Optimization Methods and Software

ISSN: 1055-6788 (Print) 1029-4937 (Online) Journal homepage: https://www.tandfonline.com/loi/goms20

An inertial extrapolation method for convex simple bilevel optimization

Yekini Shehu, Phan Tu Vuong & Alain Zemkoho

To cite this article: Yekini Shehu, Phan Tu Vuong & Alain Zemkoho (2019): An inertial extrapolation method for convex simple bilevel optimization, Optimization Methods and Software, DOI: 10.1080/10556788.2019.1619729

To link to this article: https://doi.org/10.1080/10556788.2019.1619729

Published online: 27 May 2019.

OPTIMIZATION METHODS & SOFTWARE
https://doi.org/10.1080/10556788.2019.1619729

Taylor & Francis
Taylor & Francis Group

An inertial extrapolation method for convex simple bilevel optimization

Yekini Shehu[a*], Phan Tu Vuong[b] and Alain Zemkoho[c]

[a]Department of Mathematics, University of Nigeria, Nsukka, Nigeria; [b]Faculty of Mathematics, University of Vienna, Vienna, Austria; [c]School of Mathematics, University of Southampton, Southampton, UK

ABSTRACT

We consider a scalar objective minimization problem over the solution set of another optimization problem. This problem is known as a *simple bilevel optimization problem* and has drawn a significant attention in the last few years. Our inner problem consists of minimizing the sum of smooth and non-smooth functions while the outer one is the minimization of a smooth convex function. We first formulate and give strong convergence analysis of an inertial algorithm for fixed-point problem of a non-expansive operator in an infinite dimensional Hilbert space. Then we convert the simple bilevel optimization problem to a fixed-point problem of a non-expansive operator in finite dimensional space and design the corresponding algorithm and establish its convergence. Our numerical experiments show that the proposed method in this paper outperforms the currently known best algorithm to solve the class of bilevel optimization problem considered.

ARTICLE HISTORY
Received 17 September 2018
Accepted 10 May 2019

KEYWORDS
Simple bilevel optimization; fixed-point iterative method; inertial extrapolation

1. Introduction

Our main aim in this paper is to solve a scalar objective minimization problem over the solution set of another optimization problem; i.e. precisely, the problem

$$\min h(x) \text{ s.t. } x \in X^* \subseteq \mathbb{R}^n, \tag{1}$$

where $h : \mathbb{R}^n \to \mathbb{R}$ is assumed to be strongly convex and differentiable, while X^* is the non-empty set of minimizers of the classical convex composite optimization problem

$$\min \varphi(x) := f(x) + g(x), \tag{2}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and g, an extended real-valued function on $\mathbb{R}^n$, which can be non-smooth. Problem (1)–(2) was labelled in [13] as *simple bilevel optimization problem*, as opposed to the more general version of the problem (see, e.g. [12]), where the follower's problem (2) is parametric, with the parameter representing the variable controlled by the leader, which is in turn different from the one under the control

CONTACT Alain Zemkoho a.b.zemkoho@soton.ac.uk
*Current address (May 2016 – April 2019): Institute of Mathematics, University of Würzburg, Emil-Fischer-Str. 30, 97074 Würzburg, Germany

of the follower. For more details on the vocabulary and connections of problem (1)–(2) to the standard bilevel optimization problem, see Section 2.1.

A common approach to solve problem (1)–(2) consists of the Tikhonov-type regularization [31] (indirect method), based on solving the following regularized problem

$$\min \varphi_\lambda(x) := \varphi(x) + \lambda h(x) \tag{3}$$

for some $\lambda > 0$. Note that problem (1)–(2) can be traced back to the work by Mangasarian and Meyer [22] in the process of developing efficient algorithms for large scale linear programs. The model emerged in turn as a refinement of the regularization technique introduced by Tikhonov and Arsenin [31]. The underlying idea in the related papers by Mangasarian and his co-authors is called *finite-perturbation property*, which consists of finding a parameter $\bar{\lambda}$ (*Tikhonov perturbation parameter*) such that for all $\lambda \in [0, \bar{\lambda}]$,

$$\arg\min_{x \in X^*} h(x) = \arg\min_{x \in \mathbb{R}^n} \varphi_\lambda(x). \tag{4}$$

This property, initially proven in [22] when the lower-level problem is a linear program, was later extended in [18] to the case where it is a general convex optimization problem.

To the best of our knowledge, the development of solution algorithms specifically tailored to optimization problems of the form (1)–(2) can be traced back to the work by Cabot [11], where a *proximal point method* is proposed to solve the problem and its extension to a simple hierarchical optimization problem with finitely many levels. In contrary to the latter paper, where the approximation scheme is only implicit thus making the method not easy to numerically implement, Solodov [29] proposed an explicit and more tractable proximal point method for problem (1)–(2). Since then, various proximal point algorithms have been developed to solve the problems under different types of frameworks, see, e.g. [9,23,28] and references therein.

Motivated by the results in [7], Sabach and Shtern [28] recently proposed the following scheme (with $x_0 \in \mathbb{R}^n$ as starting point), called *Bilevel Gradient Sequential Averaging Method* (abbreviated as *BiG-SAM*), to solve problem (1)–(2):

$$\begin{aligned} s_n &= \text{prox}_{\lambda g}(x_{n-1} - \lambda \nabla f(x_{n-1})) \\ z_n &= x_{n-1} - \gamma \nabla h(x_{n-1}) \\ x_n &= \alpha_n z_n + (1 - \alpha_n) s_n, \ n \geq 1 \end{aligned} \tag{5}$$

with $\lambda \in (0, 1/L_f]$, $\gamma \in (0, 2/(L_h + \sigma)]$, where L_f and L_h are the Lipschitz constants for the gradients of f and h, and $\{\alpha_n\}$ satisfying the conditions assumed in [33]. Sabach and Shtern [28] obtained a non-asymptotic $O(1/n)$ global rate of convergence in terms of the inner objective function values and showed that BiG-SAM (5) appears simpler and cheaper than the method proposed in [7]. The numerical example in [28] also showed that BiG-SAM (5) outperforms the method in [7] for solving problem (1)–(2). The algorithm in [28] seems to be the most efficient method developed so far for convex simple bilevel optimization problems.

Inspired by recent results on *inertial extrapolation* type algorithms for solving optimization problem (see, e.g. [1,6,8,24] and references therein), our aim in this paper is to solve problems (1)–(2) by introducing an inertial extrapolation step to BiG-SAM (5) (which we

shall call *iBiG-SAM*). The strategy is to provide an equivalent fixed-point formulation to this bilevel optimization problems and then establish the global convergence of the proposed method under reasonable assumptions. The proposed method can be considered as an initial version of the iterative method for solving fixed point problem of non-expansive mapping studied in [33]. Numerical experiments show that the new proposed method outperforms the BiG-SAM (5) introduced in [28].

For the remainder of the paper, first note that there is a striking similarity between the exact penalization model (3) and a corresponding partial penalization approach based on the *partial calmness concept* [35] often used to solve the general bilevel optimization problem. Both approaches seem to have originated from completely different sources and their development also seems to be occurring independently from each other till now. In Section 2.1, we clarify this similarity and discuss some strong relationships between the two problem classes. In Section 2.2, we recall some basic definitions and results that will play an important role in the paper. The proposed method and its convergence analysis are presented in Section 3. Some numerical experiments are given in Section 4. We conclude the paper with some final remarks in Section 5.

2. General context and mathematical tools

2.1. Standard bilevel optimization

In this subsection, we provide a discussion to place the simple bilevel optimization introduced above in a general context of bilevel optimization. To proceed, we consider a simple optimistic version of the latter class of problem, which aligns suitably with problem (1)–(2), i.e.

$$\min_{x,y} h(x,y) \ \text{ s.t. } \ y \in S(x), \tag{6}$$

where $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ represents the upper level objective function and the set-valued mapping S defines the set of optimal solutions of the lower level problem

$$\min_{y} \varphi(x,y) \tag{7}$$

($\varphi : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$) for any fixed upper level variable x. Obviously, problem (1)–(2) is a special case of problem (6)–(7), where the optimal solution of the leader is simply picked among the optimal solutions of the lower level problem, which in turn are obtained without any influence from the leader as it is the case in the latter problem.

