CARPATHIAN J. MATH. Online version at https://www.carpathian.cunbm.utcluj.ro
36 (2020), No. 1, 35 - 44 Print Edition: ISSN 1584 - 2851 Online Edition: ISSN 1843 - 4401

Dedicated to Prof. Hong-Kun Xu on the occasion of his 60^{th} anniversary

A parallel inertial S-iteration forward-backward algorithm for regression and classification problems

LIMPAPAT BUSSABAN[1], SUTHEP SUANTAI[2] and ATTAPOL KAEWKHAO[3]

ABSTRACT. In this paper, a novel algorithm, called parallel inertial S-iteration forward-backward algorithm (PISFBA) is proposed for finding a common fixed point of a countable family of nonexpansive mappings and convergence behavior of PISFBA is analyzed and discussed. As applications, we apply PISFBA to estimate the weight connecting the hidden layer and output layer in a regularized extreme learning machine. Finally, the proposed learning algorithm is applied to solve regression and data classification problems.

1. INTRODUCTION

In the past decade, *Extreme learning machine* (ELM) [7], a new learning algorithm for single-hidden layer feedforward networks (SLFNs), has been extensively studied in various research topics for machine learning and artificial intelligence such as face classification, image segmentation, regression and data classification problems. ELM was proved in theory that it has extremely fast learning speed and good performance better than the gradient-based learning such as backpropagation in most of the cases. The target of this model is to find the parameter β that solves the following minimization problem, called *ordinary least square* (OLS),

$$\min_{\beta} \|\mathbf{H}\beta - \mathbf{T}\|_2^2 , \tag{1.1}$$

where $\|\cdot\|_2$ is l_2-norm defined by $\|x\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$, $\mathbf{T} \in \mathbb{R}^{N\times m}$ is the target of data, $\beta \in \mathbb{R}^{M\times m}$ is a weight which connects hidden layer and output layer and $\mathbf{H} \in \mathbb{R}^{N\times M}$ is the hidden layer output matrix. In general mathematical modeling, there are several methods to estimate the solution of (1.1), in this case, the solution β obtained by $\beta = \mathbf{H}^{\dagger}\mathbf{T}$ where $\mathbf{H}^{\dagger}$ is the MoorePenrose generalized inverse of $\mathbf{H}$. However, in real situation, the number of unknown variable M is much more than the number of training data N which causes the network may lead to overfitting. On the other hand, the accuracy is low while the number of hidden nodes M is small. Thus, in order to improve (1.1), several *regularization methods*, were introduced. The classical two standard techniques for improving (1.1) are subset selection and ridge regression (sometimes called Tikhonov regularization) [20]. In this paper, we focus on the following problem, called *least absolute shrinkage and selection operator* (LASSO) [19],

$$\min_{\beta} \|\mathbf{H}\beta - \mathbf{T}\|_2^2 + \lambda \|\beta\|_1 , \tag{1.2}$$

Received: 26.02.2019. In revised form: 31.07.2019. Accepted: 07.08.2019

2010 *Mathematics Subject Classification.* 47H09, 90C25, 65F22.

Key words and phrases. *S-iteration, forward-backward algorithm, nonexpansive operator, regression, data classification problems.*

Corresponding author: Suthep Suantai; suthep.s@cmu.ac.th

where $\|\cdot\|_1$ is l_1-norm defined by $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ and $\lambda > 0$, called regularization parameter. LASSO tries to retain the good features of both subset selection and ridge regression [19]. After the regularization methods and the original ELM were introduced for improving performance of OLS, 5 years later, the *regularized extreme learning machine* [11] was proposed and applied to solve regression problems. It is noted that LASSO is a special case of eq.(5) [11] by setting $\alpha = 1$. In general, (1.2) can be rewritten as minimization of sum of two functions,

$$\min_{x} F(x) := f(x) + g(x), \tag{1.3}$$

where g is a convex smooth (or possible non-smooth) function and f is a smooth convex loss function with gradient having Lipschitz constant L. By using Fermats rule, Theorem 16.3 in [3], the solution of (1.3) can be characterized as follows: $\bar{x}$ minimizing $(f+g)$ if and only if $0 \in \partial g(\bar{x}) + \nabla f(\bar{x})$ where $\partial g(\bar{x})$ and $\nabla f(\bar{x})$ refer to the subdifferential and gradient of g and f, respectively. In fixed point theory, the solution of (1.3) can be characterized [14] as follows: $\bar{x}$ is a minimizer of $f + g$ if and only if

$$\bar{x} = prox_{cg}(I - c\nabla f)(\bar{x}) = J_{c\partial g}(I - c\nabla f)(\bar{x}), \tag{1.4}$$

where $c > 0$, I is an identity operator, $prox_{cg}$ is the proximity operator of cg and $J_{\partial g}$ is the resolvent of ∂g defined by $J_{\partial g} = (I + \partial g)^{-1}$, more description of these operators will be mentioned in Section 2. For convenience, (1.4) can be rewritten as:

