Article

A Novel Inertial Viscosity Algorithm for Bilevel Optimization Problems Applied to Classification Problems

Kobkoon Janngam [1], Suthep Suantai [2], Yeol Je Cho [3], Attapol Kaewkhao [2] and Rattanakorn Wattanataweekul [4,*]

[1] Graduate Ph.D. Degree Program in Mathematics, Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand; kobkoon_jan@cmu.ac.th
[2] Research Center in Optimization and Computational Intelligence for Big Data Prediction, Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand; suthep.s@cmu.ac.th (S.S.); akaewkhao@yahoo.com (A.K.)
[3] Department of Mathematics Education, Gyeogsang National University, Jinju 52828, Republic of Korea; yjchomath@gmail.com
[4] Department of Mathematics, Statistics and Computer, Faculty of Science, Ubon Ratchathani University, Ubon Ratchathani 34190, Thailand
[*] Correspondence: rattanakorn.w@ubu.ac.th

Abstract: Fixed-point theory plays many important roles in real-world problems, such as image processing, classification problem, etc. This paper introduces and analyzes a new, accelerated common-fixed-point algorithm using the viscosity approximation method and then employs it to solve convex bilevel optimization problems. The proposed method was applied to data classification with the Diabetes, Heart Disease UCI and Iris datasets. According to the data classification experiment results, the proposed algorithm outperformed the others in the literature.

Keywords: classification problems; convex bilevel optimization; forward–backward algorithm

MSC: 47H09; 90C25; 65K10

1. Introduction

Many real-world hierarchical problems can be modeled mathematically as bilevel problems and appear in many practical applications. They are often encountered in the fields of production and capacity planning [1,2], traffic and transportation [3,4], chemistry [5,6] and management science [7,8], as well as energy networks and markets [9,10]. In addition, Nimana et al. [11] proposed an algorithm combining the incremental proximal gradient method with a smooth penalization technique to solve convex bilevel problems and applied it to image inpainting and binary classification problem.

Nowadays, we are in a world with various types of big data. In order to obtain the benefits of such data, we need to integrate advanced knowledge concerning both theory and methods from many areas, such as mathematics, computer science, statistics, medicine, etc. In mathematics, optimization plays a very important role in classifying and predicting large amounts of data because it can provide deep machine learning algorithms with high accuracy. Among optimization models for machine learning, the bilevel optimization model approach is an efficient one that makes it possible to create intelligent machine learning algorithms for data prediction and classification.

In this work, we study a bilevel problem that is an optimization problem where the constraint is another optimization problem. This problem is formulated as follows:

$$\min_{x \in S_*} \omega(x), \tag{1}$$

where $\omega : \mathbb{R}^n \to \mathbb{R}$ is assumed to be strongly convex and differentiable, while S_* is a nonempty set of inner-level problem minimizers given by

$$\min_{x \in \mathbb{R}^n} \{f(x) + g(x)\}, \tag{2}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable and convex function such that ∇f is L_f-Lipschitz-continuous and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is a lower-semicontinuous, proper, convex function.

It can be observed that the above bilevel optimization model contains both inner- and outer-level minimization problems (Equations (1) and (2)). Normally, the minimization problem in Equation (2) can be applied to data prediction and classification; see [12,13]. However, among the solutions to the inner-level problem (Equation (2)), we use the objective function ω to select solutions that are minimizers of ω. This method can provide more accuracy for data prediction and classification than Equation (2) alone.

The inner-level optimization problem is a constraint on the outer-level optimization problem. There are several algorithms for solving the problem in Equation (2); see [12,14,15].

The proximal gradient (PG) method, also known as the proximal forward–backward technique, is the basic algorithm used to solve the problem in Equation (2) (see [16,17]). It is defined by

$$x_{n+1} = prox_{\alpha_n g}(I - \alpha_n \nabla f)(x_n), \tag{3}$$

where $\alpha_n > 0$ is the step size, $prox_g$ is the proximity operator of g and ∇f is the gradient of f. The algorithm in Equation (3), which is also known as the forward–backward splitting algorithm (FBSA) [14], is suitable to solve Equation (2) if ∇f is L-Lipschitz-continuous. The FBSA is also called an iterative denoising method [18] or a fixed-point continuation algorithm [19].

One of the most well-known first-order optimization schemes is the fast iterative shrinkage-thresholding algorithm (FISTA). Beck and Teboulle [15] proposed the FISTA to solve the problem in Equation (2) by using an inertial technique as follows:

$$\begin{cases} w_n = prox_{\frac{1}{L}g}(I - \frac{1}{L}\nabla f)(x_n), \\ p_{n+1} = (1 + \sqrt{1 + 4p_n^2})/2, \\ x_{n+1} = y_n + \left(\frac{p_n - 1}{p_{n+1}}\right)(w_n - w_{n-1}), \ n \geq 1, \end{cases} \tag{4}$$

where $x_1 = w_0 \in \mathbb{R}^N$ and $p_1 = 1$. They applied the FISTA to image restoration problems and showed that the rate of convergence of the FISTA was better than other existing algorithms. The generated sequence's weak convergence was then proved by Liang and Schonlieb [20], who modified the FISTA by setting $p_{n+1} = (u + \sqrt{v + sp_n^2})/2$, where $u, v > 0$ and $0 < s \leq 4$.

It may be noticed that the convex minimization problem and fixed-point problem are related. If $0 < \alpha < 2/L$, then we know that a forward–backward operator $T := prox_{\alpha g}(I - \alpha \nabla f)$ is nonexpansive. It is known that $Fix(T) = argmin\{f(x) + g(x)\}$. Fixed-point problems with nonexpansive mappings have been investigated by many authors using the method of viscosity approximation [21–24]. This method provides a strong convergence result and it is defined by the following:

$$x_{n+1} = \beta_n S(x_n) + (1 - \beta_n)Tx_n, \ n \geq 1, \tag{5}$$

where $x_1 \in H$, $S : H \to H$ is a contraction when H is a Hilbert space and $\{\beta_n\} \in (0, 1)$. We can also call Equation (5) the viscosity forward–backward algorithm if $T := prox_{\alpha g}(I - \alpha \nabla f)$.

In 2014, Beck et al. [25] introduced a new, direct first-order method to solve the problem in Equation (1) and established its convergence results under some suitable conditions, as well as the rate of convergence of the sequence of function values. After that, Sabach and Shtern [26] proposed the following algorithm, called the bilevel gradient sequential

averaging method (BiG-SAM), to solve the problems in Equations (1) and (2). The iterative process can be defined as follows

$$
\begin{cases}
u_n = prox_{cg}(x_{n-1} - c\nabla f(x_{n-1})), \\
v_n = x_{n-1} - \lambda \nabla \omega(x_{n-1}), \\
x_{n+1} = \gamma_n v_n + (1 - \gamma_n)u_n, \ \ n \geq 1,
\end{cases}
\tag{6}
$$

where $x_0 \in \mathbb{R}^n$, $c \in (0, 1/L_f]$, $\lambda \in (0, 2/(L_\omega + \sigma)]$, in which L_f and L_ω are the Lipschitz constants of ∇f and $\nabla \omega$, and $\{\gamma_n\}$ satisfies certain conditions from [22]. In terms of the values of the inner objective function, the authors of [22] studied and analyzed the convergence behavior of the BiG-SAM with a nonasymptotic $O(1/n)$ global rate of convergence.