On the other hand problem (6)–(7) can be equivalently written as the following optimization problem over an efficient set

$$\min_{x,y} h(x,y) \ \text{ s.t. } \ (x,y) \in E\left(\mathbb{R}^n \times \mathbb{R}^m,\ \bar{\varphi},\ \preccurlyeq\right), \tag{8}$$

where $E(\mathbb{R}^n \times \mathbb{R}^m,\ \bar{\varphi},\ \preccurlyeq)$ denotes the efficient set (i.e. optimal solution set) of the problem of minimizing a multiobjective function $\bar{\varphi}$ (based on φ (7)) over $\mathbb{R}^n \times \mathbb{R}^m$ w.r.t. a certain order relation $\preccurlyeq$; for examples of choices of the latter function and corresponding order relations, see the papers [16,19]. For more details on the relationship between problems (6)–(7) and (8), see [27]. Obviously, an optimization problem over an efficient set is a

generalization of the simple bilevel optimization problem (1)–(2), and has been extensively investigated since the seminal work by Philip [26]; see [34] for a literature review on the topic.

One common approach to transform problem (6)–(7) into a single-level optimization problem is the so-called lower-level optimal value function (LLVF) reformulation

$$\min_{x,y} h(x,y) \ \text{ s.t. } \ \varphi(x,y) \leq \varphi^*(x), \tag{9}$$

where the function $\varphi^*(x) = \min_y \varphi(x,y)$ represents the optimal value function of the lower level problem (7). Recall that this reformulation is an underlying feature in the development of the link (4) between the simple bilevel optimization problem (1)–(2) and the penalized problem (3) as outlined in the corresponding publications; see, e.g. [18]. However, we instead want to point out here an interesting similarity between the finite termination property (4) and the *partial calmness* concept [35] commonly used in the context of standard bilevel optimization. To highlight this, let $(\bar{x}, \bar{y})$ be a local optimal solution of (1)–(2). The problem is partially calm at $(\bar{x}, \bar{y})$ if and only if there exists $\lambda > 0$ such that $(\bar{x}, \bar{y})$ is also a local optimal solution of the penalized problem

$$\min_{x,y} h(x,y) + \lambda \left(\varphi(x,y) - \varphi^*(x)\right). \tag{10}$$

The partial calmness concept does not automatically hold for the simple bilevel optimization problem (1)–(2). To see this, consider the example of convex simple bilevel optimization problem of minimizing $(x-1)^2$ subject to $x \in \arg\min y^2$. It is clear that 0 is the only optimal solution of this problem. But for the corresponding penalized problem (10) to minimize $(x-1)^2 + \lambda x^2$, we can easily check that the optimal solution is the number $x(\lambda) := 1/(1+\lambda)$ for all $\lambda > 0$. Clearly, $x(\lambda) \neq 0$ for all $\lambda > 0$.

It is also important to note that, possibly unlike the finite termination property (4), the partial calmness concept was introduced as a qualification condition to derive necessary optimality conditions for problem (9); see [15,35] for some papers where this concept is used, and also the papers [14,17] for new results on simple bilevel optimization problems from the perspective of standard bilevel optimization.

2.2. Basic mathematical tools

We state the following well-known lemmas which will be used in our convergence analysis in the sequel. In the following H is an arbitrary Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$.

We use the symbol '$a_n \to a$' to denote 'a_n converges strongly to a' while '$a_n \rightharpoonup a$' is used to denote 'a_n converges weakly to a'. Recall that in finite dimensional space $\mathbb{R}^n$, strong convergence coincides with weak convergence.

Lemma 2.1: *The following well-known results hold in H:*

(i) $\|x+y\|^2 = \|x\|^2 + 2\langle x, y\rangle + \|y\|^2, \ \forall x, y \in H$;
(ii) $\|x+y\|^2 \leq \|x\|^2 + 2\langle y, x+y\rangle, \ \forall x, y \in H$;
(iii) $\|tx+sy\|^2 = t(t+s)\|x\|^2 + s(t+s)\|y\|^2 - st\|x-y\|^2, \ \forall x, y \in H, \ s, t \in \mathbb{R}$.

Lemma 2.2 (see, e.g. [32]): *Let $\{a_n\}$ and $\{\gamma_n\}$ be sequences of non-negative real numbers, $\{\alpha_n\}$ a sequence in $(0,1)$ and $\{\sigma_n\}$ a sequence in $\mathbb{R}$ satisfying the following relation:*

$$a_{n+1} \leq (1-\alpha_n)a_n + \sigma_n + \gamma_n,\ n \geq 1.$$

Assume $\sum \gamma_n < \infty$. Then the following results hold:

(i) *If $\sigma_n \leq \alpha_n M$ for some $M \geq 0$, then $\{a_n\}$ is a bounded sequence.*
(ii) *If $\sum \alpha_n = \infty$ and $\limsup \sigma_n/\alpha_n \leq 0$, then $\lim a_n = 0$.*

We state the formal definition of some classes of operators that play an essential role in our analysis in the sequel.

Definition 2.3: An operator $T : H \to H$ is called

(a) *non-expansive* if and only if $\|Tx - Ty\| \leq \|x - y\|$ for all $x, y \in H$;
(b) *averaged* if and only if it can be written as the average of the identity mapping I and a non-expansive operator, i.e. $T := (1-\beta)I + \beta S$ with $\beta \in (0,1)$ and $S : H \to H$ being a non-expansive operator. More precisely, we say that T is β-averaged;
(c) *firmly non-expansive* if and only if $2T-I$ is non-expansive, or equivalently,

$$\langle Tx - Ty, x - y\rangle \geq \|Tx - Ty\|^2, \quad \forall x, y \in H.$$

Alternatively, T is said to be firmly non-expansive if and only if it can be expressed as $T := \frac{1}{2}(I + S)$, where $S : H \to H$ is non-expansive.

We can see from above that firmly non-expansive operators (in particular, projections) are $\frac{1}{2}$-averaged.

Lemma 2.4 ([20]): *Let $T : H \to H$ be a non-expansive operator. Let $\{x_n\}$ be a sequence in H and x be a point in H. Suppose that $x_n \rightharpoonup x$ as $n \to \infty$ and that $x_n - Tx_n \to 0$ as $n \to \infty$. Then, $x \in F(T)$, where $F(T)$ is the set of fixed points of T.*

Next, we provide some relevant properties of averaged operators.

Proposition 2.5 (see, e.g. [10]): *For given operators S, T, and V defined from H to H, the following statements are satisfied:*

(a) *If $T = (1-\alpha)S + \alpha V$ for some $\alpha \in (0,1)$ and if S is averaged and V is non-expansive, then the operator T is averaged.*
(b) *The operator T is firmly non-expansive if and only if the complement $I-T$ is also firmly non-expansive.*
(c) *If $T = (1-\alpha)S + \alpha V$ for some $\alpha \in (0,1)$ and if S is firmly non-expansive and V is non-expansive, then T is averaged.*
(d) *The composite of finitely many averaged operators is averaged. That is, if for each $i = 1, \ldots, N$, the operator T_i is averaged, then so is the composite operator $T_1 \ldots T_N$. In particular, if T_1 is α_1-averaged and T_2 is α_2-averaged, where $\alpha_1, \alpha_2 \in (0,1)$, then the composite $T_1 T_2$ is α-averaged, where $\alpha = \alpha_1 + \alpha_2 - \alpha_1\alpha_2$.*

Finally, for the last proposition of this section, we recall the definition of monotonicity of nonlinear operators.

Definition 2.6: Given a nonlinear operator A with domain $D(A)$ in H and β, ν are positive constants. Then A is called

(a) *monotone* on $D(A)$ if $\langle Ax - Ay, x - y\rangle \geq 0$ for all $x, y \in D(A)$;
(b) β*-strongly monotone* if $\langle Ax - Ay, x - y\rangle \geq \beta\|x - y\|^2$ for all $x, y \in D(A)$;
(c) ν*-inverse strongly monotone* (ν-ism, for short) if $\langle Ax - Ay, x - y\rangle \geq \nu\|Ax - Ay\|^2$ for all $x, y \in D(A)$.

The following proposition gathers some useful results on the relationship between averaged operators and inverse strongly monotone operators.