$$\bar{x} = T\bar{x}, \tag{1.5}$$

where $T := prox_{cg}(I - c\nabla f)$ which is called *forward-backward operator*. It is observed that a solution of (1.5) is a fixed point of T and T is a nonexpansive mapping when $c \in (0, \frac{2}{L})$. The existence of a fixed point of nonexpansive mappings was guaranteed by Browder's theorem, see [1] for detail. In order to find a point $\bar{x}$ satisfying (1.5), many researchers proposed various methods for finding the approximate solution. In this paper, we consider one of iterative method for finding fixed piont of an operator T, called *S-iteration process* [2], defined as follows:

$$\begin{cases} y_n = (1-\beta_n)x_n + \beta_n T x_n, \\ x_{n+1} = (1-\alpha_n)Tx_n + \alpha_n T y_n, n \geq 1, \end{cases} \tag{1.6}$$

where initial point x_1 is chosen randomly and $\{\alpha_n\}$, $\{\beta_n\}$ are sequences in $[0, 1]$. In [2], Agarwal, O'Regan and Sahu proved that this iteration process is independent of Mann and Ishikawa iteration process and converges faster than both of them. However, the speed of convergence of S-iteration process is needed to be improved. Thus, to speed up, the technique for improving speed and giving a better convergence behavior was introduced firstly by [15] by adding an *inertial step*. The following iterative methods improve performance of Forward-backward algorithm by adding inertial step.

A fast iterative shrinkage-thresholding algorithm (FISTA), [4], is defined by:

$$\begin{cases} y_n = Tx_n, \\ t_{n+1} = \dfrac{1 + \sqrt{1 + 4t_n^2}}{2}, \qquad \theta_n = \dfrac{t_n - 1}{t_{n+1}}, \\ x_{n+1} = y_n + \theta_n(y_n - y_{n-1}), \end{cases} \tag{1.7}$$

where $x_1 = y_0 \in \mathbb{R}^n$, $t_1 = 1$, $T := prox_{\frac{1}{L}g}(I - \frac{1}{L}\nabla f)$ and θ_n is called inertial step size. FISTA was suggested by Beck and Teboulle [4]. They proved that rate of convergence of FISTA is better than that of ISTA and applied FISTA to image deblurring problems [4]. The inertial step size θ_n of FISTA was firstly introduced by Nesterov [13]. Generally, FISTA was modified for improving its performance by replacing t_{n+1} with appropriate sequences. For example, Chambolle and Dossal [5] turned out t_{n+1} to be $\frac{n+a}{a}$ for $a > 2$, Liang and Schönlieb [8] interpolated t_{n+1} into a general form as $t_{n+1} = \frac{p+\sqrt{q+rt_n^2}}{2}$ where $p, q > 0$ and $0 < r \leq 4$ and proved weak convergence theorem of FISTA.

A new accelerated proximal gradient algorithm (nAGA) [21] was defined by Verma and Shukla as the following:

$$\begin{cases} y_n = x_n + \theta_n(x_n - x_{n-1}), \\ x_{n+1} = T_n[(1-\alpha_n)y_n + \alpha_n T_n y_n], \end{cases} \tag{1.8}$$

where $\{\theta_n\}$, $\{\alpha_n\}$ are sequences in $(0,1)$ and $\dfrac{\|x_n - x_{n-1}\|_2}{\theta_n} \to 0$. They proved a convergence theorem of nAGA and applied this method for solving the non-smooth convex minimization problem with sparsity-inducing regularizes for the multitask learning framework.

Motivated by those works mentioned above, in this paper, a novel iterative method for sloving (1.3) is proposed by employing the concepts of S-iteration process together with the inertial step for a countable family of nonexpansive mappings. This paper is organized as follows: The basic concept and mathematical background will be given in Section 2. A weak convergence theorem will be proved in Section 3. Moreover, in Section 4, we apply the proposed method for solving regression and data classification problems.

2. Preliminaries

2.1. **Mathmatical Background.** Let H be a real Hilbert space with norm $\|\cdot\|$ and inner product $\langle\cdot|\cdot\rangle$. A mapping $T : H \to H$ is said to be *L-Lipschtiz operator* if there exists $L > 0$ such that $\|Tx - Ty\| \leq L\|x - y\|$ for any $x, y \in H$. An L-Lipschitz operator is called *nonexpansive operator* if $L = 1$. A mapping $A : H \to 2^H$ is called *monotone operator* if

$$\langle x - y | u - v\rangle \geq 0,$$

for any $(x, u), (y, v) \in graA$, where $graA = \{(x, y) \in H \times H : x \in H, y \in Ax\}$ is the graph of A. A monotone operator A is called *maximal monotone operator* if the graph $graA$ is not properly contained in the graph of any other monotone operator. It is known that A is maximal monotone operator if and only if $R(I + \lambda A) = H$ for every $\lambda > 0$.

