 1.17] that

$$x_\infty\in\operatorname{arg\,min}h(\,\cdot\,;x_\infty,\alpha_\infty,0)=\operatorname{prox}_{\alpha_\infty g^{(2)}}\big(x_\infty-\alpha_\infty\nabla f^{(2)}(x_\infty)\big),$$

this condition being equivalent to $x_\infty\in\operatorname{arg\,min}(f^{(2)}+g^{(2)})\overset{(def)}{=}\mathcal{X}^{(2)}$. We next show that $\bar\varphi_k(x_k)\to 0$.

For any $x_\star\in\mathcal{X}_\star$, the subdifferential characterization of $x_{k+1}\in\operatorname{prox}_{\alpha_{k+1}g_{k+1}}(x_k-\alpha_{k+1}\nabla f_{k+1}(x_k))$

$$\tfrac{x_{k-1}-x_k}{\alpha_k}-(\nabla f_k(x_{k-1})-\nabla f_k(x_k))\in\partial\varphi_k(x_k)$$

implies that

$$\varphi_k(x_\star)\ge\varphi_k(x_k)+\langle\overbrace{\tfrac{x_{k-1}-x_k}{\alpha_k}-(\nabla f_k(x_{k-1})-\nabla f_k(x_k))}^{\to 0\text{ by Lemma 3.4(iii)}},\overbrace{x_\star-x_k}^{\text{bounded}}\rangle,$$

hence that

$$\bar\phi^{(2)}=\lim_{k\to\infty}\varphi_k(x_\star)\ge\limsup_{k\to\infty}\varphi_k(x_k)=\limsup_{k\to\infty}\Big(\sigma_k\bar\varphi^{(1)}(x_k)+\overbrace{\varphi^{(2)}(x_k)}^{\ge\bar\phi^{(2)}}\Big)\ge\limsup_{k\to\infty}\sigma_k\bar\varphi^{(1)}(x_k)+\bar\phi^{(2)}.$$

Here, the second equality is obtained by adding and subtracting $\sigma_k\bar\phi^{(1)}$ along with the fact that $\sigma_k\to 0$. Since $\sigma_k\bar\varphi^{(1)}\ge 0$, necessarily $\sigma_k\bar\varphi^{(1)}(x_k)\to 0$ and consequently $\limsup_{k\to\infty}\varphi^{(2)}(x_k)=\bar\phi^{(2)}$. Similarly, since $\varphi^{(2)}\ge\bar\phi^{(2)}$ this necessarily implies that $\varphi^{(2)}(x_k)\to\bar\phi^{(2)}$. This shows that $\bar\varphi^{(2)}(x^k)\to 0$ as claimed.

We next show that $(W_k)_{k\in\mathbb{N}}$ too is vanishing. Observe that

$$\begin{aligned}W_k\overset{(def)}{=}&\tfrac{1-4\alpha_k\sigma_k\ell_k^{(2)}\Delta_{1/\sigma_k}}{4}\overbrace{\|x_k-x_{k-1}\|^2}^{\to 0\text{ by Lemma 3.4(iii)}}+\tfrac{\sigma_{k+1}}{\sigma_k}\alpha_{k+1}\rho_{k+1}\overbrace{\bar\varphi^{(2)}(x_{k-1})}^{\to 0}\\&+\underbrace{\sigma_k\alpha_k\Delta_{1/\sigma_k}\bar\varphi^{(2)}(x_k)}_{\to 0}+\alpha_k(1+\rho_k)\underbrace{\sigma_k\big(\varphi^{(1)}(x_{k-1})-\phi_\star\big)}_{\to 0}\to 0\quad\text{as }k\to\infty,\end{aligned}$$

where we also used the fact that $(\alpha_k)_{k\in\mathbb{N}}\le\alpha_{\max}$ and $(\rho_k)_{k\in\mathbb{N}}\le\rho_{\max}$ as in Lemma 3.4*(i)*.

The asserted sublinear rate is a consequence of (3.13), the characterization of the residual in Fact 3.2*(i)* (see (3.2)), and the upper bounds for ℓ_k, L_k in Fact 3.2*(ii)*:

$$\begin{aligned}\textstyle\sum_{k=0}^K\operatorname{dist}^2(0,\partial\varphi_{k+1}(x_{k+1}))&\le\textstyle\sum_{k=0}^K\frac{1}{\alpha_{k+1}^2}\|\operatorname{H}_{k+1}(x_k)-\operatorname{H}_{k+1}(x_{k+1})\|^2\\&\le\tfrac{M_{\max}^2}{\alpha_{\min}}\textstyle\sum_{k=0}^K\frac{1}{\alpha_{k+1}}\|x_k-x_{k+1}\|^2\le\frac{M_{\max}^2}{\alpha_{\min}(1-\nu)}\bar\varphi_0(x_0),\end{aligned}$$

where $M_{\max}=\sup_{k\in\mathbb{N}}M_k\le 1+\alpha_{\max}L_{f_0,\mathcal{V}}$, see Fact 3.2.

♠ 3.5*(i)* We begin by observing that, since $\bar{\varphi}^{(2)}(x^k) \to 0$, by lower semicontinuity all limit points of $(x^k)_{k\in\mathbb{N}}$ belong to $\mathcal{X}^{(2)}$. As done in the proof of Lemma 3.4*(iv)* we consider two possible cases.

♢ *Case 1: $\varphi^{(1)}(x_k) \geq \phi_\star$ holds for k large enough.*

In this case, we will actually show that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to a solution of (1.1). We first argue that there exists an optimal limit point; to this end, since all limit points are feasible and because of lower semicontinuity, it suffices to show that $\liminf_{k\to\infty} \varphi^{(1)}(x_k) = \phi_\star$. A telescoping argument on Lemma 3.4*(ii)* along with Property $\mathsf{p}_\alpha 3$ yields

$$\alpha_{\min} \sum_{k\in\mathbb{N}} \sigma_k(1+\rho_k-\rho_{k+1}^2)\big(\varphi^{(1)}(x_{k-1})-\phi_\star\big) \leq \sum_{k\in\mathbb{N}} \sigma_k\alpha_k(1+\rho_k-\rho_{k+1}^2)\big(\varphi^{(1)}(x_{k-1})-\phi_\star\big) < \infty. \tag{3.19}$$

Since $\sum_{k\in\mathbb{N}} \sigma_k = \infty$, necessarily

$$\liminf_{k\to\infty}(1+\rho_k-\rho_{k+1}^2)(\varphi^{(1)}(x_{k-1})-\phi_\star) = 0. \tag{3.20}$$

Notice that $\limsup_{k\to\infty}(1+\rho_k-\rho_{k+1}^2) > 0$, for otherwise $1+\rho_k-\rho_{k+1}^2 \to 0$, implying that $\liminf_{k\to\infty}\rho_k > 0$, and consequently that $\alpha_k\sigma_k$ eventually increases exponentially, which contradicts $\sigma_k\alpha_k \leq \sigma_k\alpha_{\max} \to 0$. Hence, $\limsup_{k\to\infty}(1+\rho_k-\rho_{k+1}^2) > 0$, which along with (3.20) implies that $\liminf_{k\to\infty}\varphi^{(1)}(x_{k-1}) = \phi_\star$. Therefore, an optimal limit point exists, be it x_∞. Since $W_k \to 0$ by assertion 3.5*(ii)* and since $\mathcal{L}_k(x_\infty)$ converges, it follows that $\frac{1}{2}\|x_k-x_\infty\|^2 = \mathcal{L}_k(x_\infty)-W_k$ converges as well. Since along a subsequence $\frac{1}{2}\|x_k-x_\infty\|^2$ converges to zero, necessarily $\lim_{k\to\infty}\frac{1}{2}\|x_k-x_\infty\|^2 = 0$, proving that the entire sequence $(x_k)_{k\in\mathbb{N}}$ converges to x_∞.