Proposition 2.7 ([10]): *If $T : H \to H$ is an operator, then the following statements hold:*

(a) *T is non-expansive if and only if the complement $I-T$ is $\frac{1}{2}$-ism;*
(b) *If T is ν-ism, then for $\gamma > 0$, γT is ν/γ-ism;*
(c) *T is averaged if and only if the complement $I-T$ is ν-ism for some $\nu > 1/2$. Indeed, for $\alpha \in (0, 1)$, T is α-averaged if and only if $I-T$ is $1/2\alpha$-ism.*

3. The algorithm and convergence analysis

This section is dedicated to the algorithmic formulation and analysis of the convex bilevel optimization problem (1)–(2). To this end, we first propose an inertial version of the viscosity approximation method for approximating fixed points of a non-expansive operator in an infinite dimensional real Hilbert space, a result which has its own interest. More precisely, let $T : H \to H$ be a non-expansive operator. We are interested in finding a fixed point of T, i.e. $x^* \in H$ such that $x^* = Tx^*$. Combining inertial extrapolation with viscosity techniques, we introduce a new iterative scheme that converges strongly to a fixed point of T in Algorithm 1.

Assumption 3.1: Suppose $\{\alpha_n\}_{n=1}^{\infty}$ is a sequence in (0,1) and $\{\epsilon_n\}_{n=1}^{\infty}$ is a positive sequence satisfying the following conditions:

(a) $\lim_{n\to\infty} \alpha_n = 0$ and $\sum_{n=1}^{\infty} \alpha_n = \infty$.
(b) $\epsilon_n = o(\alpha_n)$, i.e. $\lim_{n\to\infty}(\epsilon_n/\alpha_n) = 0$ (e.g. $\epsilon_n = 1/(n+1)^2, \alpha_n = 1/(n+1)$).

Remark 3.2: Observe that from Assumption 3.1 and Algorithm 1 we have that

$$\lim_{n\to\infty} \theta_n\|x_n - x_{n-1}\| = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{\theta_n}{\alpha_n}\|x_n - x_{n-1}\| = 0.$$

Algorithm 1 Inertial Viscosity Method

Step 0: Choose sequences $\{\alpha_n\}_{n=1}^{\infty}$ and $\{\epsilon_n\}_{n=1}^{\infty}$ such that the conditions in Assumption 3.1 hold. Select arbitrary points $x_0, x_1 \in H$, $\beta \in (0,1)$ and $\alpha \geq 3$. Set $n := 1$.

Step 1: Given the iterates x_{n-1} and x_n (with $n \geq 1$), choose θ_n such that we have $0 \leq \theta_n \leq \bar{\theta}_n$ with $\bar{\theta}_n$ defined by

$$\bar{\theta}_n := \begin{cases} \min\left\{\frac{n-1}{n+\alpha-1}, \ \frac{\epsilon_n}{\|x_n - x_{n-1}\|}\right\} & \text{if } \ x_n \neq x_{n-1}, \\ \frac{n-1}{n+\alpha-1} & \text{otherwise.} \end{cases} \tag{11}$$

Step 2: Proceed with the following computations

$$\begin{cases} y_n = x_n + \theta_n(x_n - x_{n-1}), \\ x_{n+1} = \alpha_n S(y_n) + (1-\alpha_n)(1-\beta)y_n + \beta(1-\alpha_n)Ty_n, \ \ n \geq 1, \end{cases} \tag{12}$$

Also note that Step 1 in our Algorithm 1 is easily implemented in numerical computation since the value of $\|x_n - x_{n-1}\|$ is a priori known before choosing θ_n.

Theorem 3.3: *Let H be a real Hilbert space and $T : H \to H$ a non-expansive operator such that $F(T) \neq \emptyset$. Let S be a contraction mapping with constant $\eta \in [0,1)$. Let Assumption* 3.1 *hold. Then the sequence $\{x_n\}$ generated by Algorithm* 1 *converges strongly z, which is the unique solution of the variational inequality*

$$\langle (I-S)z, x - z\rangle \geq 0, \quad \forall x \in F(T). \tag{13}$$

Proof: From (12), for any $z \in F(T)$, we have

$$\begin{aligned} \|x_{n+1} - z\| &\leq \alpha_n\|S(y_n) - z\| + (1-\alpha_n)(1-\beta)\|y_n - z\| + \beta(1-\alpha_n)\|Ty_n - z\| \\ &\leq \alpha_n\left(\|S(y_n) - S(z)\| + \|S(z) - z\|\right) + (1-\alpha_n)\|y_n - z\| \\ &\leq \alpha_n\|S(z) - z\| + (1 - \alpha_n(1-\eta))\|y_n - z\| \\ &\leq \alpha_n\|S(z) - z\| + (1 - \alpha_n(1-\eta))(\|x_n - z\| + \theta_n\|x_n - x_{n-1}\|) \\ &= (1 - \alpha_n(1-\eta))\|x_n - z\| + (1 - \alpha_n(1-\eta))\theta_n\|x_n - x_{n-1}\| \\ &\quad + \alpha_n\|S(z) - z\| \\ &= (1 - \alpha_n(1-\eta))\|x_n - z\| + \alpha_n\left((1 - \alpha_n(1-\eta))\frac{\theta_n}{\alpha_n}\|x_n - x_{n-1}\|\right. \\ &\quad \left. + \|S(z) - z\|\right). \end{aligned} \tag{14}$$

Observe that $\sup_{n\geq 1}(1 - \alpha_n(1-\eta))(\theta_n/\alpha_n)\|x_n - x_{n-1}\|$ exists by Remark 3.2 and take

$$M := \max\left\{\|S(z) - z\|, \ \sup_{n\geq 1}(1 - \alpha_n(1-\eta))\frac{\theta_n}{\alpha_n}\|x_n - x_{n-1}\|\right\}.$$

Then (14) becomes

$$\|x_{n+1} - z\| \leq (1 - \alpha_n(1 - \eta))\|x_n - z\| + \alpha_n M.$$

By Lemma 2.2, we get that $\{x_n\}$ is bounded. As a consequence, $\{y_n\}$ is also bounded.

Observing that

$$\|y_n - z\|^2 = \|x_n - z\|^2 + 2\theta_n\langle x_n - x_{n-1}, x_n - z\rangle + \theta_n^2\|x_n - x_{n-1}\|^2. \tag{15}$$

From Lemma 2.1 (i) it holds

$$2\langle x_n - x_{n-1}, x_n - z\rangle = -\|x_{n-1} - z\|^2 + \|x_n - z\|^2 + \|x_n - x_{n-1}\|^2. \tag{16}$$

Substituting (16) into (15), we obtain

$$\begin{aligned}
\|y_n - z\|^2 &= \|x_n - z\|^2 + \theta_n(-\|x_{n-1} - z\|^2 + \|x_n - z\|^2 + \|x_n - x_{n-1}\|^2) \\
&\quad + \theta_n^2\|x_n - x_{n-1}\|^2 \\
&= \|x_n - z\|^2 + \theta_n(\|x_n - z\|^2 - \|x_{n-1} - z\|^2) \\
&\quad + \theta_n(1 + \theta_n)\|x_n - x_{n-1}\|^2 \\
&\leq \|x_n - z\|^2 + \theta_n(\|x_n - z\|^2 - \|x_{n-1} - z\|^2) \\
&\quad + 2\theta_n\|x_n - x_{n-1}\|^2,
\end{aligned} \tag{17}$$

where the last inequality follows from the fact that $\theta_n \in [0, 1)$. Using Lemma 2.1 (ii) and (iii) and the fact that T is non-expansive, we obtain from (12) that

$$\begin{aligned}
\|x_{n+1} - z\|^2 &= \|\alpha_n(S(y_n) - z) + (1 - \alpha_n)(1 - \beta)(y_n - z) + \beta(1 - \alpha_n)(Ty_n - z)\|^2 \\
&\leq \|(1 - \alpha_n)(1 - \beta)(y_n - z) + \beta(1 - \alpha_n)(Ty_n - z)\|^2 \\
&\quad + 2\langle\alpha_n(S(y_n) - z), x_{n+1} - z\rangle \\
&= (1 - \alpha_n)^2(1 - \beta)\|y_n - z\|^2 + \beta(1 - \alpha_n)^2\|Ty_n - z\|^2 \\
&\quad - \beta(1 - \beta)(1 - \alpha_n)^2\|y_n - Ty_n\|^2 + 2\alpha_n\langle S(y_n) - z, x_{n+1} - z\rangle \\
&\leq (1 - \alpha_n)^2(1 - \beta)\|y_n - z\|^2 + \beta(1 - \alpha_n)^2\|y_n - z\|^2 \\
&\quad - \beta(1 - \beta)(1 - \alpha_n)^2\|y_n - Ty_n\|^2 + 2\alpha_n\langle S(y_n) - z, x_{n+1} - z\rangle \\
&= (1 - \alpha_n)^2\|y_n - z\|^2 - \beta(1 - \beta)(1 - \alpha_n)^2\|y_n - Ty_n\|^2 \\
&\quad + 2\alpha_n\langle S(y_n) - z, x_{n+1} - z\rangle.
\end{aligned} \tag{18}$$