In 2019, Shehu et al. [27] introduced an inertial extrapolation step into BiG-SAM (Equation (6)), calling the result the inertial bilevel gradient sequential averaging method (iBiG-SAM), to solve the problems in Equations (1) and (2). This iterative scheme is defined by

$$
\begin{cases}
s_n = x_n + \theta_n(x_n - x_{n-1}), \\
u_n = prox_{cg}(I - c\nabla f)(s_n), \\
v_n = s_n - \lambda \nabla \omega(s_n), \\
x_{n+1} = \gamma_n v_n + (1 - \gamma_n)u_n, \ \ n \geq 1.
\end{cases}
\tag{7}
$$

In their study, they presented a strong convergence analysis of an inertial algorithm that can be used to approximate fixed points of nonexpansive mappings in infinite-dimensional real Hilbert spaces. Furthermore, they converted the bilevel optimization problems into a fixed-point problem of nonexpansive mappings and showed its convergence under certain conditions.

In 2022, Duan and Zhang [28] introduced an alternated inertial step into BiG-SAM to create an alternated inertial bilevel gradient sequential averaging method (aiBiG-SAM) for solving convex bilevel optimization problems. It is defined as

$$
s_n = \begin{cases}
x_n & \text{if } n \text{ is even,} \\
x_n + \theta_n(x_n - x_{n-1}) & \text{if } n \text{ is odd,}
\end{cases}
$$

and

$$
\begin{cases}
u_n = prox_{cg}(I - c\nabla f)(s_n), \\
v_n = s_n - \lambda \nabla \omega(s_n), \\
x_{n+1} = \gamma_n v_n + (1 - \gamma_n)u_n, \ \ n \geq 1.
\end{cases}
\tag{8}
$$

They proved that the aiBiG-SAM converges strongly to a solution for the problem and extended the method into a more general alternating inertial acceleration method.

Recently, in [29,30], the authors proposed new bilevel optimization methods within the framework of Hilbert spaces and proved the strong convergence of their algorithms using the viscosity approximation technique.

In this paper, motivated by these results, we present a novel accelerated algorithm using the viscosity approximation method and the inertial parameter of the FISTA to solve the convex bilevel optimization problem. We then demonstrate the efficacy of this algorithm in solving data classification problems.

The paper is organized as follows. In Section 2, we present the preliminaries in terms of definitions, notations and lemmas for proving the main results. The new accelerated viscosity-type algorithm is introduced and studied, and then we apply it to solving the convex bilevel optimization problems described in Section 3. Then, in Section 4, we present mathematical models for the classification of datasets and the application of the

results obtained in the previous section, and we provide numerical experimental results in Section 5. Finally, we present the conclusions and future work in Section 6.

2. Preliminaries

In this section, we present fundamental ideas and principles that will be utilized in the the rest of the research.

Throughout the present paper, H denotes a real Hilbert space with norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$, $\mathbb{R}$ is the set of real numbers and $\mathbb{N}$ is the set of positive integers. I denotes the identity operator on H. Let C be a nonempty subset of H and let T be a mapping of C into itself. The strong convergence of a sequence $\{x_n\}$ in H to $x \in H$ is denoted by $x_n \to x$, weak convergence by $x_n \rightharpoonup x$ and $Fix(T)$ symbolizes the set of all fixed points of T.

For this work, nonlinear mappings from the following classes were essentially needed.

Definition 1. *The mapping $T : C \to C$ is said to be L-Lipschitz with $L \geq 0$ if*

$$\|Tu - Tv\| \leq L\|u - v\|$$

for all $u, v \in C$.

An L-Lipschitz mapping T is said to be a contraction mapping if $L \in [0,1)$, and it is nonexpansive if $L = 1$.

Definition 2 ([31]). *Let T be a nonexpansive mapping of C into itself and let $T_n : C \to C$ be a family of nonexpansive mappings such that $\varnothing \neq Fix(T) \subset \Gamma := \bigcap_{n=1}^{\infty} Fix(T_n)$, where $Fix(T_n)$ is the set of all fixed points of T_n for each $n \geq 1$. Then, $\{T_n\}$ is said to satisfy the NST-condition (I) with T if, for any bounded sequence $\{x_n\} \subset C$,*

$$\lim_{n\to\infty} \|x_n - T_n x_n\| = 0 \ \text{ implies } \ \lim_{n\to\infty} \|x_n - T x_n\| = 0.$$

Definition 3 ([32,33]). *For any bounded sequence $\{u_n\}$ in H, a family $\{T_n\}$ of nonexpansive mappings $T_n : C \to C$ with $\bigcap_{n=1}^{\infty} Fix(T_n) \neq \varnothing$ is said to satisfy the condition (Z) if*

$$\lim_{n\to\infty} \|u_n - T_n u_n\| = 0.$$

Then, every weak cluster point of $\{u_n\} \in \bigcap_{n=1}^{\infty} Fix(T_n)$.

Using the demicloseness of $I - T$ where $T : C \to C$ is a nonexpansive mapping, we obtain the following remark.

Remark 1. *Let T be a nonexpansive mapping. Then, $\{T_n\}$ satisfies the condition (Z) if $\{T_n\}$ is a family of nonexpansive mappings that satisfies NST-condition (I) with T.*

The metric projection P_C from H onto C is defined by

$$P_C x = argmin\{\|x - y\| : y \in C\}$$

for all $x \in H$, where C is a nonempty closed convex subset of H. It is known that $v = P_C x$ if and only if $\langle x - v, y - v \rangle \leq 0$ for all $y \in C$.

Let us recall the definition of the proximity operator and its properties.

Definition 4 ([34,35]). *Let $f : H \to \mathbb{R} \cup \{\infty\}$ be a convex, proper and lower-semicontinuous function. The proximity operator of f, denoted by $prox_f$, is defined as follows:*

$$prox_f = \min_{y \in H} f(y) + \frac{1}{2}\|x - y\|^2$$

and it can be formulated in the equivalent form:

$$prox_f = (I + \partial f)^{-1},$$

where ∂f is the subdifferential of f defined by

$$\partial f(x) := \{v \in H : f(x) + \langle v, u - x \rangle \le f(u) \ \text{for all} \ u \in H\}$$

for all $x \in H$. For $\rho > 0$, we also know that $prox_{\rho f}$ is firmly nonexpansive and

$$Fix(prox_{\rho f}) = Argminf := \{v \in H : f(v) \le f(u) \ \text{for all} \ u \in H\}.$$

Let C be closed convex with $\varnothing \ne C \subset H$. In particular, if $f := i_C$, the indicator function on C is defined by

$$i_C(x) = \begin{cases} 0 & \text{if} \ x \in C, \\ \infty & \text{otherwise}. \end{cases}$$

Then, $prox_{\rho f} = P_C$.

The following lemmas are well known; see [13,36,37].