♢ *Case 2: $\varphi^{(1)}(x_k) < \phi_\star$ holds infinitely often.*

Recall the index i_k and the set K' defined in (3.14) and (3.15). Having established boundedness of the entire sequence, limit points of $(x_{i_k})_{k\in\mathbb{N}}$ exist and, as shown above, belong to $\mathcal{X}^{(2)}$. Moreover, it follows from (3.14) that $\limsup_{k\to\infty}\varphi^{(1)}(x_{i_k}) \leq \phi_\star$, and a lower semicontinuity argument then yields that all the limit points of $(x_{i_k})_{k\in\mathbb{N}}$ attain the optimal cost and are therefore optimal. Since $x_{i_k} - x_{i_k+1} \to 0$ by Lemma 3.4*(iii)*, the same is also true for $(x_{i_k+1})_{k\in\mathbb{N}}$, and in particular $\operatorname{dist}(x_{i_k+1},\mathcal{X}_\star) \to 0$ as $k\to\infty$.

For each k let $\bar{x}_k := \Pi_{\mathcal{X}_\star} x_k$, which is well defined since $\mathcal{X}_\star \neq \emptyset$ is closed and convex. Recalling that $\mathcal{L}_k(x_\star) = W_k + \frac{1}{2}\|x_k - x_\star\|^2$, for $k' \in K'$ we have

$$\begin{aligned}
\tfrac{1}{2}\operatorname{dist}(x_{k'},\mathcal{X}_\star)^2 &\leq \tfrac{1}{2}\|x_{k'}-\bar{x}_{i_{k'}+1}\|^2 = \mathcal{L}_{k'}(\bar{x}_{i_{k'}+1}) - W_{k'} \\
&\overset{\text{by (3.18)}}{\leq} \mathcal{L}_{i_{k'}+2}(\bar{x}_{i_{k'}+1}) - W_{k'} = \tfrac{1}{2}\|x_{i_{k'}+2}-\bar{x}_{i_{k'}+1}\|^2 + W_{i_{k'}+2} - W_{k'} \\
&\leq \tfrac{1}{2}\big(\operatorname{dist}(x_{i_{k'}+1},\mathcal{X}_\star) + \|x_{i_{k'}+1}-x_{i_{k'}+2}\|\big)^2 + W_{i_{k'}+2} - W_{k'} \to 0
\end{aligned}$$

as $K' \ni k' \to \infty$, where the limit follows from Lemma 3.4*(iii)*, the vanishing of $(W_k)_{k\in\mathbb{N}}$ established in assertion 3.5*(ii)*, and the fact that $i_k \to \infty$ as $k\to\infty$. Since $K' \cup \{i_k, i_k+1 \mid k\in\mathbb{N}\} = \mathbb{N}$, we conclude that all the limit points of $(x^k)_{k\in\mathbb{N}}$ are optimal. □

4 Simulations

In this section the performance of the proposed algorithms is evaluated through a series of simulations on standard problems on both synthetic data and standard datasets from the LIBSVM dataset [10]. All the algorithms are implemented in the Julia programming language and are available online.[4] An overview of the algorithms included in the simulations is provided in the following subsection.

In accounting for the difference in iteration complexity among the methods, the simulations report the progress against the number of calls to $\nabla f^{(2)}$, since in all problems $\nabla f^{(1)}$ and proximal operations have negligible cost. As explained in Section 4.1.1, this criterion favors the method SEDM, as it ignores the cost of the backtracks which involve function evaluations. On the contrary, all the backtracking steps included in adaBiM, which involve gradient evaluations, are fully accounted for in the comparisons.

[4] https://github.com/pylat/adaptive-bilevel-optimization

4.1 Compared algorithms

When applicable, other than adaBiM and staBiM the algorithms involved in the simulations are SEDM, BiGSAM, and iterative-3D. For iterative-3D, $\sigma_k = 1/(k+1)^2$ was used, for Bi-SG-II, $\sigma_k = 1/(k+1)^p$ with $p = 0.95$, and $\sigma_k = 1/k+1$ was adopted for the rest. Although only BiGSAM and Bi-SG-II require $\sigma_k \in (0, 1]$, this limitation was applied across all methods to maintain more uniform comparisons.

It is also worth noting that adaBiM is not sensitive to the choice of initial stepsize α_0, as future values are automatically adjusted during the iterations. In all the simulations, the parameter $\nu = 0.99$ was used, and the parameter $\alpha_{\max}$ appearing in (2.5) of adaBiM was set as a large constant; as remarked before, it is only of theoretical significance. As we will see, the same parameter in SEDM is instead crucial for dictating the algorithmic performance. These facts are better detailed in the following brief description of the algorithms compared against in the simulations.

4.1.1 Solodov's explicit descent method (SEDM-r)

This is Solodov's explicit descent method [37, Alg. 2.1] already outlined in (1.4). In the simulations, the suffix "-r" is used to distinguish different choices for the value of $\widehat{\alpha}_0$ therein; namely, whenever $f^{(2)}$ is $L_{f^{(2)}}$-Lipschitz differentiable, we set $\widehat{\alpha}_0 := \frac{r}{L_{f^{(2)}}}$. In addition to Assumption I, the algorithm requires:

- $g^{(1)} = 0$;
- $g^{(2)} = \delta_D$ for a nonempty, closed, and convex set $D \subseteq \mathbb{R}^n$.

The backtracks involved in SEDM do not require additional gradient evaluations, but instead require function evaluations which are not reflected in the comparisons in terms of total number of gradients. For this reason, in the top row of Figure 1 we provided a sample plot for three selected applications demonstrating the higher number of backtracks that it incurs compared to adaBiM. As evident in the figures, in practice SEDM is sensitive to parameter tuning; while selecting a larger $\widehat{\alpha}_0$ can lead to larger stepsizes and consequently faster convergence speed in terms of number of iterations (gradient evaluations), it results in a higher number of backtracks (each requiring one cost evaluation), and vice versa.

In all the simulations, both for SEDM and adaBiM we used the backtrack parameter $\eta = 1/2$, and the linesearch related parameter $\nu = 0.99$.