Combining (17) and (18), we get

$$\begin{aligned}
\|x_{n+1} - z\|^2 \leq{}& (1 - \alpha_n)^2\|x_n - z\|^2 - \beta(1 - \beta)(1 - \alpha_n)^2\|y_n - Ty_n\|^2 \\
&+ \theta_n(1 - \alpha_n)^2(\|x_n - z\|^2 - \|x_{n-1} - z\|^2) \\
&+ 2\theta_n(1 - \alpha_n)^2\|x_n - x_{n-1}\|^2 \\
&+ 2\alpha_n\langle S(y_n) - z, x_{n+1} - z\rangle.
\end{aligned} \tag{19}$$

Setting $\Gamma_n := \|x_n - z\|^2$ for all $n \geq 1$, it follows from (19) that

$$\begin{aligned}\Gamma_{n+1} \leq (1-\alpha_n)^2\Gamma_n - \beta(1-\beta)(1-\alpha_n)^2\|y_n - Ty_n\|^2 + \theta_n(1-\alpha_n)^2(\Gamma_n - \Gamma_{n-1}) \\ + 2\theta_n(1-\alpha_n)^2\|x_n - x_{n-1}\|^2 + 2\alpha_n\langle S(y_n) - z, x_{n+1} - z\rangle. \end{aligned} \tag{20}$$

We consider two cases for the rest of the proof.

Case 1: Suppose there exists a natural number n_0 such that $\Gamma_{n+1} \leq \Gamma_n$ for all $n \geq n_0$. Therefore, $\lim_{n\to\infty}\Gamma_n$ exists. From (20), we have

$$\begin{aligned}&\beta(1-\beta)(1-\alpha_n)^2\|y_n - Ty_n\|^2 \\ &\quad\leq (\Gamma_n - \Gamma_{n+1}) + \theta_n(1-\alpha_n)^2(\Gamma_n - \Gamma_{n-1}) \\ &\qquad + 2\theta_n(1-\alpha_n)^2\|x_n - x_{n-1}\|^2 + 2\alpha_n\langle S(y_n) - z, x_{n+1} - z\rangle. \end{aligned} \tag{21}$$

Using Assumption 3.1 and Remark 3.2 (noting that $\lim_{n\to\infty}\theta_n\|x_n - x_{n-1}\| = 0$ and $\{x_n\}$, $\{y_n\}$ are bounded), we have

$$\lim_{n\to\infty}\beta(1-\beta)(1-\alpha_n)^2\|y_n - Ty_n\| = 0.$$

Observe that $\liminf_{n\to\infty}\beta(1-\beta)(1-\alpha_n)^2 = \lim_{n\to\infty}\beta(1-\beta)(1-\alpha_n)^2 = \beta(1-\beta) > 0$ and this immediately implies that

$$\lim_{n\to\infty}\|Ty_n - y_n\| = 0.$$

Since $\{x_n\}$ is bounded, take a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that $x_{n_k} \rightharpoonup p \in H$ and

$$\begin{aligned}\limsup_{n\to\infty}\langle S(z) - z, x_n - z\rangle &= \lim_{k\to\infty}\langle S(z) - z, x_{n_k} - z\rangle \\ &= \langle S(z) - z, p - z\rangle. \end{aligned} \tag{22}$$

From $y_n = x_n + \theta_n(x_n - x_{n-1})$, we get

$$\|y_n - x_n\| = \theta_n\|x_n - x_{n-1}\| \to 0.$$

Since $x_{n_k} \rightharpoonup p$, we have $y_{n_k} \rightharpoonup p$. Lemma 2.4 then guarantees that $p \in F(T)$. Furthermore, we have from (13) and (24) that

$$\limsup_{n\to\infty}\langle S(z) - z, x_n - z\rangle \leq 0. \tag{23}$$

From the contraction of S and (17), we can write

$$\begin{aligned}2\alpha_n\langle S(y_n) - z, x_{n+1} - z\rangle &= 2\alpha_n\langle S(y_n) - S(z) + S(z) - z, x_{n+1} - z\rangle \\ &\leq 2\alpha_n\eta\|y_n - z\|\|x_{n+1} - z\| + 2\alpha_n\langle S(z) - z, x_{n+1} - z\rangle \\ &\leq \alpha_n\eta(\|y_n - z\|^2 + \|x_{n+1} - z\|^2) + 2\alpha_n\langle S(z) - z, x_{n+1} - z\rangle \\ &\leq \alpha_n\eta(\Gamma_n + \theta_n(\Gamma_n - \Gamma_{n-1}) + 2\theta_n\|x_n - x_{n-1}\|^2) \\ &\quad + 2\alpha_n\langle S(z) - z, x_{n+1} - z\rangle + \alpha_n\eta\|x_{n+1} - z\|^2. \end{aligned}$$

Therefore from (20) it holds

$$\begin{aligned}
\Gamma_{n+1} &\leq (1-\alpha_n)^2\Gamma_n + \theta_n(1-\alpha_n)^2(\Gamma_n - \Gamma_{n-1}) \\
&\quad + 2\theta_n(1-\alpha_n)^2\|x_n - x_{n-1}\|^2 + 2\alpha_n\langle S(y_n) - z, x_{n+1} - z\rangle \\
&\leq ((1-\alpha_n)^2 + \alpha_n\eta)\Gamma_n + \theta_n((1-\alpha_n)^2 + \alpha_n\eta)(\Gamma_n - \Gamma_{n-1}) \\
&\quad + 2\theta_n((1-\alpha_n)^2 + \alpha_n\eta)\|x_n - x_{n-1}\|^2 + 2\alpha_n\langle S(z) - z, x_{n+1} - z\rangle \\
&\quad + \alpha_n\eta\|x_{n+1} - z\|^2 \\
&\leq ((1-\alpha_n)^2 + \alpha_n\eta)\Gamma_n + \theta_n((1-\alpha_n)^2 + \alpha_n\eta)\|x_n - x_{n-1}\|(\sqrt{\Gamma_n} + \sqrt{\Gamma_{n-1}}) \\
&\quad + 2\theta_n((1-\alpha_n)^2 + \alpha_n\eta)\|x_n - x_{n-1}\|^2 + 2\alpha_n\langle S(z) - z, x_{n+1} - z\rangle \\
&\quad + \alpha_n\eta\|x_{n+1} - z\|^2 \\
&= ((1-\alpha_n)^2 + \alpha_n\eta)\Gamma_n + \theta_n\|x_n - x_{n-1}\|M_2 + \alpha_n\eta\|x_{n+1} - z\|^2 \\
&\quad + 2\alpha_n\langle S(z) - z, x_{n+1} - z\rangle, \qquad (24)
\end{aligned}$$

where

$$M_2 := \sup_{n\geq 1}\left((1-\alpha_n)^2 + \alpha_n\eta)(\sqrt{\Gamma_n} + \sqrt{\Gamma_{n-1}}) + 2((1-\alpha_n)^2 + \alpha_n\eta)\|x_n - x_{n-1}\|\right).$$