Lemma 1 ([13]). *Let $\mathfrak{g} : H \to \mathbb{R} \cup \{\infty\}$ be a lower-semicontinuous, proper and convex function and let $\mathfrak{f} : H \to \mathbb{R}$ be differentiable and convex such that $\nabla \mathfrak{f}$ is L-Lipschitz-continuous. Let*

$$T_n := prox_{\rho_n \mathfrak{g}}(I - \rho_n \nabla \mathfrak{f}) \quad and \quad T := prox_{\rho \mathfrak{g}}(I - \rho \nabla \mathfrak{f}),$$

where $\rho_n, \rho \in (0, 2/L)$ with $\rho_n \to \rho$ as $n \to \infty$. Then, $\{T_n\}$ satisfies the NST-condition (I) with T.

Lemma 2 ([36]). *Let $\eta, \mu \in H$ and $\zeta \in [0, 1]$. Then, the following properties hold for H:*
(1) $\|\eta + \mu\|^2 \le \|\eta\|^2 + 2\langle \mu, \eta + \mu \rangle$;
(2) $\|\eta \pm \mu\|^2 = \|\eta\|^2 \pm 2\langle \eta, \mu \rangle + \|\mu\|^2$;
(3) $\|\zeta \eta + (1 - \zeta)\mu\|^2 = \zeta \|\eta\|^2 + (1 - \zeta)\|\mu\|^2 - \zeta(1 - \zeta)\|\eta - \mu\|^2$.

Lemma 3 ([37]). *Let $\{c_n\} \subset \mathbb{R}_+$, $\{b_n\} \subset \mathbb{R}$ and $\{t_n\} \subset (0, 1)$ such that $\sum_{n=1}^{\infty} t_n = \infty$. Suppose that*

$$c_{n+1} \le (1 - t_n)c_n + t_n b_n$$

for all $n \in \mathbb{N}$. If $\limsup_{i \to \infty} b_{n_i} \le 0$, and for any subsequence $\{c_{n_i}\}$ of $\{c_n\}$ satisfying

$$\liminf_{i \to \infty}(c_{n_i+1} - c_{n_i}) \ge 0,$$

then $\lim_{n \to \infty} c_n = 0$.

In the next section, we introduce an inertial viscosity modified SP algorithm and its application to the convex bilevel optimization problem.

3. Proposed Method

We first present a new inertial viscosity algorithm and prove a strong convergence theorem under mild conditions as follows.

Let C be closed convex with $\varnothing \ne C \subset H$ and the mapping $S : C \to C$ be a k-contraction where $0 < k < 1$. Let $\{T_n\}$ be a family of nonexpansive mappings of C onto itself satisfying the condition (Z) such that $\Gamma := \bigcap_{n=1}^{\infty} Fix(T_n) \ne \varnothing$.

Many mathematicians often use inertial-type extrapolation [38,39] in optimization problems to speed up the convergence of iterative methods by using the technical term $\theta_n(x_n - x_{n-1})$. The momentum $x_n - x_{n-1}$ is controlled by the parameter θ_n, also known as an inertial parameter.

In 2012, Phuengrattana and Suantai [40] introduced an SP algorithm and showed that

its convergence behavior is better than Mann and Ishikawa iterations [41,42]. By using the idea of the SP algorithm, in this paper, we introduce an inertial viscosity modified SP algorithm (IVMSPA) for obtaining a common fixed point for $\{T_n\}$ as follows.

The following theorem establishes strong convergence for the proposed algorithm.

Theorem 1. *A sequence $\{x_n\}$ generated by Algorithm 1 converges strongly to an element $\breve{a} \in \Gamma$, where $\breve{a} = P_\Gamma S(\breve{a})$, provided that the sequences $\{\alpha_n\}$, $\{\beta_n\}$, $\{\gamma_n\}$ and $\{\tau_n\}$ satisfy the following conditions:*

(C1) $0 < a_1 \leq \beta_n \leq a_2 < 1$;

(C2) $0 < \alpha_n, \gamma_n < 1$, $\lim_{n \to \infty} \alpha_n = 0$ and $\sum_{n=1}^\infty \alpha_n = \infty$;

(C3) $\lim_{n \to \infty} \tau_n = 0$.

Algorithm 1 An Inertial Viscosity Modified SP Algorithm (IVMSPA)

Initialization: Let $\{\alpha_n\}$, $\{\beta_n\}$, $\{\gamma_n\}$ and $\{\tau_n\}$ be sequences of positive real numbers. Take $x_0, x_1 \in H$ arbitrarily.

Iterative steps: For $n \geq 1$, calculate x_{n+1} as follows:

Step 1. Compute an inertial parameter

$$\theta_n = \begin{cases} \min\left\{\dfrac{p_n - 1}{p_{n+1}}, \dfrac{\alpha_n \tau_n}{\|x_n - x_{n-1}\|}\right\} & \text{if } x_n \neq x_{n-1}, \\[2mm] \dfrac{p_n - 1}{p_{n+1}} & \text{otherwise,} \end{cases}$$

where $p_1 = 1$ and $p_{n+1} = \dfrac{1 + \sqrt{1 + 4p_n^2}}{2}$.

Step 2. Compute

$$\begin{aligned} y_n &= x_n + \theta_n(x_n - x_{n-1}), \\ z_n &= (1 - \alpha_n)y_n + \alpha_n S(y_n), \\ w_n &= (1 - \beta_n)z_n + \beta_n T_n z_n, \\ x_{n+1} &= (1 - \gamma_n)w_n + \gamma_n T_n w_n. \end{aligned}$$

Proof. Let $\breve{a} = P_\Gamma S(\breve{a})$. Then, $\breve{a} \in \bigcap_{n=1}^\infty F(T_n)$. First of all, we show that $\{x_n\}$ is bounded. From Algorithm 1, we have

$$\begin{aligned} \|w_n - \breve{a}\| &\leq (1 - \beta_n)\|z_n - \breve{a}\| + \beta_n\|T_n z_n - \breve{a}\| \\ &\leq (1 - \beta_n)\|z_n - \breve{a}\| + \beta_n\|z_n - \breve{a}\| \\ &= \|z_n - \breve{a}\| \end{aligned} \tag{9}$$

and

$$\begin{aligned} \|x_{n+1} - \breve{a}\| &\leq \gamma_n\|w_n - \breve{a}\| + (1 - \gamma_n)\|T_n w_n - \breve{a}\| \\ &\leq \gamma_n\|w_n - \breve{a}\| + (1 - \gamma_n)\|w_n - \breve{a}\| \\ &= \|w_n - \breve{a}\| \\ &\leq \|z_n - \breve{a}\|. \end{aligned} \tag{10}$$