4.1.2 Bilevel gradient sequential averaging method (BiGSAM)

Proposed in [35], BiGSAM addresses strongly convex bilevel methods under global Lipschitz differentiability assumptions. Specifically, in addition to Assumption I the algorithm requires that

- $f^{(1)}$ is $L_{f^{(1)}}$-Lipschitz differentiable and $\mu_{f^{(1)}}$-strongly convex;
- $f^{(2)}$ is $L_{f^{(2)}}$-Lipschitz differentiable;
- $g^{(1)} = 0$.

BiGSAM iterates

$$\begin{cases} x_{k+1}^{(1)} = x_k^{(1)} - \alpha^{(1)} \nabla f^{(1)}\left(x_k^{(1)}\right) \\ x_{k+1}^{(2)} = \operatorname{prox}_{\alpha^{(2)} g^{(2)}}\left(x_k^{(1)} - \alpha^{(2)} \nabla f^{(2)}\left(x_k^{(2)}\right)\right) \\ x_{k+1} = \sigma_{k+1} x_{k+1}^{(1)} + (1 - \sigma_{k+1}) x_{k+1}^{(2)}, \end{cases}$$

where $\alpha^{(1)} \leq \frac{2}{L_{f^{(1)}} + \mu_{f^{(1)}}}$, $\alpha^{(2)} \leq \frac{1}{L_{f^{(2)}}}$. Although in BiGSAM the sequence $(\sigma_k)_{k \in \mathbb{N}}$ has a different interpretation that the one in (1.2), it must still comply with (1.3) and in addition $\sigma_1 \leq 1$. For this reason we opted to use the same notation.

4.1.3 Iterative-3D

This method, presented in [16], is specialized to linear inverse problems and operating on the dual formulation, which complicates the comparison with the other methods. A strongly convex upper layer cost is required, but the iterations only involve gradient (and not proximal) evaluations on its (Lipschitz-differentiable) conjugate. Similarly, the lower level cost is an infimal convolution between a prox-friendly and a strongly convex function, making its dual the sum of a prox-friendly and a Lipschitz-differentiable terms. The requirements on the primal formulation are roughly as follows:

- $\varphi^{(1)}$ is $\mu_{\varphi^{(1)}}$-strongly convex;
- $\varphi^{(2)}$ is a coercive "data-fit" function.

In referring the reader to [16] for a rigorous account on the problem formulation and its requirements, we point out that among our simulations iterative-3D is only applicable to the linear inverse problem of Section 4.2.2 with ℓ^2-norm upper layer cost. In that setting, initializing with $x_0 \in \operatorname{range} A^\top$ the method performs the following iterations

$$x_{k+1} = x_k - \gamma \nabla f_k(x_k) = x_k - A^\top (Ax_k - b) - \gamma \sigma_k x_k,$$

each increasing the total gradient count by one. Remarkably, in this setting it does not constrain $(\sigma_k)_{k\in\mathbb{N}}$ to a nonsummable decay; see [16, Rem. 10]. For this reason, in the simulations inverse penalties $\sigma_k = 1/(k+1)^2$ were used for iterative-3D, while $\sigma_k = 1/k+1$ for all other methods.

4.1.4 Bi-Sub-Gradient method (Bi-SG-II)

Proposed in [27], this method can also cope with nondifferentiable terms in the upper level. It comes in two versions, depending on the type of operations on the upper level; we here consider the second one, as it relies on milder assumptions and is compatible with our proximal-gradient setting (the first one involves subgradient operations on the upper level). The standing assumptions are the following:

- $f^{(1)}$ is $L_{f^{(1)}}$-Lipschitz differentiable;
- $f^{(2)}$ is $L_{f^{(2)}}$-Lipschitz differentiable;
- $\varphi^{(1)}$ is coercive.

(An additional technical assumption of real valuedness of $\varphi^{(1)}$ is also imposed which nevertheless does not cause any loss of generality.) Notice that both our Assumptions I.A3 and I.A4 are implied by (the existence of solutions and) the coercivity requirement on $\varphi^{(1)}$. The method alternates proximal-gradient operations with constant stepsize:

$$\begin{cases} y_{k+1} = \operatorname{prox}_{\alpha^{(2)} g^{(2)}} \left(x_k - \alpha^{(2)} \nabla f^{(2)}(x_k) \right) \\ x_{k+1} = \operatorname{prox}_{\sigma_{k+1}\alpha^{(1)} g^{(1)}} \left(y_{k+1} - \sigma_{k+1}\alpha^{(1)} \nabla f^{(1)}(y_{k+1}) \right), \end{cases}$$

with $\alpha^{(1)} = \frac{1}{\max\{1, L_{f^{(1)}}\}}$, $\alpha^{(2)} = \frac{1}{L_{f^{(2)}}}$, and $\sigma_{k+1} = \frac{1}{(k+1)^p}$ for some $p \in (\frac{1}{2}, 1)$. As suggested in [27], $p = 0.95$ was used in all the simulations.

4.2 Numerical experiments

We compare the algorithms on three benchmark bilevel problems with Lipschitz differentiable and strongly convex upper level cost $\varphi^{(1)}$, so as to satisfy the requirements of all the algorithms being compared. For two of these we also consider minimum ℓ^1-norm versions in which the upper level $\varphi^{(1)} = \|\cdot\|_1$ is neither smooth nor strongly convex (but is nevertheless coercive). For these latter ones, only adaBiM, staBiM and Bi-SG-II are applicable.

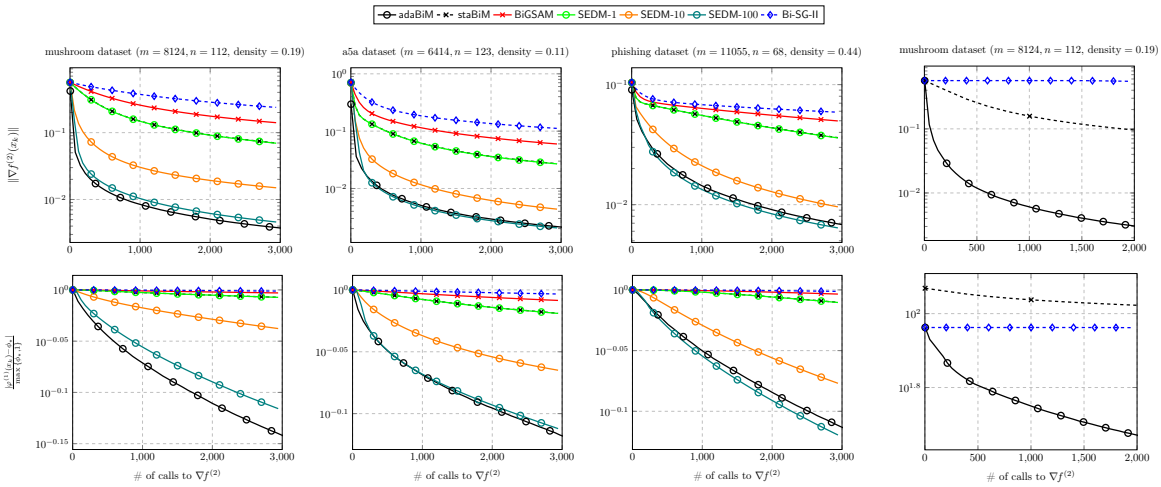
4.2.1 Logistic regression

We consider the logistic regression problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \ \varphi^{(1)}(x) \tag{4.1a}$$