Let $A : H \to 2^H$ be a maximal monotone operator and $c > 0$. The *resolvent* of A is defined by $J_{cA} = (I + cA)^{-1}$ where I is an identity operator. If $A = \partial f$ for some $f \in \Gamma_0(H)$, $\Gamma_0(H)$ is denoted by the set of proper lower semicontinuous convex functions from H to $(-\infty, +\infty]$, then $J_{cA} = prox_{cf}$ where $prox_f$ is *proximity operator* [3] of f given by

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

If $f = \|\cdot\|_1$, then $prox_{cf}$ can be represented by

$$prox_{c\|\cdot\|_1}(x) = sgn(x)max\{\|x\|_1 - c, 0\},$$

see Chapter 24 in [3] for detial.

Let $\{T_n\}$ and $\mathcal{T}$ be a families of nonexpansive operators such that $\emptyset \neq F(\mathcal{T}) \subset \bigcap_{n=1}^{\infty} F(T_n)$, where $F(\mathcal{T})$ is the set of all common fixed points of $T \in \mathcal{T}$. Then, $\{T_n\}$ is said to satisfy *NST-condition(I) with* $\mathcal{T}$ [12, 17] if for each bounded sequence $\{x_n\}$,

$$\lim_{n\to\infty} \|x_n - T_n x_n\| = 0 \text{ implies } \lim_{n\to\infty} \|x_n - Tx_n\| = 0 \text{ for all } T \in \mathcal{T}.$$

If $\mathcal{T}$ is singleton, i.e. $\mathcal{T} = \{T\}$, then $\{T_n\}$ is said to satisfy NST-condition(I) with T.

Proposition 2.1. *Let H be a Hilbert space and let $A : H \to 2^H$ be a maximal monotone operator and $B : H \to H$ be an L-Lipschitz operator. Let $\alpha > 0$ and $x, p \in H$. Setting $\tilde{A}_\alpha = \frac{1}{\alpha}(I - J_{\alpha A}(I - \alpha B))$. Then, the following hold:*

(i) $\tilde{A}_\alpha x \in AJ_{\alpha A}(I - \alpha B)x + Bx$.

(ii) $p \in \tilde{A}_\alpha x$ *if and only if* $(x - \alpha p, p - Bx) \in graA$.

(iii) $\left\|\tilde{A}_\alpha x\right\| \leq \|Ax + Bx\|$ *where* $\|Ax + Bx\| := \inf\{\|z\| : z \in Ax + Bx\}$.

Proof.

(i) Let $u = \tilde{A}_\alpha x$. Then $x - \alpha u = J_{\alpha A}(I - \alpha B)x$ which implies that $u - Bx \in A(x - \alpha u)$. Thus, $\tilde{A}_\alpha x \in AJ_{\alpha A}(I - \alpha B)x + Bx$.

(ii) By using definition of $\tilde{A}_\alpha$ and $J_{\alpha A}$, we have

$$\begin{aligned} p = \tilde{A}_\alpha x \Leftrightarrow x - \alpha p = J_{\alpha A}(I - \alpha B)x \Leftrightarrow (I - \alpha B)x \in (I + \alpha A)(x - \alpha p) \\ \Leftrightarrow p - Bx \in A(x - \alpha p) \Leftrightarrow (x - \alpha p, p - Bx) \in graA. \end{aligned}$$

(iii) Let $w = \tilde{A}_\alpha x$ and $u \in Ax + Bx$. Then, by monotonic of A, we have

$$\langle u - w | w \rangle = \frac{1}{\alpha}\langle (u - Bx) - (w - Bx) | x - (x - \alpha w) \rangle \geq 0.$$

By CauchySchwarz inequality, we obtain that $\|w\| \leq \|u\|$. Thus,

$$\left\|\tilde{A}_\alpha x\right\| = \inf\{\|z\| : z \in \tilde{A}_\alpha x\} \leq \inf\{\|z\| : z \in Ax + Bx\} = \|Ax + Bx\|.$$

□

Lemma 2.1. ([18]) *Let $\{a_n\}, \{b_n\}$ and $\{\delta_n\}$ be sequences of nonnegative numbers such that*

$$a_{n+1} \leq (1 + \delta_n)a_n + b_n, \forall n \in \mathbb{N}.$$

If $\sum_{n=1}^{\infty} \delta_n < \infty$ and $\sum_{n=1}^{\infty} b_n < \infty$, then $\lim_{n\to\infty} a_n$ exists.

Lemma 2.2 (Opial lemma). *Let H be a Hilbert space and $\{x_n\}$ be a sequence in H such that there exists a nonempty subset Ω of H satisfying the following conditions:*

- *for all $y \in \Omega$, $\lim_{n\to\infty} \|x_n - y\|$ exists,*
- *Any weak-cluster point of $\{x_n\}$ belongs to Ω.*

Then, there exists $\bar{x} \in \Omega$ such that $x_n \rightharpoonup \bar{x}$.