From the definition of y_n and z_n, we obtain

$$\begin{aligned} \|z_n - \breve{a}\| &\leq \alpha_n\|S(y_n) - \breve{a}\| + (1 - \alpha_n)\|y_n - \breve{a}\| \\ &\leq \alpha_n\|S(y_n) - S(\breve{a})\| + \alpha_n\|S(\breve{a}) - \breve{a}\| + (1 - \alpha_n)\|y_n - \breve{a}\| \\ &\leq \alpha_n k\|y_n - \breve{a}\| + \alpha_n\|S(\breve{a}) - \breve{a}\| + (1 - \alpha_n)\|y_n - \breve{a}\| \\ &= \left[1 - \alpha_n(1 - k)\right]\|y_n - \breve{a}\| + \alpha_n\|S(\breve{a}) - \breve{a}\| \end{aligned}$$

$$\leq \left[1 - \alpha_n(1-k)\right]\left[\|x_n - \breve{a}\| + \theta_n\|x_{n-1} - x_n\|\right] + \alpha_n\|S(\breve{a}) - \breve{a}\|$$

$$\leq \left[1 - \alpha_n(1-k)\right]\|x_n - \breve{a}\| + \alpha_n\left(\frac{\theta_n}{\alpha_n}\|x_{n-1} - x_n\| + \|S(\breve{a}) - \breve{a}\|\right).$$

From (C3), we have

$$\frac{\theta_n}{\alpha_n}\|x_{n-1} - x_n\| \to 0 \quad \text{as} \quad n \to \infty. \tag{11}$$

From Equation (11), we know that there exists $M > 0$ such that $\frac{\theta_n}{\alpha_n}\|x_{n-1} - x_n\| \leq M$ for all $n \geq 1$. Thus,

$$\|z_n - \breve{a}\| \leq \left[1 - (1-k)\alpha_n\right]\|x_n - \breve{a}\| + (1-k)\alpha_n\left(\frac{\|S(\breve{a}) - \breve{a}\| + M}{1-k}\right)$$

$$\leq \max\left\{\|x_n - \breve{a}\|, \frac{\|S(\breve{a}) - \breve{a}\| + M}{1-k}\right\}.$$

From Equation (10) and the above inequality, we obtain

$$\|x_{n+1} - \breve{a}\| \leq \max\left\{\|x_n - \breve{a}\|, \frac{\|S(\breve{a}) - \breve{a}\| + M}{1-k}\right\}.$$

Using mathematical induction, we have

$$\|x_n - \breve{a}\| \leq \max\left\{\|x_1 - \breve{a}\|, \frac{\|S(\breve{a}) - \breve{a}\| + M}{1-k}\right\}$$

for all $n \geq 1$. It follows that $\{x_n\}$ is bounded and, hence, $\{z_n\}$ is bounded. According to part (3) of Lemma 2, we obtain

$$\begin{aligned}
\|x_{n+1} - \breve{a}\|^2 &= \gamma_n\|T_n w_n - \breve{a}\|^2 + (1-\gamma_n)\|w_n - \breve{a}\|^2 - (1-\gamma_n)\gamma_n\|w_n - T_n w_n\|^2 \\
&\leq (1-\gamma_n)\|w_n - \breve{a}\|^2 + \gamma_n\|w_n - \breve{a}\|^2 \\
&= \|w_n - \breve{a}\|^2
\end{aligned} \tag{12}$$

and

$$\begin{aligned}
\|w_n - \breve{a}\|^2 &= \beta_n\|T_n z_n - \breve{a}\|^2 + (1-\beta_n)\|z_n - \breve{a}\|^2 - (1-\beta_n)\beta_n\|z_n - T_n z_n\|^2 \\
&= \|z_n - \breve{a}\|^2 - (1-\beta_n)\beta_n\|z_n - T_n z_n\|^2.
\end{aligned} \tag{13}$$

Using Lemma 2, we obtain

$$\begin{aligned}
\|z_n - \breve{a}\|^2 &\leq \|(1-\alpha_n)(y_n - \breve{a}) + \alpha_n(S(y_n) - S(\breve{a}))\|^2 + 2\alpha_n\langle S(\breve{a}) - \breve{a}, z_n - \breve{a}\rangle \\
&\leq (1-\alpha_n)\|y_n - \breve{a}\|^2 + \alpha_n\|S(y_n) - S(\breve{a})\|^2 + 2\alpha_n\langle S(\breve{a}) - \breve{a}, z_n - \breve{a}\rangle \\
&\leq (1-\alpha_n)\|y_n - \breve{a}\|^2 + \alpha_n k\|y_n - \breve{a}\|^2 + 2\alpha_n\langle S(\breve{a}) - \breve{a}, z_n - \breve{a}\rangle \\
&= \left[1 - \alpha_n(1-k)\right]\|y_n - \breve{a}\|^2 + 2\alpha_n\langle S(\breve{a}) - \breve{a}, z_n - \breve{a}\rangle
\end{aligned} \tag{14}$$

and

$$\|y_n - \breve{a}\|^2 \leq \|x_n - \breve{a}\|^2 + 2\theta_n\|x_n - \breve{a}\|\|x_{n-1} - x_n\| + \theta_n^2\|x_{n-1} - x_n\|^2. \tag{15}$$

From Equations (12)–(15), we obtain

$$\|x_{n+1} - \breve{a}\|^2 \leq \|z_n - \breve{a}\|^2 - (1 - \beta_n)\beta_n\|z_n - T_n z_n\|^2$$

$$= \left[1 - (1-k)\alpha_n\right]\|y_n - \breve{a}\|^2 + 2\alpha_n\langle S(\breve{a}) - \breve{a}, z_n - \breve{a}\rangle$$

$$- (1 - \beta_n)\beta_n\|z_n - T_n z_n\|^2$$

$$\leq \left[1 - (1-k)\alpha_n\right]\|x_n - \breve{a}\|^2 + 2\theta_n\|x_n - \breve{a}\|\|x_{n-1} - x_n\| + \theta_n^2\|x_{n-1} - x_n\|^2$$

$$+ 2\alpha_n\langle S(\breve{a}) - \breve{a}, z_n - \breve{a}\rangle - (1 - \beta_n)\beta_n\|z_n - T_n z_n\|^2 \tag{16}$$

$$= \left[1 - (1-k)\alpha_n\right]\|x_n - \breve{a}\|^2 - (1 - \beta_n)\beta_n\|z_n - T_n z_n\|^2 + (1-k)\alpha_n b_n,$$

where

$$b_n = \frac{1}{1-k}\left[2\langle S(\breve{a}) - \breve{a}, z_n - \breve{a}\rangle + \theta_n\|x_{n-1} - x_n\|\left(\frac{\theta_n}{\alpha_n}\|x_{n-1} - x_n\|\right)\right.$$

$$\left. + 2\|x_n - \breve{a}\|\left(\frac{\theta_n}{\alpha_n}\|x_{n-1} - x_n\|\right)\right].$$

It follows that

$$(1 - \beta_n)\beta_n\|z_n - T_n z_n\|^2 \leq \|x_n - \breve{a}\|^2 - \|x_{n+1} - \breve{a}\|^2 + (1-k)\alpha_n B', \tag{17}$$

where $B' = \sup\{b_n : n \in \mathbb{N}\}$.