$$\text{subject to}\ \ x \in \underset{w\in\mathbb{R}^n}{\arg\min}\left\{\tfrac{1}{m}\textstyle\sum_{i=1}^m \left(y_i \log(s_i(w)) + (1-y_i)\log(1-s_i(w))\right)\right\}, \tag{4.1b}$$

where m, n are the number of samples and features, the pair $a_i \in \mathbb{R}^{n+1}$ denotes the i-th sample (up to absorbing the bias terms), $y_i \in \{-1, 1\}$ is the associated label, and $s_i(x) = (1+\exp(-a_i^\top x))^{-1}$ is the logistic sigmoid function. In the simulations we used $\varphi^{(1)} = \frac{1}{2}\|\cdot\|^2$ (Figure 2a) and $\varphi^{(1)} = \|\cdot\|_1$ (Figure 2b); for the latter, only one dataset is reported, as the plots for other ones are very similar. Note also that in the simulations for adaBiM, staBiM and Bi-SG-II we set $g^{(1)} = \varphi^{(1)}$ and $f^{(1)} \equiv 0$. For other methods (applicable only when $\varphi^{(1)} = \frac{1}{2}\|\cdot\|^2$) the upper level cost is captured using $f^{(1)}$ with $L_{f^{(1)}} = 1$ and its strong convexity modulus equal to 1 (in the case of BiGSAM). For methods that require Lipschitz modulus of $\nabla f^{(2)}$, $L_{f^{(2)}} = \frac{1}{4m}\|A\|^2$ was used where A is the data matrix that is the concatenation of $(a_j)_{j=1,\dots,m}$.



(a) $\varphi^{(1)} = \frac{1}{2}\|\cdot\|^2$ (b) $\varphi^{(1)} = \|\cdot\|_1$

Figure 2: Logistic regression problems of Section 4.2.1 with minimum ℓ^p-norm solution, $p = 1, 2$.

4.2.2 Linear inverse problems with simulated data

In a series of experiments we consider the special cases of the following problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \ \varphi^{(1)}(x) \tag{4.2a}$$

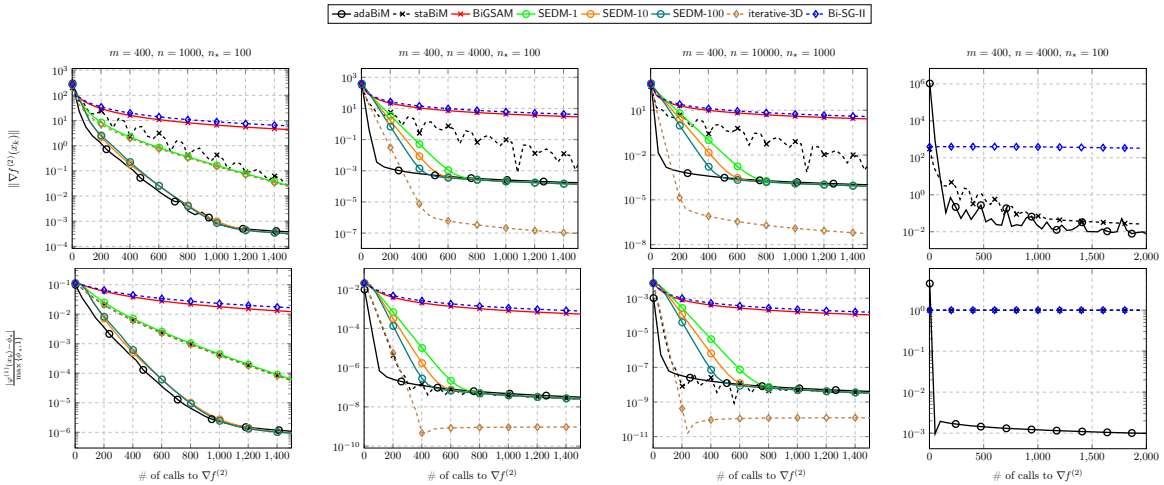
$$\text{subject to}\ \ x \in \underset{w\in\mathbb{R}^n}{\arg\min}\ \tfrac{1}{2}\|Aw - b\|^2, \tag{4.2b}$$

where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are generated based on the procedure described in [29, §6], and $n_\star$ denotes the number of nonzero elements of the solution. For the upper level cost $\varphi^{(1)}$, we consider two sets of experiments:

(i) $\varphi^{(1)} = \frac{1}{2}\|\cdot\|^2$, corresponding to the Moore–Penrose solution, see Figure 3a;

(ii) least ℓ^1-norm solutions corresponding to $\varphi^{(1)} = \|\cdot\|_1$, see Figure 3b (the behavior of the algorithms is consistent with these plots for other values of m, n).

As also done for the logistic regression problems, for adaBiM, staBiM and Bi-SG-II we set $g^{(1)} = \varphi^{(1)}$ and $f^{(1)} \equiv 0$; for other methods the (smooth) upper level cost is captured by using $f^{(1)}$ with $L_{f^{(1)}} = 1$ and its strong convexity modulus equal to 1 (in the case of BiGSAM). For methods that require Lipschitz modulus of $\nabla f^{(2)}$, $L_{f^{(2)}} = \|A\|^2$.



(a) $\varphi^{(1)} = \frac{1}{2}\|\cdot\|^2$ (b) $\varphi^{(1)} = \|\cdot\|_1$

Figure 3: Linear inverse problems of Section 4.2.2 with minimum ℓ^p-norm solution, $p = 1, 2$.

4.2.3 Solution of integral equations

We consider the solution of integral equations using the setting described in [6, §5.2]. The corresponding bilevel problem is the following:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \tfrac{1}{2}\|x\|_Q^2 \tag{4.3a}$$

$$\text{subject to} \quad x \in \underset{w \geq 0}{\arg\min}\, \tfrac{1}{2}\|Aw - b\|^2. \tag{4.3b}$$

The data matrix A in (4.2) is generated using *philips, foxgood, baart* functions. Let L denote the discrete gradient operator, and let $Q_1 = L^\top L$ and $Q = Q_1 + \mathrm{I}$. In the simulations for BiGSAM and SEDM, the upper level cost is captured using $f^{(1)}$, while for adaBiM, staBiM and Bi-SG-II we used $f^{(1)} = \frac{1}{2}\langle x, Q_1 x\rangle$ and $g^{(1)} = \frac{1}{2}\|\cdot\|^2$. (As a rule of thumb, considering formulating the problem using