2.2. **Extreme Learning Machine.** Let $D = \{(\mathbf{x}_i, \mathbf{t}_i) : \mathbf{x}_i \in \mathbb{R}^n, \mathbf{t}_i \in \mathbb{R}^m, i = 1, 2, \ldots, N\}$ be a training set with N distinct samples, $\mathbf{x}_i$ and $\mathbf{t}_i$ is called *input data* and *target*, respectively. A standard SLFNs with M hidden nodes and activation function $\Phi(x)$, e.g. sigmoid, are mathematically modeled as

$$\sum_{j=1}^{M} \beta_j \Phi(\langle \mathbf{w}_j | \mathbf{x}_i \rangle + b_j) = \mathbf{o}_i, i = 1, \ldots, N, \tag{2.9}$$

where $\mathbf{w}_j$ is the weight vector connecting the jth hidden node and the input node, β_j is the weight vector connecting the jth hidden node and the output node, and b_j is the threshold of the jth hidden node. The target of standard SLFNs is to approximate these N samples with zero error means that $\sum_{i=1}^{N} \|\mathbf{o}_i - \mathbf{t}_i\| = 0$, i.e., there exist $\beta_j, \mathbf{w}_j, b_j$ such that

$$\sum_{j=1}^{M} \beta_j \Phi(\langle \mathbf{w}_j | \mathbf{x}_i \rangle + b_j) = \mathbf{t}_i, i = 1, \ldots, N. \tag{2.10}$$

From above N equations, we can formulate a simple equation as

$$\mathbf{H}\beta = \mathbf{T}, \tag{2.11}$$

where

$$\mathbf{H} = \begin{bmatrix} \Phi(\langle \mathbf{w}_1 | \mathbf{x}_1 \rangle + b_1) & \cdots & \Phi(\langle \mathbf{w}_M | \mathbf{x}_1 \rangle + b_M) \\ \vdots & \ddots & \vdots \\ \Phi(\langle \mathbf{w}_1 | \mathbf{x}_N \rangle + b_1) & \cdots & \Phi(\langle \mathbf{w}_M | \mathbf{x}_N \rangle + b_M) \end{bmatrix}_{N\times M},$$

$$\beta = \left[\beta_1^T, \dots, \beta_M^T\right]_{m\times M}^T, \mathbf{T} = \left[\mathbf{t}_1^T, \dots, \mathbf{t}_N^T\right]_{m\times N}^T.$$

The goal of a standard SLFNs is to estimate $\beta_j, \mathbf{w}_j$ and b_j for solving (2.11) while ELM aim to find only β_j with randomly $\mathbf{w}_j$ and b_j. An original ELM algorithm is defined as follows.

ELM algorithm: Given a training set $D = \{(\mathbf{x}_i, \mathbf{t}_i) : \mathbf{x}_i \in \mathbb{R}^n, \mathbf{t}_i \in \mathbb{R}^m, i = 1, 2, \dots, N\}$, activation function $g(x)$, and hidden node number M,

Step 1: Randomly $\mathbf{w}_i$ and $b_i, i = 1, \dots, M$.
Step 2: Calculate the hidden layer output matrix $\mathbf{H}$.
Step 3: Calculate β by

$$\beta = \mathbf{H}^\dagger \mathbf{T},$$

where $\mathbf{H}^\dagger$ is the MoorePenrose generalized inverse of matrix $\mathbf{H}$ [16] and $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_N]^T$.

3. Main results

In this section, we propose a new iterative method, called *parallel inertial S-iteration forward-backward algorithm* (PISFBA), for findind a fixed point of a countable family of nonexpansive mappings. Convergence theorems of PISFBA are proved in 3.1 and the proposed learning algorithm base on ELM will be discussed in 3.2.

3.1. Convergence Theorems of PISFBA.

Theorem 3.1. *Let H be a Hilbert space, $\{T_n\}$ be a family of nonexpansive operators of H into itself and $T : H \to H$ be a nonexpansive operator such that $\{T_n\}$ satisfies NST-condition(I) with T. Suppose that $\emptyset \neq F(T) \subset \cap_{n=1}^{\infty} F(T_n)$. Let $\{x_n\}$ be a sequence in H generated by*

$$\begin{cases} x_0, x_1 \in H, \\ y_n = x_n + \theta_n(x_n - x_{n-1}), \\ z_n = (1-\beta_n)x_n + \beta_n T_n x_n, \\ x_{n+1} = (1-\alpha_n)T_n y_n + \alpha_n T_n z_n, \end{cases} \tag{3.12}$$

where $0 < q < \alpha_n \leq 1$, $0 < s < \beta_n < r < 1$, $0 \leq \theta_n \leq 1$ and $\sum_{n=1}^{\infty} \theta_n \|x_n - x_{n-1}\| < \infty$. Then, $\{x_n\}$ converges weakly to a point in $F(T)$.