We next show that $\{x_n\}$ converges strongly to $\breve{a}$. To apply Lemma 3, we set $a_n := \|x_n - \breve{a}\|^2$ and $t_n := \alpha_n(1-k)$. From Equation (16), we obtain

$$a_{n+1} \leq (1 - t_n)a_n + t_n b_n.$$

Suppose that $\{a_{n_i}\}$ is a subsequence of $\{a_n\}$ such that $\liminf_{i\to\infty}(a_{n_i+1} - a_{n_i}) \geq 0$. Using Equation (17) and (C2), we obtain

$$\limsup_{i\to\infty} \beta_{n_i}(1 - \beta_{n_i})\|z_{n_i} - T_{n_i}z_{n_i}\|^2 \leq \limsup_{i\to\infty}\left(a_{n_i} - a_{n_i+1} + \alpha_{n_i}(1-k)B'\right)$$

$$= \limsup_{i\to\infty}(a_{n_i+1} - a_{n_i})$$

$$\leq 0. \tag{18}$$

From (C1) and Equation (18), we obtain

$$\lim_{i\to\infty}\|z_{n_i} - T_{n_i}z_{n_i}\| = 0. \tag{19}$$

Next, we show that $\limsup_{i\to\infty} b_{n_i} \leq 0$. Obviously, it suffices to show that

$$\limsup_{i\to\infty}\langle S(\breve{a}) - \breve{a}, z_{n_i} - \breve{a}\rangle \leq 0.$$

Since $\{z_{n_i}\}$ is bounded, there exists a subsequence $\{z_{n_{i_j}}\}$ of $\{z_{n_i}\}$ and $y \in H$ such that $z_{n_{i_j}} \rightharpoonup y$ as $j \to \infty$ and

$$\limsup_{i\to\infty}\langle S(\breve{a}) - \breve{a}, z_{n_i} - \breve{a}\rangle = \lim_{j\to\infty}\left\langle S(\breve{a}) - \breve{a}, z_{n_{i_j}} - \breve{a}\right\rangle.$$

Since $\{T_n\}$ satisfies the condition (Z) and Equation (19), we obtain $y \in \Gamma$. Using $\breve{a} = P_\Gamma S(\breve{a})$, we get

$$\lim_{j\to\infty}\left\langle S(\breve{a}) - \breve{a}, z_{n_{i_j}} - \breve{a}\right\rangle \leq 0.$$

So, we have

$$\limsup_{i\to\infty} \langle S(\breve{a}) - \breve{a}, z_{n_i} - \breve{a} \rangle \leq 0.$$

Thus, in view of Lemma 3, $\{x_n\}$ converges to $\breve{a}$ as required. $\quad\square$

We next establish our inertial bilevel gradient modified SP algorithm (IBiG-MSPA) to solve the convex bilevel optimization problems in Equations (1) and (2) by applying Algorithm 1 and present its strong convergence. We use the following assumptions in order to solve this problem:

- $f : H \to \mathbb{R}$ is a convex and differentiable function such that ∇f is Lipschitz-continuous with constant $f > 0$ and $g : H \to (-\infty, \infty]$ are proper, lower-semicontinuous and convex functions;
- $\omega : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with a parameter σ such that $\nabla\omega$ is L_ω-Lipschitz continuous and $s \in (0, \frac{2}{L_\omega + \sigma})$.

  Our IBiG-MSPA algorithm is defined as follows.

  The following useful result was proved by Sabach and Shtern [26].

Proposition 1. *Suppose that $\omega : \mathbb{R}^n \to \mathbb{R}$ is strongly convex with $\sigma > 0$ and $\nabla\omega$ is Lipschitz-continuous with constant L_ω. Then, the mapping defined by $S_s = I - s\nabla\omega$ is a contraction for all $s \in (0, \frac{2}{\sigma + L_\omega})$. Thus,*

$$\|x - s\nabla\omega(u) - (v - s\nabla\omega(v))\| \leq \sqrt{1 - \frac{2s\sigma L_\omega}{\sigma + L_\omega}} \|u - v\|$$

for all $u, v \in \mathbb{R}^n$.

Combining Theorem 1 and Proposition 1, we obtain the following result.

Theorem 2. *Let Λ be the set of all solutions to Equation (1) and $\breve{a} = P_{S_*}(I - s\nabla\omega)(\breve{a})$ and let (C1)–(C3) in Theorem 1 hold. Then, $\{x_n\}$ generated by Algorithm 2 converges strongly to $\breve{a} \in \Lambda$.*

Algorithm 2 An Inertial Bilevel Gradient Modified SP Algorithm (IBiG-MSPA)

Initialization: Let $\{\alpha_n\}$, $\{\beta_n\}$, $\{\gamma_n\}$, $\{\tau_n\}$ and $\{c_n\}$ be sequences of positive real numbers. Take $x_0, x_1 \in H$ arbitrarily.

Iterative steps: For $n \geq 1$, calculate x_{n+1} as follows:

Step 1. Compute an inertial parameter

$$\theta_n = \begin{cases} \min\left\{\frac{p_n - 1}{p_{n+1}}, \frac{\alpha_n \tau_n}{\|x_n - x_{n-1}\|}\right\} & \text{if } x_n \neq x_{n-1}, \\ \frac{p_n - 1}{p_{n+1}} & \text{otherwise,} \end{cases}$$

where $p_1 = 1$ and $p_{n+1} = \frac{1 + \sqrt{1 + 4p_n^2}}{2}$.

Step 2. Compute

$$\begin{aligned} y_n &= x_n + \theta_n(x_n - x_{n-1}), \\ z_n &= (1 - \alpha_n)y_n + \alpha_n(I - s\nabla\omega)y_n, \\ w_n &= (1 - \beta_n)z_n + \beta_n prox_{c_n g}(I - c_n\nabla f)z_n, \\ x_{n+1} &= (1 - \gamma_n)w_n + \gamma_n prox_{c_n g}(I - c_n\nabla f)w_n. \end{aligned}$$

Proof. Set $S = I - s\nabla\omega$ and $T_n = prox_{c_n g}(I - c_n\nabla f)$. Then, according to Proposition 1, S is a contraction mapping. We also know that T_n is nonexpansive. Using Theorem 1,

we can conclude that $x_n \to \breve{a} \in \Gamma$, where $\breve{a} = P_\Gamma S(\breve{a})$. It can be noted that, in this case, $\Gamma = \bigcap_{n=1}^{\infty} F(T_n) = S_*$. Then, for all $x \in S_*$, we have

$$0 \geq \langle S(\breve{a}) - \breve{a}, x - \breve{a} \rangle = \langle \breve{a} - s\nabla\omega(\breve{a}) - \breve{a}, x - \breve{a} \rangle = \langle -s\nabla\omega(\breve{a}), x - \breve{a} \rangle.$$

Dividing the above inequalities by $-s$, we obtain

$$\langle \nabla\omega(\breve{a}), x - \breve{a} \rangle \geq 0$$

for all $x \in S_*$. Then, $x_n \to \breve{a} \in \Lambda$. This completes the proof. $\square$

Using our main results (Theorems 1 and 2), we apply the IBiG-MSPA in the next section to solve a classification problem.

4. Applications with Classification Problems

There are several mathematical models used for the classification of datasets. For this paper, we use the extreme learning machine model and present the advantages of this model as follows.