Therefore

$$\begin{aligned}
\Gamma_{n+1} &\leq \frac{(1-\alpha_n)^2 + \alpha_n\eta}{1-\alpha_n\eta}\Gamma_n + \frac{\theta_n\|x_n - x_{n-1}\|M_2}{1-\alpha_n\eta} \\
&\quad + 2\frac{\alpha_n}{1-\alpha_n\eta}\langle S(z) - z, x_{n+1} - z\rangle \\
&= \left(1 - \frac{2(1-\eta)\alpha_n}{1-\alpha_n\eta}\right)\Gamma_n + \frac{\theta_n\|x_n - x_{n-1}\|M_2}{1-\alpha_n\eta} \\
&\quad + 2\frac{\alpha_n}{1-\alpha_n\eta}\langle S(z) - z, x_{n+1} - z\rangle + \frac{\alpha_n^2}{1-\alpha_n\eta}\Gamma_n \\
&\leq \left(1 - \frac{2(1-\eta)\alpha_n}{1-\alpha_n\eta}\right)\Gamma_n \\
&\quad + \frac{2(1-\eta)\alpha_n}{1-\alpha_n\eta}\left\{\frac{\theta_n\|x_n - x_{n-1}\|M_2}{2(1-\eta)\alpha_n} + \frac{\alpha_n\Gamma_{n_0}}{2(1-\eta)} + \frac{1}{1-\eta}\langle S(z) - z, x_{n+1} - z\rangle\right\} \\
&= (1-\delta_n)\Gamma_n + \delta_n\sigma_n, \quad \forall n \geq n_0, \qquad (25)
\end{aligned}$$

where $\delta_n := 2(1-\eta)\alpha_n/1-\alpha_n\eta$ and

$$\sigma_n := \frac{\theta_n\|x_n - x_{n-1}\|M_2}{2(1-\eta)\alpha_n} + \frac{\alpha_n\Gamma_{n_0}}{2(1-\eta)} + \frac{1}{1-\eta}\langle S(z) - z, x_{n+1} - z\rangle.$$

Let us verify conditions in Lemma 2.2 (ii). Since

$$\sum \delta_n = \sum \frac{2(1-\eta)\alpha_n}{1-\alpha_n\eta} \geq \sum 2(1-\eta)\alpha_n = \infty,$$

we have $\sum \delta_n = \infty$. Furthermore, using Remark 3.2, Assumption 3.1 (a) and (23) we obtain $\limsup \sigma_n = 0$. Now we can apply Lemma 2.2 (ii) and Assumption 3.1 in (25) to get $\Gamma_n = \|x_n - z\| \to 0$ and thus $x_n \to z$ as $n \to \infty$.

Case 2: Assume that there is no $n_0 \in \mathbb{N}$ such that $\{\Gamma_n\}_{n=n_0}^{\infty}$ is monotonically decreasing. Let $\tau : \mathbb{N} \to \mathbb{N}$ be a mapping defined for all $n \geq n_0$ (for some n_0 large enough) by

$$\tau(n) := \max \left\{k \in \mathbb{N} : k \leq n, \Gamma_k \leq \Gamma_{k+1}\right\},$$

i.e. $\tau(n)$ is the largest number k in $\{1, \ldots, n\}$ such that Γ_k increases at $k = \tau(n)$; note that, in view of Case 2, this $\tau(n)$ is well-defined for all sufficiently large n. Clearly, τ is a non-decreasing sequence [21] such that $\tau(n) \to \infty$ as $n \to \infty$ and

$$0 \leq \Gamma_{\tau(n)} \leq \Gamma_{\tau(n)+1}, \quad \forall n \geq n_0.$$

Using similar techniques as in (21), it is easy to show that

$$\lim_{n\to\infty} \|Ty_{\tau(n)} - y_{\tau(n)}\| = \lim_{n\to\infty} \|y_{\tau(n)} - x_{\tau(n)}\| = \lim_{n\to\infty} \|Ty_{\tau(n)} - x_{\tau(n)}\| = 0.$$

Furthermore, using the boundedness of $\{x_n\}$, $\{y_n\}$ and Assumption 3.1, we get

$$\begin{aligned}\|x_{\tau(n)+1} - x_{\tau(n)}\| \leq \alpha_{\tau(n)}\|S\Big(y_{\tau(n)}\Big) - x_{\tau(n)}\| + (1 - \alpha_{\tau(n)})\|y_{\tau(n)} - x_{\tau(n)}\| \\ + (1 - \alpha_{\tau(n)})\|Ty_{\tau(n)} - x_{\tau(n)}\| \longrightarrow 0 \text{ as } n \to \infty. \end{aligned} \tag{26}$$

Since $\{x_{\tau(n)}\}$ is bounded, there exists a subsequence of $\{x_{\tau(n)}\}$, still denoted by $\{x_{\tau(n)}\}$, which converges weakly to some $p \in F(T)$. Similarly, as in Case 1 above, we can show that we have $\limsup_{n\to\infty} \langle S(z) - z, x_{\tau(n)+1} - z\rangle \leq 0$. Following (25), we obtain

$$\Gamma_{\tau(n)+1} = (1 - \delta_{\tau(n)})\Gamma_{\tau(n)} + \delta_{\tau(n)}\sigma_{\tau(n)}, \tag{27}$$

which implies that $\|x_{\tau(n)} - z\|^2 \leq \sigma_{\tau(n)}$ while noting that $\Gamma_{\tau(n)} \leq \Gamma_{\tau(n)+1}$ and $\alpha_{\tau(n)} > 0$ hold. This leads to $\limsup_{n\to\infty} \|x_{\tau(n)} - z\|^2 \leq 0$. Thus, we have

$$\lim_{n\to\infty} \|x_{\tau(n)} - z\| = \lim_{n\to\infty} \Gamma_{\tau(n)} = 0,$$

which in turn implies $\lim_{n\to\infty} \|x_{\tau(n)+1} - z\| = 0$. Furthermore, for $n \geq n_0$, it is easy to see that $\Gamma_n \leq \Gamma_{\tau(n)+1}$ (observe that $\tau(n) \leq n$ for $n \geq n_0$ and consider the three cases: $\tau(n) = n$, $\tau(n) = n - 1$ and $\tau(n) < n - 1$. For the first and second cases, it is obvious that $\Gamma_n \leq \Gamma_{\tau(n)+1}$ for $n \geq n_0$. For the third case $\tau(n) \leq n - 2$, we have from the definition of $\tau(n)$ and for any integer $n \geq n_0$ that $\Gamma_j \geq \Gamma_{j+1}$ for $\tau(n) + 1 \leq j \leq n - 1$. Thus, $\Gamma_{\tau(n)+1} \geq \Gamma_{\tau(n)+2} \geq \cdots \geq \Gamma_{n-1} \geq \Gamma_n$). As a consequence, we obtain for all sufficiently large n that $0 \leq \Gamma_n \leq \Gamma_{\tau(n)+1}$. Hence $\lim_{n\to\infty} \Gamma_n = 0$. Therefore, $\{x_n\}$ converges to z. ■

We are now in position to state the *inertial Bilevel Gradient Sequential Averaging Method* (iBiG-SAM) for bilevel optimization problem (1)–(2) and present its convergence

analysis. We first state the assumptions that will be needed throughout the rest of this paper.

Assumption 3.4: Considering problem (1)–(2), let the following hold:

(a) $f : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable such that its gradient is Lipschitz continuous with constant L_f.
(b) $g : \mathbb{R}^n \to (-\infty, \infty]$ is proper, lower semicontinous and convex.
(c) $h : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with parameter $\sigma > 0$ and continuously differentiable such that its gradient is Lipschitz continuous with constant L_h.
(d) The set X^* of all optimal solutions of problem (2) is non-empty.

Assumption 3.5: Suppose $\{\alpha_n\}_{n=1}^{\infty}$ is a sequence in (0,1) and $\{\epsilon_n\}_{n=1}^{\infty}$ is a positive sequence satisfying the following conditions:

(a) $\lim_{n\to\infty} \alpha_n = 0$ and $\sum_{n=1}^{\infty} \alpha_n = \infty$.
(b) $\epsilon_n = o(\alpha_n)$, i.e. $\lim_{n\to\infty}(\epsilon_n/\alpha_n) = 0$.
(c) $\lambda \in (0, (2/L_f))$ and $\gamma \in (0, (2/L_h + \sigma)]$.

Remark 3.6: Note that the stepsize λ in Assumption (c) above is chosen in a larger interval than that of [28]. Also, our Assumption (a) is weaker than Assumption C of [28] since $\{\alpha_n\}_{n=1}^{\infty}$ is not required in our Assumption (a) to satisfy $\lim_{n\to\infty}(\alpha_{n+1}/\alpha_n) = 1$ as assumed in Assumption C of [28]. Take, for example, $\alpha_n = 1/\sqrt{n}$, when n is odd and $\alpha_n = 1/n$, when n is even. We see that $\{\alpha_n\}$ satisfies Assumption (a) but $\alpha_{n+1}/\alpha_n \not\to 1$.