Proof. Let $x^* \in F(T)$ and let $\{x_n\} \subset H$ be generated by (3.12). Then,

$$\begin{aligned} \|x_{n+1} - x^*\| &\leq (1-\alpha_n)\|T_n y_n - x^*\| + \alpha_n \|T_n z_n - x^*\| \\ &\leq (1-\alpha_n)\|y_n - x^*\| + \alpha_n \|z_n - x^*\| \\ &\leq (1-\alpha_n)\|x_n - x^*\| + (1-\alpha_n)\theta_n \|x_n - x_{n-1}\| + \alpha_n \|z_n - x^*\| \\ &\leq \|x_n - x^*\| + \theta_n \|x_n - x_{n-1}\|. \end{aligned}$$

Thus by Lemma 2.1, $\lim_{n\to\infty}\|x_n - x^*\|$ exists which implies that $\{x_n\}$ is bounded. By (3.12), we have

$$\begin{aligned}\|y_n - x^*\|^2 &= \|x_n - x^*\|^2 + \theta_n^2\|x_n - x_{n-1}\|^2 + 2\theta_n\langle x_n - x^* | x_n - x_{n-1}\rangle\\ &\le \|x_n - x^*\|^2 + \theta_n^2\|x_n - x_{n-1}\|^2 + 2\theta_n\|x_n - x^*\|\|x_n - x_{n-1}\|.\end{aligned}$$

Then,

$$\begin{aligned}\|z_n - x^*\|^2 &= (1-\beta_n)\|x_n - x^*\|^2 + \beta_n\|T_nx_n - x^*\|^2 - \beta_n(1-\beta_n)\|x_n - T_nx_n\|^2\\ &\le \|x_n - x^*\|^2 - \beta_n(1-\beta_n)\|x_n - T_nx_n\|^2,\end{aligned}$$

and

$$\begin{aligned}\|x_{n+1} - x^*\|^2 &= (1-\alpha_n)\|T_ny_n - x^*\|^2 + \alpha_n\|T_nz_n - x^*\|^2 - \alpha_n(1-\alpha_n)\|T_ny_n - T_nz_n\|^2\\ &\le (1-\alpha_n)\|T_ny_n - x^*\|^2 + \alpha_n\|T_nz_n - x^*\|^2\\ &\le (1-\alpha_n)\|y_n - x^*\|^2 + \alpha_n\|z_n - x^*\|^2\\ &\le \|x_n - x^*\|^2 + (1-\alpha_n)\theta_n^2\|x_n - x_{n-1}\|^2\\ &\quad + 2(1-\alpha_n)\theta_n\|x_n - x^*\|\|x_n - x_{n-1}\| - \alpha_n\beta_n(1-\beta_n)\|x_n - T_nx_n\|^2.\end{aligned}$$

As $n\to\infty$, we have $\lim_{n\to\infty}\|x_n - T_nx_n\| = 0$. Since $\{x_n\}$ is bounded and $\{T_n\}$ satisfies NST-condition(I) with T, we obtain that $\|x_n - Tx_n\|\to 0$. Thus, by using Opial lemma, $\{x_n\}$ weakly converges to some point in $F(T)$. □

Lemma 3.3. *Let H be a Hilbert space. Let $A : H \to 2^H$ be a maximal monotone operator and $B : H \to H$ be an L-Lipschitz operator. Let $\alpha, \beta > 0$. Then,*

$$\frac{1}{\beta}\|J_{\alpha A}(I-\alpha B)x - J_{\beta A}(I-\beta B)J_{\alpha A}(I-\alpha B)x\| \le \frac{1+\alpha L}{\alpha}\|x - J_{\alpha A}(I-\alpha B)x\|,$$

for every $x\in H$.

Proof. Let $x\in H$. Set $\tilde{A}_\alpha = \frac{1}{\alpha}(I - J_{\alpha A}(I-\alpha B))$. Then, by using Proposition 2.1, we obtain

$$\begin{aligned}&\frac{1}{\beta}\|J_{\alpha A}(I-\alpha B)x - J_{\beta A}(I-\beta B)J_{\alpha A}(I-\alpha B)x\|\\ &= \left\|\tilde{A}_\beta J_{\alpha A}(I-\alpha B)x\right\|\\ &\le \|AJ_{\alpha A}(I-\alpha B)x + BJ_{\alpha A}(I-\alpha B)x\|\\ &\le \|AJ_{\alpha A}(I-\alpha B)x + Bx\| + \|Bx - BJ_{\alpha A}(I-\alpha B)x\|\\ &\le \left\|\tilde{A}_\alpha x\right\| + L\|x - J_{\alpha A}(I-\alpha B)x\|\\ &= \frac{1+\alpha L}{\alpha}\|x - J_{\alpha A}(I-\alpha B)x\|.\end{aligned}$$