The advantages of feedforward neural networks have led to their widespread use in diverse fields over the past few decades. Stated concisely, feedforward neural networks allow for the approximation of complex nonlinear mappings directly from input samples and provide models for numerous natural and artificial phenomena that are difficult to deal with using classical parametric techniques. However, the rendering of feedforward neural networks is time-consuming due to the dependence of the parameters of the different layers and the requirement to configure all of the parameters. One of the widely used feedforward neural networks is the single hidden layer feedforward neural network (SLFN). It has been widely studied in terms of both theory and application because of its learning ability and anti-error ability (see [43–45] for more detail).

In order to increase the effectiveness of SLFNs, a development model of a neural learning algorithm called the extreme learning machine (ELM) [46] was recently established. The advantage of the ELM is that hidden node learning parameters, such as input weights and biases, are generated at random and do not need to be adjusted, whereas the output weights can be obtained analytically by using a simply generalized inverse operation. The ELM has been effectively used in several real-world applications, including regression and classification problems [47,48].

We next examine some aspects of the ELM regarding the classification of data.

Let $\{(x_l, t_l) : x_l \in \mathbb{R}^n, t_l \in \mathbb{R}^m, l = 1, 2, \ldots, N\}$ be a set of training data taken from different samples with a total sample size N, where $x_l = [x_{l1}, x_{l2}, \ldots, x_{ln}] \in \mathbb{R}^n$ and $t_l = [t_{l1}, t_{l2}, \ldots, t_{lm}] \in \mathbb{R}^m$ are the input data and target data, respectively. The mathematical formula for an ELM with M hidden nodes is as follows:

$$\sum_{r=1}^{M} K_r \mathfrak{E}(\langle v_r, x_s \rangle + a_r) = o_s, \ s = 1, \ldots, N,$$

where $\mathfrak{E}(x)$ represents the activation function, $K_r = [K_{r1}, K_{r2}, \ldots, K_{rm}]^T$ is the weight that links the r-th output node and the hidden node, $v_r = [v_{r1}, v_{r2}, \ldots, v_{rn}]^T$ is the weight that links the r-th input nodes and the hidden node and a_r is a bias.

The purpose of SLFNs is the prediction of N output nodes that satisfy $\sum_{s=1}^{N} \|o_s - t_s\| = 0$. This means that there exist K_r, v_r and a_r such that

$$\sum_{r=1}^{M} K_r \mathfrak{E}(\langle v_r, x_s \rangle + a_r) = t_s, \ s = 1, \ldots, N.$$

From the above system of linear equations, we can rewrite the following:

$$\mathbf{H}K = \mathbf{T},$$

where

$$\mathbf{H} = \begin{bmatrix} \mathfrak{E}(\langle v_1, x_1 \rangle + a_1) & \cdots & \mathfrak{E}(\langle v_M, x_1 \rangle + a_M) \\ \vdots & \ddots & \vdots \\ \mathfrak{E}(\langle v_1, x_N \rangle + a_1) & \cdots & \mathfrak{E}(\langle v_M, x_N \rangle + a_M) \end{bmatrix},$$

$K = [K_1^T, \dots, K_M^T]_{m \times M}^T$, and $\mathbf{T} = [t_1^T, \dots, t_N^T]_{m \times N}^T$.

As the Moore–Penrose generalized inverse $\breve{H}$ of $\mathbf{H}$ exists, K can be obtained from $K = \breve{H}\mathbf{T}$ (see [46]). If $\breve{H}$ does not exist, then it could be impossible to find K using this approach. To solve this issue, we determine K as a minimizer of the ordinary least squares minimization problem (OLS):

$$\min_{K} \|\mathbf{H}K - \mathbf{T}\|_2^2, \tag{20}$$

where $\mathbf{H} \in \mathbb{R}^{N \times M}$ is called the hidden layer output matrix, $K \in \mathbb{R}^{M \times m}$ is the weight of the output layer, $\mathbf{T} \in \mathbb{R}^{N \times m}$ is the training data target matrix, M is the number of hidden nodes and N is the number of training samples.

However, in a real situation, the use of OLS (Equation (20)) may cause an overfitting problem. To overcome such problems, the output weight K can be approximated with the least absolute shrinkage and selection operator (lasso) (see [49]):

$$\min_{K} \|\mathbf{H}K - \mathbf{T}\|_2^2 + \lambda \|K\|_1, \tag{21}$$

where λ is a regularization parameter. Now, let S_* be the set of all solutions to Equation (21). Among the solutions in S_*, we would like to select a solution $K^* \in S_*$ in such a way that K^* is a minimizer of

$$\min_{K \in S_*} \frac{1}{2} \|K\|^2. \tag{22}$$

Our aim in the next section is to employ the IBiG-MSPA to solve the convex bilevel optimization problems in Equations (21) and (22) and to use the obtained optimal weight for classification of the Diabetes [50], Heart Disease UCI [51] and Iris datasets [52]. These databases are widely used as benchmarks in many research works in the area of data classification.

5. Numerical Experiments

In this section, we present the experimental results from applying our proposed algorithm to classify the Diabetes, Heart Disease UCI and Iris datasets.

We employed our algorithm (IBiG-MSPA) to solve the convex bilevel optimization problems in Equations (21) and (22) by setting $\omega(K) = \frac{1}{2}\|K\|_2^2$, $f(K) = \|\mathbf{H}K - \mathbf{T}\|_2^2$, $g(K) = \lambda\|K\|_1$ and $\mathfrak{E}(x)$ as sigmoid.

The parameters selected for this experiment are shown in Table 1, where $L_f = 2\|\mathbf{H}\|^2$. We measured the efficiency of each algorithm using the output data accuracy as follows:

$$\text{accuracy} = 100 \times \frac{\text{correct prediction}}{\text{total cases}}.$$

Next, we used the Diabetes, Heart Disease UCI and Iris datasets for classification, as described below.

Diabetes dataset [50]: There are nine features in the dataset. We categorized two data classes in this dataset.

Heart Disease UCI dataset [51]: The dataset has 14 features. Patients' heart disease data are presented in this dataset. We wanted to categorize the data into two groups.

Iris dataset [52]: The dataset contains four features and three classes. We aimed to classify the data into three types of iris plants.

Testing and training data are shown in Table 2.

Table 1. Parameter selection for the IBIG-MSPA, BiG-SAM, iBiG-SAM and aiBiG-SAM.

Methods	Settings
IBiG-MSPA	$s = 0.01, c_n = \frac{1}{L_f}, \alpha_n = \frac{1}{50n}, \beta_n = \gamma_n = 0.5, \tau_n = \frac{10^{20}}{n}$
BiG-SAM	$\lambda = 0.01, c = \frac{1}{L_f}, \gamma_n = \frac{2(0.1)}{1 - \frac{2+cL_f}{4}}$
iBiG-SAM	$\lambda = 0.01, c = \frac{1}{L_f}, \alpha = 3, \gamma_n = \frac{2(0.1)}{1 - \frac{2+cL_f}{4}}, \beta_n = \frac{\gamma_n}{n^{0.01}}$ $\theta_n = \begin{cases} \min\left\{\frac{n}{n+\alpha-1}, \frac{\beta_n}{\|x_n - x_{n-1}\|}\right\} & \text{if } x_n \neq x_{n-1}, \\ \frac{n}{n+\alpha-1} & \text{otherwise.} \end{cases}$
aiBiG-SAM	$\lambda = 0.01, c = \frac{1}{L_f}, \gamma_n = \frac{2(0.1)}{1 - \frac{2+cL_f}{4}}, \beta_n = \frac{\gamma_n}{n^{0.01}}$ $\theta_n = \begin{cases} \min\left\{\frac{n}{n+\alpha-1}, \frac{\beta_n}{\|x_n - x_{n-1}\|}\right\} & \text{if } x_n \neq x_{n-1}, \\ \frac{n}{n+\alpha-1} & \text{otherwise.} \end{cases}$

Table 2. Diabetes, Heart Disease UCI and Iris datasets, with 30% of each dataset used for testing and 70% for training.