We next give a precise statement of our *inertial Bilevel Gradient Sequential Averaging Method* (iBiG-SAM) as follows.

Algorithm 2 iBiG-SAM

Step 0: Choose sequences $\{\alpha_n\}_{n=1}^{\infty}$ and $\{\epsilon_n\}_{n=1}^{\infty}$ such that the conditions in Assumption 3.1 hold. Select arbitrary points $x_0, x_1 \in \mathbb{R}^n$ and $\alpha \geq 3$. Set $n := 1$.

Step 1: Given the iterates x_{n-1} and x_n (with $n \geq 1$), choose θ_n such that we have $0 \leq \theta_n \leq \bar{\theta}_n$ with $\bar{\theta}_n$ defined by

$$\bar{\theta}_n := \begin{cases} \min\left\{\frac{n-1}{n+\alpha-1}, \ \frac{\epsilon_n}{\|x_n - x_{n-1}\|}\right\} & \text{if } \ x_n \neq x_{n-1}, \\ \frac{n-1}{n+\alpha-1} & \text{otherwise.} \end{cases} \tag{28}$$

Step 2: Proceed with the following computations

$$\begin{cases} y_n = x_n + \theta_n(x_n - x_{n-1}), \\ s_n = \text{prox}_{\lambda g}(y_n - \lambda \nabla f(y_n)), \\ z_n = y_n - \gamma \nabla h(y_n), \\ x_{n+1} = \alpha_n z_n + (1 - \alpha_n)s_n, \ \ n \geq 1. \end{cases} \tag{29}$$

We are now in the position to discuss the convergence of iBIG-SAM. Let us define

$$T_\lambda := \text{prox}_{\lambda g}(I - \lambda \nabla f). \tag{30}$$

The next lemma shows that the prox-grad mapping T_λ is averaged. This is an improvement over Lemma 1(i) of [28].

Lemma 3.7: *The prox-grad mapping T_λ (30) is $(2 + \lambda L_f)/4$-averaged for all $\lambda \in (0, 2/L_f)$.*

Proof: Observe that the Lipschitz condition on ∇f implies that ∇f is $1/L_f$-ism (see [3]), which then implies that $\lambda \nabla f$ is $1/\lambda L_f$-ism. Hence, by Proposition 2.7(c), $I - \lambda \nabla f$ is $(\lambda L_f/2)$-averaged. Since $\text{prox}_{\lambda f}$ is firmly non-expansive and hence 1/2-averaged, we see from Proposition 2.5(d) that the composite $\text{prox}_{\lambda g}(I - \lambda \nabla f)$ is $(2 + \lambda L_f)/4$-averaged for $\lambda \in (0, 2/L_f)$. Hence we have that, $T_\lambda = \text{prox}_{\lambda g}(I - \lambda \nabla f)$ is $(2 + \lambda L_f)/4$-averaged. Therefore, we can write

$$T_\lambda = \text{prox}_{\lambda g}(I - \lambda \nabla f) = \left(\frac{2 - \lambda L_f}{4}\right) I + \left(\frac{2 + \lambda L_f}{4}\right) T \tag{31}$$

$$= (1 - \beta) I + \beta T, \tag{32}$$

where $\beta := \frac{2+\lambda L_f}{4} \in [a, b] \subset (1/2, 1)$ and T is a non-expansive mapping. ■

Lemma 1(ii) of [28] showed the equivalence between the fixed points of prox-grad mapping T_λ (30) and optimal solutions of problem (2). That is, $x \in X^*$ if and only if $x = T_\lambda x$. This equivalence will be needed in our convergence analysis in this paper.

Lemma 3.8 ([28]): *Suppose that Assumption* 3.4 *(c) holds. Then, the mapping S_γ, defined by $S_\gamma := I - \gamma \nabla h$, is a contraction for all $\gamma \in (0, 2/(L_h + \sigma))$. That is,*

$$\|S_\gamma(x) - S_\gamma(y)\| \leq \eta \|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

Here, I represents the identity operator and $\eta := \sqrt{1 - 2\gamma\sigma L_h/(\sigma + L_h)}$.

By the statements of Lemma 3.7 and Lemma 3.8, we can re-write (29) as

$$\begin{aligned} y_n &= x_n + \theta_n(x_n - x_{n-1}), \\ x_{n+1} &= \alpha_n S_\gamma(y_n) + (1 - \alpha_n)(1 - \beta) y_n + \beta(1 - \alpha_n) T y_n, \ n \geq 1, \end{aligned} \tag{33}$$

where T is a non-expansive mapping, S_γ is a contraction mapping and $\beta := (2 + \lambda L_f)/4$.

Before we proceed with the main result of this section, we first show that the iterative sequence generated by our algorithm is bounded.

Lemma 3.9: *Let Assumptions* 3.4 *and* 3.5 *be satisfied. Then the sequence $\{x_n\}$ generated by Algorithm* 1 *is bounded.*

Proof: First observe that $x \in X^*$ if and only if $x \in F(T_\lambda) = F(T)$. Using the same arguments in the earlier part of the proof of Theorem 3.3 replacing S in Theorem 3.3 with S_γ, we obtain the desired result that $\{x_n\}$ generated by Algorithm 2 is bounded. ■

Theorem 3.10: *Let Assumptions* 3.4 *and* 3.5 *hold. Then the sequence* $\{x_n\}$ *generated by Algorithm* 2 *converges to a point* $z \in X^*$ *satisfying*

$$\langle \nabla h(z), x - z \rangle \geq 0 \quad \forall x \in X^* \tag{34}$$

and therefore, $z = z_{mn}$, *where* z_{mn} *is the unique optimal solution of problem (1).*

Proof: Follows from Theorem 3.3, while replacing S in Theorem 3.3 with S_γ in Lemma 3.8. ■

Remark 3.11: Suppose that Assumption 3.4(c) is replaced with the following milder condition: '$h : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with parameter $\sigma > 0$ and L_h-Lipschitz continuous'. Then the step involving z_n in Algorithm 2 can be replaced by

$$\begin{aligned} z_n &= y_n - \gamma \nabla M_{\gamma h}(y_n) \\ &= y_n - \gamma \frac{1}{\gamma}(y_n - \text{prox}_{\gamma h}(y_n)) \\ &= \text{prox}_{\gamma h}(y_n), \end{aligned}$$

where $M_{\gamma h}$ is the Moreau envelop of h, defined by

$$M_{\gamma h}(x) := \min_{u \in \mathbb{R}^n} \left\{ h(u) + \frac{1}{2\gamma} \|u - x\|^2 \right\},$$

which is continuously differentiable (see [5]) with $\nabla M_{\gamma h}(x) = (1/\gamma)(x - \text{prox}_{\gamma h}(x))$ and global convergence is still obtained as in Theorem 3.10 using Lemma 6 of [28].

We give some brief comments on the non-asymptotic $O(1/n^2)$ convergence rate of some estimates obtained in Theorem 3.10.

Remark 3.12: Observe that for Algorithm 2, we have $\theta_n \|x_n - x_{n-1}\| \leq \epsilon_n$ for all $n \geq 1$. If we choose $\epsilon_n := c/n^2$, where $c > 0$, then $\theta_n \|x_n - x_{n-1}\| \leq c/n^2$ for all $n \geq 1$. Thus, $\theta_n \|x_n - x_{n-1}\| = O(1/n^2)$ and consequently

$$\|y_n - x_n\| = \theta_n \|x_n - x_{n-1}\| = O(1/n^2).$$

Full details on the convergence rate of the result in Theorem 3.10 is left for further careful investigation in a separate work.

4. Numerical results

For numerical implementation of our proposed method in Section 3 we consider the inverse problems tested in [28] and give numerical comparison with the proposed Algorithm 2 (iBiG-SAM) and that of BiG-SAM method in [28]. The codes are implemented in Matlab. We perform all computations on a windows desktop with an Intel(R) Core(TM) i7-2600 CPU at 3.4 GHz and 8.00 GB of memory. We take $\alpha_n = 2\kappa/n(1-\beta)$ with $\kappa = 0.1$, which is the best choice for BiG-SAM considered in [28] and $\beta \in [0, 1)$ defined as in (32) and $\theta_n = \bar{\theta}_n$ as in (28) with $\alpha = 3$ and $\epsilon_n = \alpha_n / n^{0.01}$ for iBiG-SAM.