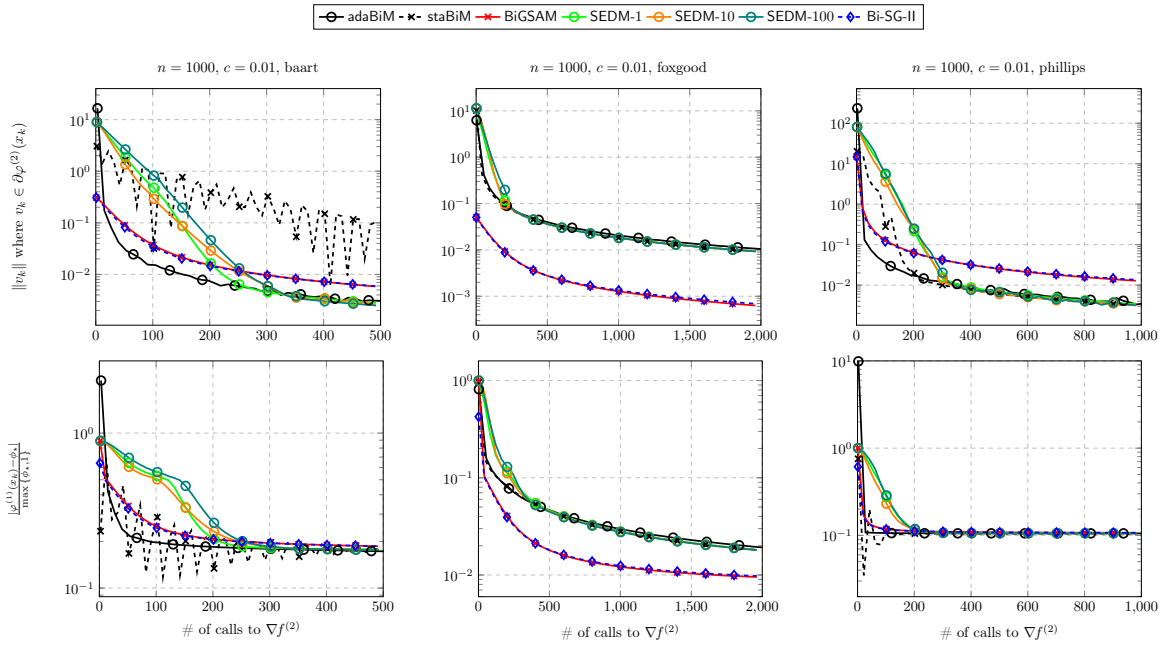


Figure 4: Solution of integral equations

the proximable term is preferable, a tweak that only our methods and Bi-SG-II can take advantage of, being the only ones that allow proximable terms in the upper level.) By using the calculus rule of [4, Prop. 24.8(i)], the proximal mapping of $g_k = \frac{\sigma_k}{2}\|\cdot\|^2 + g^{(2)}$ is given by

$$\operatorname{prox}_{\alpha_k g_k}(u) = \operatorname{prox}_{\frac{\alpha_k}{1+\alpha_k\sigma_k} g^{(2)}}\left(\tfrac{1}{1+\alpha_k\sigma_k}u\right).$$

Multiplications by Q_1 involved in calls to $\nabla f^{(1)}$ can efficiently be handled through abstract linear operators and are ignored in the gradient calls count for all methods. In order to compare the methods in a fair manner, in addition to the deviation of the upper level cost from $\phi_\star$, we also plot a measure of optimality for the lower level. For example, in the case of adaBiM, given that $g^{(1)} = \frac{1}{2}\|\cdot\|^2$, it is of immediate verification that

$$v^k := \tfrac{1}{\alpha_{k+1}}(x_k - x_{k+1}) - \sigma_k x_{k+1} + \nabla f^{(2)}(x_{k+1}) - \nabla f_{k+1}(x_k) \in \partial\varphi^{(2)}(x_{k+1}).$$

Similar computation applies to the other methods that are included in the comparisons.

5 Conclusions

This paper considered structured bilevel problems where both the upper and lower level minimizations are split as the sum of a nonsmooth and a (locally) Lipschitz differentiable function. A convergence recipe was developed for proximal gradient updates treating global and local Lipschitzian settings in a unified fashion. The aforementioned recipe depends on three properties for the stepsizes and involves a carefully designed adaptive scheme that builds upon and generalizes adaPGM [25, Alg. 1] to the bilevel setting. Notably, while in the locally Lipschitz setting our scheme involves a linesearch, it prescribes a suitable initialization for the linesearch based on Barzilai-Borwein type estimates, leading to much larger stepsizes compared to existing methods and considerably fewer backtracks in practice. Finally, the favorable convergence properties of the method were confirmed through a series of numerical simulations. Future research directions involve designing adaptive strategies for the inverse penalty parameters σ_k, deriving stopping criteria, and extensions to non-simple and possibly nonconvex bilevel settings. Relaxing the assumptions to (local) Hölder continuity of the gradients of the smooth terms is another promising direction which can leverage on recent developments on adaptive schemes [30].

A Appendix

***Proof of Theorem 2.3** (convergence of staBiM).* Once Properties $\mathsf{p}_\alpha 1$ to $\mathsf{p}_\alpha 3$ are verified, all the claims follow from Lemma 3.4*(iii)* and Theorem 3.5. Property $\mathsf{p}_\alpha 3$ is trivially satisfied for staBiM due to the underlying global Lipschitz continuity, and since $(\sigma_k)_{k\in\mathbb{N}}$ is decreasing. In what follows we consider the iterates generated by staBiM with $\sigma_1 = \sigma$ for some initial inverse penalty $\sigma > 0$. We show that for every k it holds that

$$\alpha_{k+1}\ell_{k+1} \le \nu \quad \text{and} \quad \alpha_1 = \tfrac{\nu}{\sigma_0 L_{f^{(1)}} + L_{f^{(2)}}} \le \alpha_{k+1} \le \min\{\widehat{\alpha}_{k+1}, \alpha_{\max}\},$$

where $\widehat{\alpha}_{k+1}$ is as in (2.5) with $\alpha_{\max} \ge \frac{\nu}{L_{f^{(2)}}}$ and initialization $\alpha_{-1} = \alpha_0 = \alpha_{\max}$ and $\sigma_{-1} = \sigma_0 = \sigma$. Since $\ell_{k+1} \le \sigma_{k+1}L_{f^{(1)}} + L_{f^{(2)}}$, the first inequality is obvious. Similarly, the second inequality follows from the fact that $(\sigma_{k+1})_{k\in\mathbb{N}}$ is decreasing, and thus so is $(\alpha_{k+1})_{k\in\mathbb{N}}$. Notice further that $\alpha_{k+1} \le \frac{\nu}{L_{f^{(2)}}} \le \alpha_{\max}$; moreover, since $\sigma_k\nabla f^{(1)} + \nabla f^{(2)}$ is globally Lipschitz with modulus $\sigma_k L_{f^{(1)}} + L_{f^{(2)}}$, by cocoercivity it holds that

$$\alpha_k \tfrac{L_k^2}{\ell_k} = \alpha_k \tfrac{\|\nabla f_k(x_{k-1}) - \nabla f_k(x_k)\|^2}{\langle \nabla f_k(x_{k-1}) - \nabla f_k(x_k), x_{k-1} - x_k\rangle} \le \alpha_k(\sigma_k L_{f^{(1)}} + L_{f^{(2)}}) = \nu < 1,$$

which as argued in (2.7) implies that the second term in (2.5) is infinite. Since, as already argued, $\alpha_k \le \alpha_{\max}$, to conclude it remains to show that α_{k+1} is also smaller than the first term in (2.5). To this end, observe that