□

Theorem 3.2. *Let H be a Hilbert space. Let $A : H \to 2^H$ be a maximal monotone operator and $B : H \to H$ be an L-Lipschitz operator. Let $c \in (0, \frac{2}{L})$ and $\{c_n\} \subset (0, \frac{2}{L})$ such that $c_n \to c$. Define $T_n = J_{c_nA}(I - c_nB)$. Then, $\{T_n\}$ satisfies the NST-condition(I) with T_c where $T_c = J_{cA}(I-cB)$.*

Proof. Let $\{x_n\}$ be a bounded sequence in H. Suppose that $\|x_n - T_nx_n\|\to 0$. Since T_n and T_c are nonexpansive for all $n\in\mathbb{N}$, see Theorem 26.14 in [3] for detail, by Lemma 3.3, we obtain that

$$\begin{aligned}\|x_n - T_c x_n\| &= \|x_n - J_{cA}(I - cB)x_n\| \\ &\le \|x_n - J_{c_nA}(I - c_nB)x_n\| + \|J_{c_nA}(I - c_nB)x_n - J_{cA}(I - cB)J_{c_nA}(I - c_nB)x_n\| \\ &\quad + \|J_{cA}(I - cB)J_{c_nA}(I - c_nB)x_n - J_{cA}(I - cB)x_n\| \\ &\le 2\|x_n - J_{c_nA}(I - c_nB)x_n\| + \frac{c(1 + c_nL)}{c_n}\|x_n - J_{c_nA}(I - c_nB)x_n\| \\ &= (2 + \frac{c(1 + c_nL)}{c_n})\|x_n - T_nx_n\| \to 0\end{aligned}$$

Thus, $\{T_n\}$ satisfies the NST-condition(I) with T_c. □

Corollary 3.1. *Let H be a Hilbert space. Let $A : H \to 2^H$ be maximal monotone operator and $B : H \to H$ be an L-Lipschitz operator. Let $c \in (0, \frac{2}{L})$ and $\{c_n\} \subset (0, \frac{2}{L})$ such that $c_n \to c$. Define $T_n = J_{c_nA}(I - c_nB)$ and $T = J_{cA}(I - cB)$. Suppose that $zer(A + B) \neq \emptyset$. Let $\{x_n\}$ be a sequence in H generated by (3.12). Then, $\{x_n\}$ converges weakly to a point in $zer(A + B)$.*

Proof. Using Proposition 26.1(iv)(a) in [3] , we have $F(T) = F(T_n) = zer(A + B)$ and we know that $\{T_n\}$ and T are nonexpansive operators for all n, by Proposition 26.1(iv)(d) in [3]. Then, the proof is completed by Theorem 3.1 and Theorem 3.2. □

Corollary 3.2 (PISFBA). *Let H be a Hilbert space. Let $g \in \Gamma_0(H)$ and $f : H \to \mathbb{R}$ be convex and differentiable with an L-Lipschitz continuous gradient, let $c \in (0, \frac{2}{L})$ and $\{c_n\} \subset (0, \frac{2}{L})$ such that $c_n \to c$. Define $T_n = prox_{c_ng}(I - c_n\nabla f)$ and $T = prox_{cg}(I - c\nabla f)$. Suppose that $argmin(f + g) \neq \emptyset$. Let $\{x_n\}$ be a sequence in H generated by (3.12). Then, $\{x_n\}$ converges weakly to a point in $argmin(f + g)$.*

Proof. Setting $A := \partial g$ and $B := \nabla f$, then A is maximal monotone operator, by using Theorem 20.25 in [3]. The proof is completed by Corollary 3.1. □

3.2. **Proposed Learning Algorithm.** In this section, we apply PISFBA to propose a learning algorithm base on ELM algorithm for solving (1.2).

Modified regularized ELM algorithm: Given a training set $D = \{(\mathbf{x}_i, \mathbf{t}_i) : \mathbf{x}_i \in \mathbb{R}^n, \mathbf{t}_i \in \mathbb{R}^m, i = 1, 2, \dots, N\}$, activation function $g(x)$,

Step 1: Select regularization parameter λ and hidden node number M.
Step 2: Randomly $\mathbf{w}_i$ and $b_i, i = 1, \dots, M$.
Step 3: Calculate the hidden layer output matrix **H**.
Step 4: Calculate β by using PISFBA (Corollary 3.2) or nAGA (1.8) or FISTA (1.7).

4. Performance evaluation and analysis

In this section, we predict a function sine and classify datasets by the proposed learning algorithm. All results are performed on Intel Core-i7 gen 8th with 8.00 GB RAM, windows 10, under MATLAB computing environment.

4.1. **Regression a function sine.** In order to regression a function sine, we set a training set by randomly 10 distinct data with command *unifrnd*, activation function is sigmoid, regularization parameter $\lambda = 1 \times 10^{-5}$, $\alpha_n = \beta_n = \frac{0.9n}{n+1}$ and $\theta_n = \frac{1}{2^n \|x_n - x_{n-1}\|}$, if $x_n \neq x_{n-1}$; =0, otherwise. Then, we obtain comparision result with FISTA and nAGA in Figure 1 and mean squre error (MSE) in Table 1.