Example 4.1: Following [28], the inner objective function is taking as

$$\varphi(x) := \frac{1}{2}\|Ax - b\|^2 + \delta_X(x),$$

where δ_X is the indicator function over the non-negative orthant $X := \{x \in \mathbb{R}^n : x \geq 0\}$. Furthermore, we take the outer objective function as

$$h(x) := \frac{1}{2}x^T Qx, \tag{35}$$

where Q is a positive definite matrix. It is clear that $L_f = \|A^t A\|$ and $L_h = \|Q\|$. We choose $\lambda = 1/L_f$ and $\gamma = 2/(L_h + \sigma)$.

Following [6], we consider three inverse problems, i.e. *Baart*, *Foxgood*, and *Phillips* [28]. These test problems are well known as important Fredholm integral equations of the first kind of the following form

$$\int_a^b K(s,t)f(t)\,\mathrm{d}t = g(s) \quad c \leq s \leq d,$$

where the right-hand side g and the kernel K are given and where f is the unknown solution. In *Baart* problem [2], $g(s) := 2sins/s$ and the kernel $K(s,t) := exp(scost)$ and integration intervals $s \in [0, \pi/2]$ and $t \in [0, \pi]$. *Foxgood* is a severely ill-posed test problem with both integration intervals equal to $[0, 1]$, with kernel K and right-hand side g given by

$$K(s,t) = (s^2 + t^2)^{1/2}, \quad g(s) = \frac{1}{2}\left((1 + s^2)^{3/2} - s^3\right),$$

and with the solution $f = t$ [4, p. 665]. *Phillips* is the 'famous' Fredholm integral equation of the first kind studied by D. L. Phillips [25]. Define the function

$$\phi(x) = \begin{array}{ll} 1 + cos\left(\dfrac{\pi x}{3}\right) & |x| < 3 \\ 0 & |x| \geq 3. \end{array}$$

Then the kernel K, the solution f, and the right-hand side g are given by:

$$\begin{aligned} K(s,t) &= \phi(s - t) \\ f(t) &= \phi(t) \\ g(s) &= (6 - |s|)\left(1 + \frac{1}{2}\cos\left(\frac{\pi s}{3}\right)\right) + \frac{9}{2\pi}\sin\left(\frac{\pi |s|}{3}\right), \end{aligned}$$

with both integration intervals are $[-6, 6]$.

In our experiment, the data A,b,x are the results from the discretization of the test problems *Baart*,*Foxgood*, and *Phillips*. For each of these problems, we generated the corresponding 1000 by 1000 exact linear system of the form $Ax = b$, by applying the relevant function (baart, foxgood, and phillips) from the *regularization tools* (see http://www.imm.dtu.dk/pcha/Regutools/). We then performed the simulation by adding normally distributed noise with zero mean to the right-hand-side vector b, with deviation

Table 1. Averaged over 100 runs for each problem in Example 4.1.

	iBiG-SAM		BiG-SAM	
Problem	Number of iterations	Time (s)	Number of iterations	Time (s)
Baart	119.15	1.7253	145.67	2.1089
Foxgood	122.04	1.7861	149.78	2.1885
Phillips	120.77	1.7463	148.18	2.1397

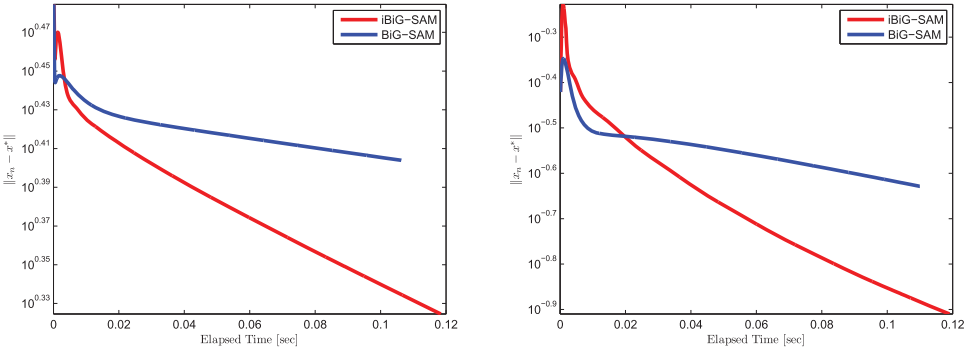

Figure 1. Distance to optimal solution vs. CPU time for problems Baart (left) and Foxgood (right) with $n = 100$.

$\rho = 0.01$. The matrix Q is defined by $Q = LL' + I$, where L is generated by the function get-l(1,000,1) from the *regularization tools* and approximates the first-derivative operator.

Following [28], we use the stopping condition $(\varphi(x_n) - \varphi^*)/\varphi^* \leq 10^{-2}$ for both methods, where φ^* is the optimal value of the inner problem computed in advance by BiG-SAM with 1000 iterations. In Table 1 we present the averaged number of iterations and time (out of 100 runs) until the algorithms reach the stopping criterion. It can be seen that iBiG-SAM outperforms BiG-SAM (on averaged about 20%) in all problems tested.

In Figure 1, we compare the behaviour of iBiG-SAM wih BiG-SAM for Baart and Foxgood problems when $n = 100$.

Example 4.2: We now look at the case when g is not an indicator function. In this case, the methods proposed in [6,18,29] cannot be applied. We still give a comparison of our method with BiG-SAM (5). The inner objective function is taking here as

$$\varphi(x) := \frac{1}{2}\|Ax - b\|_2^2 + \mu\|x\|_1,$$

where $A \in \mathbb{R}^{m\times n}$ is a given matrix, b is a given vector and μ a positive scalar. This is LASSO (Least Absolute selection and Shrinkage Operator) [30] in compressed sensing. The proximal map with $g(x) = \mu\|x\|_1$ is given as $\text{prox}_g(x) = \arg\min_u \mu\|x\|_1 + \frac{1}{2}\|u - x\|_2^2$, which is separable in indices. Thus, for $x \in \mathbb{R}^n$,

$$\text{prox}_g(x) = \left(\text{prox}_{\mu|.|_1}(x_1), \ldots, \text{prox}_{\mu|.|_1}(x_n)\right) = (\beta_1, \ldots, \beta_n),$$

where $\beta_k = \text{sgn}(x_k)\max\{|x_k| - \mu, 0\}$ for $k = 1, 2, \ldots, n$. As in Example 4.1, we take the outer objective function as in (35) with Q similarly being a positive definite matrix. We

Table 2. Averaged over 100 runs for each problem in Example 4.2.

	iBiG-SAM		BiG-SAM	
Parameters	Iterations	Time (s)	Iterations	Time (s)
$\alpha = 3, m = 100, n = 500$	43.32	0.0498	60.43	0.0697
$\alpha = 4, m = 200, n = 500$	12.25	0.017	18.65	0.0252
$\alpha = 5, m = 500, n = 1000$	12.31	0.124	18.07	0.1793

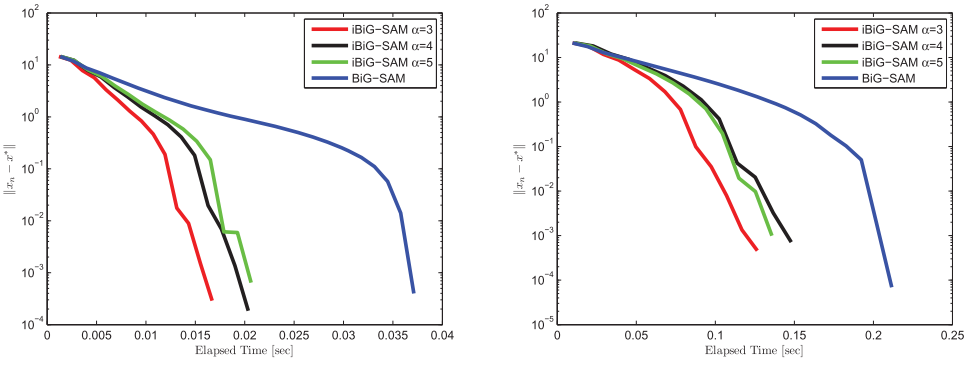

Figure 2. Distance to optimal solution vs. CPU time when $m = 100, n = 500$ (left) and $m = 500, n = 1000$ (right).

take $\mu = 0.5$, and the data b is generated as $Ax + \delta e$, where A and e are random matrices whose elements are normally distributed with zero mean and variance 1, and $\delta = 0.01$, and x is a generated sparse vector. The stopping condition is $\|x_n - x^*\| \leq \epsilon$ with $\epsilon = 10^{-3}$ and x^* computed in advance by BiG-SAM with 1000 iterations. In Table 2 we present the averaged number of iterations and time (out of 100 runs) until the algorithms reach the stopping criterion for different choices of $\alpha \geq 3$ in different dimensional spaces. Again iBiG-SAM outperforms BiG-SAM in all simulations.