$$\rho_{k+1} \overset{(def)}{=} \frac{\sigma_{k+1}\alpha_{k+1}}{\sigma_k\alpha_k} = \frac{L_{f^{(1)}} + \sigma_k^{-1}L_{f^{(2)}}}{L_{f^{(1)}} + \sigma_{k+1}^{-1}L_{f^{(2)}}} \le 1,$$

where the inequality follows from the fact that $(\sigma_{k+1})_{k\in\mathbb{N}}$ is decreasing. Similarly, $\rho_{k+1} = \frac{\sigma_{k+1}\alpha_{k+1}}{\sigma_k\alpha_k} \geq \frac{\sigma_{k+1}}{\sigma_k} \geq \frac{3}{4}$, where the first inequality follows from the fact that $(\alpha_{k+1})_{k\in\mathbb{N}}$ is increasing and the second one from the constraints on σ_{k+1} prescribed at step I.1. Overall, it follows that $\rho_{k+1} \in [3/4, 1]$ holds for every k. Therefore,

$$\sqrt{\tfrac{\sigma_k}{\sigma_{k-1}}\big(1+\rho_k\big)} \geq \sqrt{\tfrac{3}{4}\big(1+\tfrac{3}{4}\big)} > 1 \geq \rho_{k+1} \overset{(def)}{=} \tfrac{\sigma_{k+1}\alpha_{k+1}}{\sigma_k\alpha_k},$$

where the first inequality again follows from the bounds on σ_{k+1} at step I.1. Rearranging yields the sought inequality $\alpha_{k+1} \leq \sqrt{\frac{\sigma_k}{\sigma_{k-1}}\big(1+\rho_k\big)}\frac{\sigma_k}{\sigma_{k+1}}\alpha_k$. □

***Proof of Theorem 2.4** (convergence of adaBiM).* We first state a simple lemma to justify the enlargement in the definition of the set $\mathcal{V}$.

Lemma A.1. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper lsc and convex, and given $0 < \alpha^+ \leq \alpha^-$ and $d \in \mathbb{R}^n$ let $z^\pm := \operatorname{prox}_{\alpha^\pm h}(x - \alpha^\pm d)$. Then, denoting $\eta := \alpha^+/\alpha^- \in (0,1]$, it holds that*

$$\|z^- - z^+\| \leq \tfrac{1-\eta}{\eta}\|x - z^+\| \quad \textit{and} \quad \|z^- - z^+\|^2 \leq \tfrac{1-\eta}{1+\eta}\|x - z^-\|^2 - \tfrac{1-\eta}{1+\eta}\|x - z^+\|^2.$$

Proof. The proximal characterization of $z^\pm$ reads

$$\alpha^\pm h(z^\pm) + \tfrac{1}{2}\|z^\pm - x + \alpha^\pm d\|^2 \leq \alpha^\pm h(y) + \tfrac{1}{2}\|y - x + \alpha^\pm d\|^2 - \tfrac{1}{2}\|y - z^\pm\|^2 \quad \forall y \in \mathbb{R}^n.$$

By considering $y = z^\mp$ and summing the resulting inequalities we obtain

$$\begin{aligned}&\tfrac{1}{2}\|z^+ - x + \alpha^+ d\|^2 + (\alpha^- - \alpha^+)\big(h(z^-) - h(z^+)\big) + \tfrac{1}{2}\|z^- - x + \alpha^- d\|^2 \\ &\leq \tfrac{1}{2}\|z^- - x + \alpha^+ d\|^2 + \tfrac{1}{2}\|z^+ - x + \alpha^- d\|^2 - \|z^+ - z^-\|^2,\end{aligned}$$

which after expanding the squares and suitably rearranging results in

$$\tfrac{1}{\alpha^- - \alpha^+}\|z^+ - z^-\|^2 \leq h(z^+) - h(z^-) + \langle d, z^+ - z^-\rangle;$$

using the fact that $\tilde{\nabla}h(z^+) := \frac{1}{\alpha^+}(x - z^+) - d \in \partial h(z^+)$ we may further upper bound this as

$$\leq \langle \tilde{\nabla}h(z^+) + d, z^+ - z^-\rangle = \tfrac{1}{\alpha^+}\langle x - z^+, z^+ - z^-\rangle.$$

The first inequality in the statement now follows from the Cauchy-Schwarz inequality. The second one instead follows by expanding the inner product into three square norms and suitably rearranging. □

♠ *2.4(i)* We begin by showing that the stepsize sequence is well defined and strictly positive. Let $\alpha_{k+1,i}$, $x_{k+1,i}$, and $\ell_{k+1,i}$ respectively denote the value of α_{k+1}, x_{k+1}, and ℓ_{k+1} after i many backtracks, $i \geq 0$; in particular, $\alpha_{k+1,i} = \min\{\eta^i\widehat{\alpha}_{k+1}, \alpha_{\max}\}$ and $x_{k+1,i} = \operatorname{prox}_{\alpha_{k+1,i}g_{k+1}}(x_k - \alpha_{k+1,i}\nabla f_{k+1}(x_k))$. Let $i_{k+1} \geq 0$ denote the number of backtracks, or, equivalently, of failed attempts; we will show that i_{k+1} is finite for every k; by construction, this will imply that $x_{k+1} = x_{k+1,i_{k+1}}$. All the attempts $\{x_{k+1,i} \mid \mathbb{N} \ni i \leq i_{k+1}\}$ remain in a convex and compact set $\mathcal{V}_{k+1}$ over which ∇f_{k+1} has finite Lipschitz modulus, be it $L_{f_{k+1},\mathcal{V}_{k+1}}$; as such, one has that $\alpha_{k+1,i}\ell_{k+1,i} \leq \alpha_{k+1,0}\eta^i L_{f_{k+1},\mathcal{V}_{k+1}} \to 0$ as $i \to \infty$, implying that (2.6) is satisfied for i large enough.

We thus proceed by induction to show that each iteration is well defined, that is, that $\widehat{\alpha}_{k+1}$ is (well defined and) strictly positive. Equivalently, in view of the update (2.5) it suffices to show that $1 - 4\Big(1 - \frac{\sigma_k}{\sigma_{k-1}}\Big)\alpha_k\ell_k^{(2)} > 0$ holds for all k. For $k = 0$ this is true because $1 - \frac{\sigma_0}{\sigma_{-1}} = 0$ by initialization. Suppose that the claim holds for k; then, $\alpha_k\ell_k^{(2)} \leq \alpha_k\big(\sigma_k\ell_k^{(1)} + \ell_k^{(2)}\big) = \alpha_k\ell_k \leq \nu$, where the last inequality owes to the linesearch condition (2.6) at the previous step, which holds by inductive hypothesis. Therefore,

$$1 - 4\Big(1 - \tfrac{\sigma_k}{\sigma_{k-1}}\Big)\alpha_k\ell_k^{(2)} \geq 1 - 4\nu\Big(1 - \tfrac{\sigma_k}{\sigma_{k-1}}\Big) \geq 1 - \nu > 0, \tag{A.1}$$

where the last inequality owes to the bound $\sigma_k \geq \frac{3}{4}\sigma_{k-1}$ prescribed at step II.1.