Method	MSE	Computational time
PISFBA	$\mathbf{1.775601 \times 10^{-3}}$	2.159882×10^{-1}
FISTA	7.667023×10^{-2}	$\mathbf{3.903217 \times 10^{-2}}$
nAGA	1.213818×10^{-2}	1.764385×10^{-1}

TABLE 1. Numerical results of regression of a function sine.

Table 1 and Figure 1 show that PISFBA gives a better performance to predict a function sine than FISTA and nAGA while a computational time have a few difference. However, in this simple problem, the original ELM is not only the best performance by 8.824458×10^{-5} of MSE, but also the fastest computational time by 1.303190×10^{-4} of computational time.

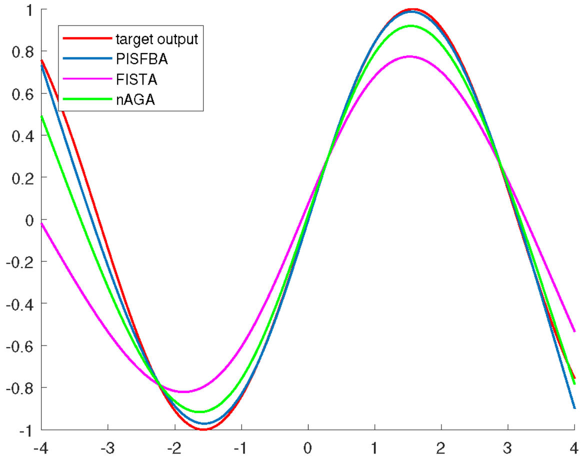

FIGURE 1. A simple regression for a function sine.

4.2. **Data classification.** In order to classify datasets, we would to thanks "https://archive.ics.uci.edu/" and "https://www.kaggle.com/" for supporting database website.

- **Parkinsons dataset** [10] This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). The aim of this dataset is to classify people who healthy and PD.
- **Heart Disease UCI dataset** [9] The original dataset contains 76 attributes, but all published experiments refer to using a subset of 14 of them. This dataset refers to the presence of heart disease in the patient. The predicted attribute is aim to classify the data into 2 classes.
- **Iris dataset** [6] This dataset contains 3 classes of 50 instances where each class refers to a type of iris plant. The aim is to separate each type of iris plant (iris setosa, iris versicolour and iris virginica) from sepal and petal length.
- **Abalone dataset** [6] This dataset aim to predict the age of abalone from physical measurements by counting the number of rings through a microscope and classification into 3 classes.

Table 2 show information about the datasets, number of attributes and number of samples for training (around 70% of data) and testing (remainder 30% of data) sets.

Dataset	# Attributes	Samples	
		# Train	# Test
Binary Classes Datasets			
Parkinsons	23	135	60
Heart Disease UCI	14	213	90
Multiple Classes Datasets			
Iris	4	105	45
Abalone	8	2924	1253

TABLE 2. Information about the datasets.

Setting all controls $(\lambda, \alpha_n, \beta_n, \theta_n)$ as in section 4.1, activation function is sigmoid, and the suitable number of hidden nodes M is selected, see Table 3. Given a training set for each dataset as mentioned in Table 2. We use our proposed learning algorithm with three difference iterative methods (PISFBA, nAGA, FISTA) to estimate the optimal weight β and then the output data $\mathbf{O}$ of training and testing sets are obtained by $\mathbf{O} = \mathbf{H}\beta$. Compared with the target data, an accuracy of output data is computed by

$$\text{accuracy} = \frac{\text{\# correct predicted data}}{\text{\# all data}} \times 100.$$

Table 3 shows the performance in term of the accuracy of each methods compared with the original ELM. However, the performance depend on the number of hidden nodes M. A huge number M may lead the prediction model to overfitting. In our experimental results, we pick the number M in the different way depend on each dataset. A suitable number M is selected when the absolutely difference of accuracy of training and accuracy of testing is small (less than 5% for this case), however, the optimal process to selected the number M remains open and dose not discussed in this paper.

Dataset	Original ELM	Regularized ELM		
		PISFBA	nAGA	FISTA
Binary Classes Datasets				
Parkinsons ($M = 37$)	66.67	75	75	48.33
Heart Disease UCI ($M = 106$)	55.56	70	68.89	65.56
Multiple Classes Datasets				
Iris ($M = 12$)	97.78	100	95.56	86.67
Abalone ($M = 100$)	66.32	63.61	62.09	60.02

TABLE 3. Performance comparison using difference methods.

From the results in Table 3, we conclude that the proposed learning algorithm under selection with the identical number of hidden nodes M has a high performance in term of the accuracy. The weight computed by PISFBA converges faster to the optimal weight and performs accuracy better than those computed by nAGA and FISTA.