In Figure 2, we compare the behaviour of BiG-SAM wih iBiG-SAM for different parameters α. It seems that iBiG-SAM with $\alpha = 3$ takes advantage over other values tested.

Interested readers can download the codes used for the experiments above via the following link (under *iBIG-SAM*), in order to proceed with their own tests on other scenarios of Examples 4.1 and 4.2 or to use corresponding adjustments for calculations on new examples: http://www.southampton.ac.uk/abz1e14/solvers.html

5. Concluding remarks

The paper has introduced and proved the global convergence of an inertial extrapolation-type method for solving simple convex bilevel optimization problems in finite dimensional Euclidean spaces. Note that our scheme is not restricted to solving simple convex bilevel optimization problems but can be applied to a general fixed point problem for non-expansive operators in infinite dimensional Hilbert spaces. Based on the numerical experiments conducted, we have shown that our method outperforms the known best algorithm recently proposed in [28] to solve problems of the form (1)–(2). Our next project in this subject area is to derive the convergence rate of the method proposed in this paper.

Acknowledgments

The research of Yekini Shehu is supported by the Alexander von Humbold-Foundation and Post-doctoral Fellowship Grant from the Institute of Science and Technology, Klosterneuburg, Vienna, Austria. The authors are indebted to two anonymous referees for their constructive remarks, which have helped us to improve the quality of the paper. We would also like to thank the Associate Editor for the efficient handling of the review process.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The work of Phan Tu Vuong is funded by the Austrian Science Fund (FWF) project number M2499-N32. The work of Alain Zemkoho is funded by the EPSRC Grant EP/P022553/1.

References

[1] H. Attouch, J. Peypouquet, and P. Redont, *A dynamical approach to an inertial forward-backward algorithm for convex minimization*, SIAM J. Optim. 24 (2014), pp. 232–256.
[2] M.L. Baart, *The use of auto-correlation for pseudo-rank determination in noisy ill-conditioned linear least-squares problems*, IMA J. Numer. Anal. 2 (1982), pp. 241–247.
[3] J.B. Baillon and G. Haddad, *Quelques proprietes des operateurs angle-bornes et n-cycliquement monotones*, Israel J. Math. 26 (1977), pp. 137–150.
[4] C.T.H. Baker, *The Numerical Treatment of Integral Equations*, Clarendon Press, Oxford, 1977.
[5] H.H. Bauschke and P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics, Springer, New York, 2011.
[6] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci. 2 (2009), pp. 183–202.
[7] A. Beck and S. Sabach, *A first order method for finding minimal norm-like solutions of convex optimization problems*, Math. Program. 147 (2014), pp. 25–46.
[8] R.I. Bot and E.R. Csetnek, *An inertial alternating direction method of multipliers*, Minimax Theory Appl. 1 (2016), pp. 29–49.
[9] R.I. Bot and D.-K. Nguyen, *A forward-backward penalty scheme with inertial effects for montone inclusions. Applications to convex bilevel programming*, 2018. Available at http://www.optimization-online.org/DB_HTML/2018/01/6416.html.
[10] C. Byrne, *Unified treatment of some algorithms in signal processing and image construction*, Inverse Probl. 20 (2004), pp. 103–120.
[11] A. Cabot, *Proximal point algorithm controlled by a slowly vanishing term: Applications to hierarchial minimization*, SIAM J. Optim. 15 (2005), pp. 555–572.
[12] S. Dempe, *Foundations of Bilevel Programming*, Kluwer Academic Publishers, New York, 2002.
[13] S. Dempe, N. Dinh, and J. Dutta, *Optimality conditions for a simple convex bilevel programming problem*, in *Variational Analysis and Generalized Differentiation in Optimization and Control*, R.S. Burachik and J.C. Yao, eds., Springer, New York, 2010, pp. 149–161.
[14] S. Dempe, N. Dinh, J. Dutta, and T. Pandit, *Simple bilevel programming and extensions: Theory and algorithms*, Preprint, Indian Institute of Technology, Kanpur, India, 2018.
[15] S. Dempe and A.B. Zemkoho, *The generalized Mangasarian-Fromowitz constraint qualification and optimality conditions for bilevel programs*, J. Optim. Theory Appl. 148 (2011), pp. 433–441.
[16] G. Eichfelder, *Multiobjective bilevel optimization*, Math. Program. 123 (2010), pp. 419–449.
[17] S. Franke, P. Mehlitz, and M. Pilecka, *Optimality conditions for the simple convex bilevel programming problem in Banach spaces*, Optimization 67 (2018), pp. 237–268.
[18] M.C. Ferris and O.L. Mangasarian, *Finite perturbation of convex programs*, Appl. Math. Optim. 23 (1991), pp. 263–273.

[19] J. Fülöp, *On the equivalency between a linear bilevel programming problem and linear optimization over the efficient set*, Tech. Rep. Hungarian Academy of Science, 1993.
[20] K. Goebel and W.A. Kirk, *Topics in Metric Fixed Point Theory*, Cambridge University Press, Cambridge, 1990.
[21] P.-E. Maingé, *A hybrid extragradient-viscosity method for monotone operators and fixed point problems*, SIAM J. Control Optim. 47 (2008), pp. 1499–1515.
[22] O.L. Mangasarian and R.R. Meyer, *Nonlinear perturbation of linear programs*, SIAM J. Control Optim. 17 (1979), pp. 745–752.
[23] Y. Malitsky, *Chambolle-Pock and Tseng's methods: Relationship and extension to the bilevel optimization*, 2017. arXiv:1706.02602.
[24] P. Ochs, T. Brox, and T. Pock, *iPiasco: Inertial proximal algorithm for strongly convex optimization*, J. Math. Imaging Vision 53 (2015), pp. 171–181.
[25] D.L. Phillips, *A technique for the numerical solution of certain integral equations of the first kind*, J. Assoc. Comput. Machinery 9 (1962), pp. 84–97.
[26] J. Philip, *Algorithms for the vector maximization problem*, Math. Program. 2 (1972), pp. 207–229.
[27] S. Ruuska, K. Miettinen, and M.M. Wiecek, *Connections between single-level and bilevel multiobjective optimization*, J. Optim. Theory Appl. 153 (2012), pp. 60–74.
[28] S. Sabach and S. Shtern, *A first order method for solving convex bilevel optimization problems*, SIAM J. Optim. 27 (2017), pp. 640–660.
[29] M.V. Solodov, *An explicit descent method for bilevel convex optimization*, J. Convex Anal. 4 (2007), pp. 227–237.
[30] R. Tibshirami, *Regression shrinkage and selection via lasso*, J. Roy. Statist. Soc. Ser. B 58 (1996), pp. 267–288.
[31] A.N. Tikhonov and V.Y. Arsenin, *Solutions of Ill-Posed Problems*, Wiley, New York, 1977.
[32] H.K. Xu, *Iterative algorithm for nonlinear operators*, J. London Math. Soc. 66(2) (2002), pp. 1–17.
[33] H.K. Xu, *Viscosity approximation methods for nonexpansive mappings*, J. Math. Anal. Appl. 298 (2004), pp. 279–291.
[34] Y. Yamamoto, *Optimization over the efficient set: overview*, J. Globe Optim. 22 (2002), pp. 285–317.
[35] J.J. Ye and D.L. Zhu, *Optimality conditions for bilevel programming problems*, Optimization 33 (1995), pp. 9–27; Erratum in *Optimization* 39 (1997) 361–366.