This concludes the proof of the well definedness of the iterations. Notice also that Properties $\mathsf{p}_\alpha 1$ and $\mathsf{p}_\alpha 2$ hold by construction, thereby ensuring through Lemmas 3.4*(iii)* and 3.4*(iv)* that $(x^k)_{k\in\mathbb{N}}$ is bounded and $\|x^{k+1} - x^k\| \to 0$. In particular, $r = \sup_{k\in\mathbb{N}}\|x^{k+1} - x^k\| < \infty$, implying that $\mathcal{V}$ as in the statement is bounded. The enlargement in its definition also guarantees that, in addition to all the iterates x_k, $\mathcal{V}$ also contains x_{k,i_k^-} where $i_k^- = [i_k - 1]_+$ ($\mathcal{V}$ contains every x_k and all the last failed attempts whenever the linsearch is not passed at the first trial). This follows from the first inequality in Lemma A.1 with $x = x_k$, $h = g_{k+1}$, $d = \nabla f_{k+1}(x_k)$, $\alpha^+ = \alpha_{k+1}$ and $\alpha^- = \alpha_{k+1,i_k^-} = \alpha_{k+1}/\eta$, for which $z^+ = x_{k+1}$ and $z^- = x_{k+1,i_{k+1}^-}$.

Whenever $\widehat{\alpha}_{k+1} < \frac{\nu}{L_{f_0,\mathcal{V}}}$ the initial stepsize $\widehat{\alpha}_{k+1}$ already complies with (2.6), and therefore $\alpha_{k+1} = \min\{\widehat{\alpha}_{k+1}, \alpha_{\max}\} = \widehat{\alpha}_{k+1}$ (the last identity follows from the fact that $\alpha_{\max} \geq \frac{1}{\ell_0} \geq \frac{1}{L_{f_0,\mathcal{V}}}$). Otherwise, the failure of the second-last backtrack implies that $\alpha_{k+1} = \eta\alpha_{k+1,i_k-1} > \frac{\eta\nu}{L_{f_0,\mathcal{V}}}$. Either way, this shows that

$$\alpha_{k+1} \geq \min\left\{\widehat{\alpha}_{k+1}, \tfrac{\eta\nu}{L_{f_0,\mathcal{V}}}\right\} \quad \forall k. \tag{A.2}$$

We show by induction that $\alpha_k \geq \alpha_{\min}$. For $k = 1$, by the choice α_{-1} used in the initialization of adaBiM and since $\sigma_0 = \sigma_{-1}$, we have

$$\begin{aligned}\widehat{\alpha}_1 &= \tfrac{\sigma_0}{\sigma_1}\alpha_0 \min\left\{\sqrt{1+\tfrac{\alpha_0}{\alpha_{-1}}}, \tfrac{1}{2\sqrt{[\alpha_0^2 L_0^2 - \alpha_0\ell_0]_+}}\right\} \geq \min\left\{\alpha_0\sqrt{1+\tfrac{\alpha_0}{\alpha_{-1}}}, \tfrac{1}{2L_0}\right\} \\ &\geq \begin{Bmatrix} \min\left\{\sqrt{2}\alpha_0, \frac{1}{2L_0}\right\} & \text{if } \alpha_0\ell_0 \geq 1/2 \\ \min\left\{\frac{1}{\ell_0}, \frac{1}{2L_0}\right\} & \text{otherwise}\end{Bmatrix} \geq \tfrac{1}{2L_{f_0,\mathcal{V}}} \geq \alpha_{\min}.\end{aligned} \tag{A.3}$$

Therefore, by (A.2) also $\alpha_1 \geq \alpha_{\min}$. Suppose that the claim holds up to k. We consider two cases.

- If $\widehat{\alpha}_{k+1}$ equals the second element in (2.5), then

$$\widehat{\alpha}_{k+1} = \alpha_k \tfrac{\sigma_k}{\sigma_{k+1}} \frac{\sqrt{1-4\left(1-\frac{\sigma_k}{\sigma_{k-1}}\right)\alpha_k\ell_k^{(2)}}}{2\sqrt{\alpha_k^2 L_k^2 - \alpha_k\ell_k}} \geq \alpha_k \frac{\sqrt{1-\nu}}{2\sqrt{\alpha_k^2 L_k^2 - \alpha_k\ell_k}} \geq \tfrac{\sqrt{1-\nu}}{2L_k} \geq \tfrac{\sqrt{1-\nu}}{2L_{f_0,\mathcal{V}}}$$

  which combined with (A.2) yields that $\alpha_{k+1} \geq \min\left\{\frac{\sqrt{1-\nu}}{2L_{f_0,\mathcal{V}}}, \frac{\eta\nu}{L_{f_0,\mathcal{V}}}\right\} \geq \alpha_{\min}$.

- Suppose that $\widehat{\alpha}_{k+1}$ equals the first element in (2.5). If $\alpha_{k+1} < \widehat{\alpha}_{k+1}$, then by (A.2) the lower bound $\alpha_{k+1} \geq \frac{\eta\nu}{L_{f_0,\mathcal{V}}} \geq \alpha_{\min}$ holds. If instead $\alpha_{k+1} = \widehat{\alpha}_{k+1}$, then

$$\alpha_{k+1} = \widehat{\alpha}_{k+1} = \tfrac{\sigma_k}{\sigma_{k+1}}\alpha_k\sqrt{\tfrac{\sigma_k}{\sigma_{k-1}}(1+\rho_k)} \geq \alpha_k\sqrt{\tfrac{3}{4}(1+\rho_k)}. \tag{A.4}$$

  We consider two subcases.
  - If $\alpha_k < \widehat{\alpha}_k$, then, as argued above, $\alpha_k \geq \frac{\eta\nu}{L_{f_0,\mathcal{V}}}$ and by (A.4) $\alpha_{k+1} \geq \alpha_k\sqrt{\frac{3}{4}(1+\rho_k)} \geq \frac{\sqrt{3}}{2}\frac{\eta\nu}{L_{f_0,\mathcal{V}}} \geq \alpha_{\min}$.
  - If instead $\alpha_k = \widehat{\alpha}_k$, then observe that $\rho_k \overset{(def)}{=} \frac{\sigma_k\alpha_k}{\sigma_{k-1}\alpha_{k-1}} = \frac{\sigma_k\widehat{\alpha}_k}{\sigma_{k-1}\alpha_{k-1}} = \sqrt{\frac{\sigma_{k-1}}{\sigma_{k-2}}(1+\rho_{k-1})}$, hence

$$\begin{aligned}\alpha_{k+1} = \widehat{\alpha}_{k+1} &= \tfrac{\sigma_k}{\sigma_{k+1}}\alpha_k\sqrt{\tfrac{\sigma_k}{\sigma_{k-1}}(1+\rho_k)} \\ &= \tfrac{\sigma_k}{\sigma_{k+1}}\alpha_k\sqrt{\tfrac{\sigma_k}{\sigma_{k-1}}\left(1+\sqrt{\tfrac{\sigma_{k-1}}{\sigma_{k-2}}(1+\rho_{k-1})}\right)} \geq \alpha_k\sqrt{\tfrac{3}{4}\left(1+\tfrac{\sqrt{3}}{2}\right)} > \alpha_k \geq \alpha_{\min},\end{aligned}$$

    where the first inequality uses $\sigma_j \geq \sigma_{j+1} \geq \frac{3}{4}\sigma_j$ for every j, and the last one holds by induction.