Acknowledgments. This research was supported by Chiang Mai University. The first and the second authors would like to thanks the Thailand Research Fund through the Royal Golden Jubilee (RGJ) PhD Programme (Grant No. PHD/0184/2560).

REFERENCES

[1] Agarwal, R., O'Regan, D. and Sahu, D., *Fixed Point Theory for Lipschitzian-type Mappings with Applications*, Topological Fixed Point Theory and Its Applications, Springer New York, 2009

[2] Agarwal, R. P., Regan, D. O' and Sahu, D. R., *Iterative construction of fixed point of nearly asymptotically nonexpansive mappings*, J. Nonlinear Convex Anal., **8** (2007), No. 1, 61–79,

[3] Bauschke, H. H. and Combettes, P. L., *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer Publishing Company, Incorporated, 2nd edition, 2017

[4] Beck, A. and Teboulle, M., *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, **2** (2009), No. 1, 183–202

[5] Chambolle, A. and Dossal, C., *On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm"*, J. Optim. Theory Appl., **166** (2015), No. 3, 968–982

[6] Dua, D. and Taniskidou, E. Karra, UCI machine learning repository, 2017

[7] Huang, G.-B., Zhu, Q.-Y. and Siew, C.-K., *Extreme learning machine: Theory and applications*, Neurocomputing, **70** (2006), No. 1, 489–501

[8] Liang, J. and Schnlieb, C.-B., *Improving fista: Faster, smarter and greedier*, abs/1811.01430, 2018

[9] Lichman, M., *UCI machine learning repository*, 2013

[10] Little, M. A., McSharry, P. E., Roberts, S. J., Costello, D. A. and Moroz, I. M., *Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection*, BioMedical Engineering OnLine, **6** (2007), No. 23

[11] Martnez-Martnez, J. M., Escandell-Montero, P., Soria-Olivas, E., Martn-Guerrero, J. D., Magdalena-Benedito, R. and Gmez-Sanchis, J., *Regularized extreme learning machine for regression problems*, Neurocomputing, **74** (2011), No. 17, 3716–3721

[12] Nakajo, K., Shimoji, K. and Takahashi, W., *Strong convergence to a common fixed point of families of nonexpansive mappings in banach spaces*, J. Nonlinear Convex Anal., **8** (2007), No. 1, 11–34

[13] Nesterov, N., *A method for solving the convex programming problem with convergence rate $o(1/k^2)$*, Dokl. Akad. Nauk. SSSR, **269** (1983), No. 3, 543–547

[14] Parikh, N. and Boyd, S., *Proximal algorithms*, Found. Trends Optim., **1** (2014), No. 3, 127–239, 2014

[15] Polyak, B., *Some methods of speeding up the convergence of iteration methods*, USSR Computational Mathematics and Mathematical Physics, **4** (1964), No. 5, 1–17

[16] Serre, D., *Matrices: Theory and Applications*, Springer-Verlag New York, New York, 2002

[17] Takahashi, W., *Viscosity approximation methods for countable families of nonexpansive mappings in banach spaces*, Nonlinear Analysis: Theory, Methods & Applications, **70** (2009), No. 2, 719–734

[18] Tan, K. and Xu, H., *Approximating fixed points of nonexpansive mappings by the ishikawa iteration process*, J. Math. Anal. Appl., **178** (1993), No. 2, 301–308

[19] Tibshirani, R., *Regression shrinkage and selection via the lasso*, J. R. Stat. Soc. Ser. B Stat. Methodol., **58** (1996) No. 1, 267–288

[20] Tikhonov, A. N. and Arsenin, V. Y., *Solutions of Ill-posed problems*, Winston, 1977, 258 pp.

[21] Verma, M. and Shukla, K., *A new accelerated proximal gradient technique for regularized multitask learning framework*, Pattern Recognition Letters, **95** (2017), 98–103

[1]GRADUATE PH.D. DEGREE PROGRAM IN MATHEMATICS
FACULTY OF SCIENCE
CHIANG MAI UNIVERSITY
239 HUAYKAEW RD, 50200, CHIANG MAI, THAILAND
E-mail address: lim.bussaban@gmail.com

[2]DEPARTMENT OF MATHEMATICS
DATA SCIENCE RESEARCH CENTER
FACULTY OF SCIENCE
CHIANG MAI UNIVERSITY
239 HUAYKAEW RD, 50200, CHIANG MAI, THAILAND
E-mail address: suthep.s@cmu.ac.th

[3]DEPARTMENT OF MATHEMATICS
CENTER OF EXCELLENCE IN MATHEMATICS AND APPLIES MATHEMATICS
FACULTY OF SCIENCE
CHIANG MAI UNIVERSITY
239 HUAYKAEW RD, 50200, CHIANG MAI, THAILAND
E-mail address: akaewkhao@yahoo.com