Datasets	Features	Sample	
		Testing Set	Training Set
Diabetes	9	230	538
Heart Disease UCI	14	90	213
Iris	4	45	105

For each dataset, the numbers of iterations and hidden nodes were determined as follows Table 3:

Table 3. Numbers of iterations and hidden nodes specified for each data collection.

Datasets	Number of Iterations ($\bar{I}$)	Number of Hidden Nodes (M)
Diabetes	200	100
Heart Disease UCI	100	60
Iris	300	30

The number of iterations for each dataset was chosen to produce the best results for each method under consideration, as can be seen in Table 2.

We conducted experiments to compare the efficiency of the IBiG-MSPA with other algorithms under consideration; namely, the BiG-SAM, iBiG-SAM and aiBiG-SAM.

As representations of the accuracy of testing and training, we use the terms Ac.Test and Ac.Train, respectively, in Table 4.

Table 4. Accuracy of predictions using various algorithms.

Dataset	IBiG-MSPA		BiG-SAM		iBiG-SAM		aiBiG-SaM	
	Ac.Train	Ac.Test	Ac.Train	Ac.Test	Ac.Train	Ac.Test	Ac.Train	Ac.Test
Diabetes	77.11	81.08	71.98	76.13	72.34	76.58	70.88	73.42
Heart Disease UCI	85.71	83.87	74.76	74.19	82.38	78.49	83.81	79.57
Iris	99.05	100.00	94.29	95.56	94.29	95.56	94.29	95.56

The results from Table 4 show that our proposed method, the IBiG-MSPA, performed better than the BiG-SAM, iBiG-SAM and aiBiG-SAM in terms of training and testing

accuracy for each dataset. Therefore, based on our study, the IBiG-MSPA could classify the chosen datasets with greater accuracy than the other methods.

We can observe that, for the Heart Disease UCI dataset, the accuracies of the existing algorithms were around 70%, while our proposed algorithm achieved accuracy over 80%. In a practical scenario, even small improvements in classification accuracy can have significant effects. For instance, in the case of medical diagnoses, for which the Heart Disease UCI dataset is often used as a benchmark, a slight increase in accuracy can lead to more reliable predictions and better patient outcomes. It may help identify more individuals at risk or improve the overall efficiency of the classification process, leading to appropriate treatments. This observation applies equally well to the other two datasets and datasets similar to them.

6. Conclusions

We first introduced an inertial viscosity modified SP algorithm (IVMSPA). Then, the strong convergence of the IVMSPA was proved under mild conditions with the control sequence. Next, we proposed the inertial bilevel gradient modified SP algorithm (IBiG-MSPA) to solve the convex bilevel optimization problem. Finally, we applied our method to classifying the Diabetes, Heart Disease UCI and Iris datasets. The numerical experiments showed that the IBiG-MSPA had higher efficiency than the BiG-SAM, iBiG-SAM and aiBiG-SAM.

The performances of the algorithms discussed in this paper depend in part on the characteristics of the datasets. In order to improve the accuracy, one needs to address issues related to the preprocessing of the data, such as feature selection, missing data and dataset imbalances. The goal of our future research is to develop new techniques or processes that can improve the efficiency of algorithms in classifying imbalanced and big datasets.

Author Contributions: Conceptualization, R.W.; Formal analysis, K.J. and R.W.; Investigation, K.J. and R.W.; Methodology, R.W.; Software, K.J.; Supervision, S.S. and Y.J.C.; Validation, S.S., Y.J.C., A.K. and R.W.; Visualization, R.W.; Writing—original draft, K.J.; Writing—review and editing, S.S., Y.J.C., A.K. and R.W. All authors have discussed the results and approved the final manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in this paper were obtained from https://archive.ics.uci.edu (accessed on 30 October 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Garcia-Herreros, P.; Zhang, L.; Misra, P.; Arslan, E.; Mehta, S.; Grossmann, I.E. Mixed-integer bilevel optimization for capacity planning with rational markets. *Comput. Chem. Eng.* **2016**, *86*, 33–47. [CrossRef]
2. Maravillo, H.; Camacho-Vallejo, J.; Puerto, J.; Labbé, M. A market regulation bilevel problem: A case study of the mexican petrochemical industry. *Omega* **2019**, *97*, 102–105. [CrossRef]
3. Marcotte, P. Network design problem with congestion effects: A case of bilevel programming. *Math. Program.* **1986**, *34*, 142–162. [CrossRef]
4. Migdalas, A. Bilevel programming in traffic planning: Models, methods and challenge. *J. Glob. Optim.* **1995**, *7*, 381–405. [CrossRef]
5. Clark, P.A.; Westerberg, A.W. Bilevel programming for steady-state chemical process design–I. Fundamentals and algorithms. *Comput. Chem. Eng.* **1990**, *14*, 87–97. [CrossRef]
6. Dempe, S. *Foundations of Bi-Level Programming*, 1st ed.; Springer: New York, NY, USA, 2002.
7. Bard, J.F. Coordination of a multidivisional organization through two levels of management. *Omega* **1983**, *11*, 457–468. [CrossRef]
8. Dan, T.; Marcotte, P. Competitive facility location with selfish users and queues. *Oper. Res.* **2019**, *67*, 479–497. [CrossRef]

9. Gabriel, S.A.; Conejo, A.J.; Fuller, J.D.; Hobbs, B.F.; Ruiz, C. *Complementarity Modeling in Energy*, 1st ed.; Springer: New York, NY, USA, 2012.

10. Wogrin, S.; Pineda, S.; Tejada-Arango, D.A. Applications of bilevel optimization in energy and electricity markets. *Bilevel Optim.* **2020**, *161*, 139–168.