  We showed that in either case $\alpha_{k+1} \geq \alpha_{\min}$ as claimed.

♠ 2.4*(ii)* and 2.4*(iii)* Having established Property $\mathsf{p}_\alpha 3$ and since Properties $\mathsf{p}_\alpha 1$ and $\mathsf{p}_\alpha 2$ hold by design, both assertions follow immediately from Lemma 3.4*(iii)* and Theorem 3.5. □

References

[1] H. Attouch. Viscosity solutions of minimization problems. *SIAM J. Optim.*, 6(3):769–806, 1996.

[2] M.A. Bahraoui and B. Lemaire. Convergence of diagonally stationary sequences in convex optimization. *Set-Valued Anal.*, 2:49–61, 1994.

[3] J. Barzilai and J.M. Borwein. Two-point step size gradient methods. *IMA J. Numer. Anal.*, 8(1):141–148, 1988.

[4] H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces.* CMS Books Math. Springer, 2017.

[5] A. Beck. *First-Order Methods in Optimization.* SIAM, Philadelphia, PA, 2017.

[6] A. Beck and S. Sabach. A first order method for finding minimal norm-like solutions of convex optimization problems. *Math. Program.*, 147(1-2):25–46, 2014.

[7] G. Bigi, L. Lampariello, and S. Sagratella. Combining approximation and exact penalty in hierarchical programming. *Optim.*, 71(8):2403–2419, 2022.

[8] Z. Borsos, M. Mutny, and A. Krause. Coresets via bilevel optimization for continual learning and streaming. *Adv. Neural Inf. Process. Syst.*, 33:14879–14890, 2020.

[9] A. Cabot. Proximal point algorithm controlled by a slowly vanishing term: applications to hierarchical minimization. *SIAM J. Optim.*, 15(2):555–572, 2005.

[10] C. Chang and C. Lin. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. (TIST))*, 2:1–27, 2011.

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

[12] S. Dempe. Bilevel optimization: theory, algorithms, applications and a bibliography. *Bilevel Optim.: Adv. Next Challenges*, pages 581–672, 2020.

[13] L. Doron and S. Shtern. Methodology and first-order algorithms for solving nonsmooth and non-strongly convex bilevel optimization problems. *Math. Program.*, pages 1–38, 2022.

[14] F. Facchinei, J. Pang, G. Scutari, and L. Lampariello. VI-constrained hemivariational inequalities: distributed algorithms and power control in ad-hoc networks. *Math. Program.*, 145(1-2):59–96, 2014.

[15] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *Int. Conf. Mach. Learn.*, pages 1568–1577. PMLR, 2018.

[16] G. Garrigos, L. Rosasco, and S. Villa. Iterative regularization via dual diagonal descent. *J. Math. Imaging Vision*, 60:189–215, 2018.

[17] R. Grazzi, M. Pontil, and S. Salzo. Bilevel optimization with a lower-level contraction: Optimal sample complexity without warm-start. *J. Mach. Learn. Res.*, 24(167):1–37, 2023.

[18] W. Guan and W. Song. A first-order method for solving bilevel convex optimization problems in Banach space. *Optim.*, pages 1–26, 2023.

[19] E.S. Helou and L.E. Simões. ϵ-subgradient algorithms for bilevel convex optimization. *Inverse Probl.*, 33(5):055020, apr 2017.

[20] M. Hong, H. Wai, Z. Wang, and Z. Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM J. Optim.*, 33(1):147–180, 2023.

[21] 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 *Int. Conf. Artif. Intell. Stat.*, pages 10305–10323. PMLR, 2023.

[22] H.D. Kaushik and F. Yousefian. A method with convergence rates for optimization problems with variational inequality constraints. *SIAM J. Optim.*, 31(3):2171–2198, 2021.

[23] L. Lampariello, G. Priori, and S. Sagratella. On the solution of monotone nested variational inequalities. *Math. Methods Oper. Res.*, 96(3):421–446, 2022.

[24] P. Latafat, A. Themelis, and P. Patrinos. On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms. *arXiv:2311.18431*, 2023.

[25] P. Latafat, A. Themelis, L. Stella, and P. Patrinos. Adaptive proximal algorithms for convex optimization under local Lipschitz continuity of the gradient. *arXiv:2301.04431*, 2023.

[26] Y. Malitsky and K. Mishchenko. Adaptive gradient descent without descent. In *Proc. 37th Int. Conf. Mach. Learn.*, volume 119, pages 6702–6712. PMLR, 13- 2020.

[27] R. Merchav and S. Sabach. Convex bi-level optimization problems with non-smooth outer objective function, 2023.

[28] A. Moudafi. Viscosity approximation methods for fixed-points problems. *J. Math. Anal. Appl.*, 241(1):46–55, 2000.

[29] Y. Nesterov. Gradient methods for minimizing composite functions. *Math. Program.*, 140(1):125–161, aug 2013.

[30] K.A. Oikonomidis, E. Laude, P. Latafat, A. Themelis, and P. Patrinos. Adaptive proximal gradient methods are universal without approximation. *arXiv:2402.06271*, 2024.

[31] F. Pedregosa. Hyperparameter optimization with approximate gradient. In *Int. Conf. Mach. Learn.*, pages 737–746. PMLR, 2016.

[32] J. Peypouquet. Coupling the gradient method with a general exterior penalization scheme for convex minimization. *J. Optim. Theory Appl.*, 153:123–138, 2012.

[33] A. Rajeswaran, C. Finn, S.M. Kakade, and S. Levine. Meta-learning with implicit gradients. *Adv. neural Inf. Process. Syst.*, 32, 2019.

[34] R.T. Rockafellar and R.J. Wets. *Variational analysis*, volume 317. Springer, 2011.

[35] S. Sabach and S. Shtern. A first order method for solving convex bilevel optimization problems. *SIAM J. Optim.*, 27(2):640–660, 2017.

[36] M.V. Solodov. A bundle method for a class of bilevel nonsmooth convex minimization problems. *SIAM J. Optim.*, 18(1):242–259, 2007.

[37] M.V. Solodov. An explicit descent method for bilevel convex optimization. *J. Convex Anal.*, 14(2):227, 2007.

[38] A. Themelis, M. Ahookhosh, and P. Patrinos. On the acceleration of forward-backward splitting via an inexact Newton method. In H.H. Bauschke, R.S. Burachik, and D.R. Luke, editors, *Splitting Algorithms, Modern Operator Theory, and Applications*, pages 363–412. Springer, 2019.

[39] H. Xu. Viscosity approximation methods for nonexpansive mappings. *J. Math. Anal. Appl.*, 298(1):279–291, 2004.