11. Nimana, N.; Petrot, N. Incremental proximal gradient scheme with penalization for constrained composite convex optimization problems. *Optimization* **2019**, *70*, 1307–1336. [CrossRef]

12. Janngam, K.; Suantai, S. An inertial modified S-Algorithm for convex minimization problems with directed graphs and their applications in classification problems. *Mathematics* **2022**, *10*, 4442. [CrossRef]

13. Bussaban, L.; Suantai, S.; Kaewkhao, A. A parallel inertial S-iteration forward-backward algorithm for regression and classification problems. *Carpathian J. Math.* **2020**, *36*, 35–44. [CrossRef]

14. Lions, P.L.; Mercier, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.* **1979**, *16*, 964–979. [CrossRef]

15. Beck, A.; Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2009**, *2*, 183–202. [CrossRef]

16. Bruck, R.E., Jr. On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space. *J. Math. Anal. Appl.* **1977**, *61*, 159–164. [CrossRef]

17. Passty, G.B. Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. *J. Math. Anal. Appl.* **1979**, *72*, 383–390. [CrossRef]

18. Figueiredo, M.; Nowak, R. An EM algorithm for wavelet-based image restoration. *IEEE Trans. Image Process.* **2003**, *12*, 906–916. [CrossRef]

19. Hale, E.; Yin, W.; Zhang, Y. *A Fixed-Point Continuation Method for l1-Regularized Minimization with Applications to Compressed Sensing*; Department of Computational and Applied Mathematics, Rice University: Houston, TX, USA, 2007.

20. Liang, J.; Schonlieb, C. B. Improving fista: Faster, smarter and greedier. *arXiv* **2018**, arXiv:1811.01430.

21. Moudafi, A. Viscosity approximation method for fixed-points problems. *J. Math. Anal. Appl.* **2000**, *241*, 46–55. [CrossRef]

22. Xu, H.K. Viscosity approximation methods for nonexpansive mappings. *J. Math. Anal. Appl.* **2004**, *298*, 279–291. [CrossRef]

23. Jailoka, P.; Suantai, S. and Hanjing, A. A fast viscosity forward-backward algorithm for convex minimization problems with an application in image recovery. *Carpathian J. Math.* **2020**, *37*, 449–461. [CrossRef]

24. Tan, B.; Zhou, Z.; Li, S. Viscosity-type inertial extragradient algorithms for solving variational inequality problems and fixed point problems. *J. Appl. Math. Comput.* **2022**, *68*, 1387–1411. [CrossRef]

25. Beck, A.; Sabach, S. A first order method for finding minimal norm-like solutions of convex optimization problems. *Math. Program.* **2014**, *147*, 25–46. [CrossRef]

26. Sabach, S.; Shtern, S. A first order method for solving convex bilevel optimization problems. *SIAM J. Optim.* **2017**, *27*, 640–660. [CrossRef]

27. Shehu, Y.; Vuong, PT.; Zemkoho, A. An inertial extrapolation method for convex simple bilevel optimization. *Optim. Methods Softw.* **2019**, *2019*, 1–19. [CrossRef]

28. Duan, P.; Zhang, Y. Alternated and multi-step inertial approximation methods for solving convex bilevel optimization problems. *Optimization* **2022**, *2022*, 1–30 . [CrossRef]

29. Jiang, R.; Abolfazli, N.; Mokhtari, A.; Hamedani, E.Y. A Conditional Gradient-based Method for Simple Bilevel Optimization with Convex Lower-level Problem. In Proceedings of the 26th International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 25–27 April 2023; pp. 10305–10323.

30. Thongsri, P.; Panyanak, B.; Suantai, S. A New Accelerated Algorithm Based on Fixed Point Method for Convex Bilevel Optimization Problems with Applications. *Mathematics* **2023**, *11*, 702. [CrossRef]

31. Nakajo, K.; Shimoji, K.; Takahashi, W. Strong convergence to a common fixed point of families of nonexpansive mappings in Banach spaces. *J. Nonlinear Convex Anal.* **2007**, *8*, 11–34.

32. Aoyama, K.; Kimura, Y. Strong convergence theorems for strongly nonexpansive sequences. *Appl. Math. Comput.* **2011**, *217*, 7537–7545. [CrossRef]

33. Aoyama, K.; Kohsaka, F.; Takahashi, W. Strong convergence theorems by shrinking and hybrid projection methods for relatively nonexpansive mappings in Banach spaces. In *Nonlinear Analysis and Convex Analysis, Proceedings of the 5th International Conference on Nonlinear Analysis and Convex Analysi, Hsinchu, Taiwan, 31 May–4 June 2007*; Yokohama Publishers: Yokohama, Japan, 2009; pp. 7–26.

34. Moreau, J.J. Fonctions convexes duales et points proximaux dans un espace hilbertien. *Comptes Rendus Acad. Sci. Paris Ser. A Math.* **1962**, *255*, 2897–2899.

35. Bauschke, H.H.; Combettes, P.L. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*; Springer: New York, NY, USA, 2011.

36. Takahashi, W. *Introduction to Nonlinear and Convex Analysis*; Yokohama Publishers: Yokohama, Japan, 2009.

37. Saejung, S.; Yotkaew, P. Approximation of zeros of inverse strongly monotone operators in Banach spaces. *Nonlinear Anal.* **2012**, *75*, 724–750. [CrossRef]

38. Polyak, B.T. Some methods of speeding up the convergence of iteration methods. *Ussr Comput. Math. Math. Phys.* **1964**, *4*, 1–17. [CrossRef]

39. Nesterov, Y. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Dokl. Akad. Nauk SSSR* **1983**, *269*, 543–547.

40. Phuengrattana, W.; Suantai, S. On the rate of convergence of Mann, Ishikawa, Noor and SP-iterations for continuous functions on an arbitrary interval. *J. Comput. Appl. Math.* **2000**, 235, 3006–3014. [CrossRef]

41. Mann, W.R. Mean value methods in iteration. *Proc. Am. Math. Soc.* **1953**, 4, 506–510. [CrossRef]

42. Ishikawa, S. Fixed points by a new iteration method. *Proc. Am. Math. Soc.* **1974**, 44, 147–150. [CrossRef]

43. Ding, S.F.; Jia, W.K.; Su, C.Y.; Zhang, L.W. Research of neural network algorithm based on factor analysis and cluster analysis. *Neural Comput. Appl.* **2011**, *20*, 297–302. [CrossRef]

44. Razavi, S.; Tolson, B.A. A new formulation for feedforward neural networks. *IEEE Trans. Neural Netw.* **2011**, *22*, 1588–1598. [CrossRef]

45. Chen, Y.; Zheng, W.X. Stochastic state estimation for neural networks with distributed delays and Markovian jump. *Neural Netw.* **2012**, *25*, 14–20. [CrossRef]

46. Huang, G.B.; Zhu, Q.Y.; Siew, C.K. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]

47. Rong, H.J.; Ong, Y.S.; Tan, A.H.; Zhu, Z. A fast pruned-extreme learning machine for classification problem. *Neurocomputing* **2008**, *72*, 359–366. [CrossRef]

48. Huang, G.B.; Ding, X.; Zhou, H. Optimization method based extreme learning machine for classification. *Neurocomputing* **2010**, *74*, 155–163. [CrossRef]

49. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B Methodol.* **1996**, *58*, 267–288. [CrossRef]

50. Smith, J.W.; Everhart, J.E.; Dickson, W.C.; Knowler, D.C.; Johannes, R.S. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Symposium on Computer Applications and Medical Care*; IEEE Computer Society Press: New York, NY, USA, 1998; pp. 261–265.

51. Lichman, M. UCI Machine Learning Repository. Available online: https://archive.ics.uci.edu (accessed on 20 October 2022).

52. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]