Comput Optim Appl
https://doi.org/10.1007/s10589-017-9975-9

Two smooth support vector machines for ε-insensitive regression

Weizhe Gu[1] · **Wei-Po Chen**[2] · **Chun-Hsu Ko**[3] ·
Yuh-Jye Lee[4] · **Jein-Shan Chen**[2]

Received: 6 June 2017

Abstract In this paper, we propose two new smooth support vector machines for ε-insensitive regression. According to these two smooth support vector machines, we construct two systems of smooth equations based on two novel families of smoothing functions, from which we seek the solution to ε-support vector regression (ε-SVR). More specifically, using the proposed smoothing functions, we employ the smoothing Newton method to solve the systems of smooth equations. The algorithm is shown to be globally and quadratically convergent without any additional conditions. Numerical comparisons among different values of parameter are also reported.

J.-S. Chen work is supported by Ministry of Science and Technology, Taiwan.

✉ Jein-Shan Chen
jschen@math.ntnu.edu.tw

Weizhe Gu
weizhegu@yahoo.com.cn

Wei-Po Chen
weaper@gmail.com

Chun-Hsu Ko
chko@isu.edu.tw

Yuh-Jye Lee
yuhjye@math.nctu.edu.tw

[1] Department of Mathematics, School of Science, Tianjin University, Tianjin 300072, People's Republic of China

[2] Department of Mathematics, National Taiwan Normal University, Taipei 11677, Taiwan

[3] Department of Electrical Engineering, I-Shou University, Kaohsiung 840, Taiwan

[4] Department of Applied Mathematics, National Chiao Tung University, Hsinchu 300, Taiwan

Published online: 18 December 2017

Keywords Support vector machine · ε-insensitive loss function · ε-smooth support vector regression · Smoothing Newton algorithm

1 Introduction

Support vector machine (SVM) is a popular and important statistical learning technology [1,7,8,16–19]. Generally speaking, there are two main categories for support vector machines (SVMs): support vector classification (SVC) and support vector regression (SVR). The model produced by SVR depends on a training data set $S = \{(A_1, y_1), \ldots, (A_m, y_m)\} \subseteq \mathbb{R}^n \times \mathbb{R}$, where $A_i \in \mathbb{R}^n$ is the input data and $y_i \in \mathbb{R}$ is called the observation. The main goal of ε-insensitive regression with the idea of SVMs is to find a linear or nonlinear regression function f that has at most ε deviation from the actually obtained targets y_i for all the training data, and at the same time is as flat as possible. This problem is called ε-support vector regression (ε-SVR).

For pedagogical reasons, we begin with the linear case, in which the regression function $f(\varpi)$ is defined as

$$f(\varpi) = \varpi^T x + b \quad \text{with} \quad x \in \mathbb{R}^n, \ b \in \mathbb{R}. \tag{1}$$

Flatness in the case of (1) means that one seeks a small x. One way to ensure this is to minimize the norm of x, then the problem ε-SVR can be formulated as a constrained minimization problem:

$$\begin{array}{ll} \min & \frac{1}{2}x^T x + C \sum_{i=1}^{m}(\xi_i + \xi_i^*) \\ \text{s.t.} & \begin{cases} y_i - A_i^T x - b \le \varepsilon + \xi_i \\ A_i^T x + b - y_i \le \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0, \ i = 1, \ldots, m \end{cases} \end{array} \tag{2}$$

The constant $C > 0$ determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated. This corresponds to dealing with a so called ε-insensitive loss function $|\xi|_\varepsilon$ described by

$$|\xi|_\varepsilon = \max\{0, |\xi| - \varepsilon\}.$$

The formulation (2) is a convex quadratic minimization problem with $n + 1$ free variables, $2m$ nonnegative variables, and $2m$ inequality constraints, which enlarges the problem size and could increase computational complexity.

In fact, the problem (2) can be reformulated as an unconstrained optimization problem:

$$\min_{(x,b)\in\mathbb{R}^{n+1}} \frac{1}{2}\left(x^T x + b^2\right) + \frac{C}{2}\sum_{i=1}^{m}\left|A_i^T x + b - y_i\right|_\varepsilon^2 \tag{3}$$

This formulation has been proposed in active set support vector regression [11] and solved in its dual form. The objective function is strongly convex, hence, the problem has a unique global optimal solution. However, according to the fact that the objective function is not twice continuously differentiable, Newton-type algorithms cannot be applied to solve (3) directly.

Lee, Hsieh and Huang [7] apply a smooth technique for (3). The smooth function

$$f_\varepsilon(x,\alpha) = x + \frac{1}{\alpha}\log(1+e^{-\alpha x}), \tag{4}$$

which is the integral of the sigmoid function $\frac{1}{1+e^{-\alpha x}}$, is used to smooth the plus function $[x]_+$. More specifically, the smooth function $f_\varepsilon(x,\alpha)$ approaches to $[x]_+$, when α goes to infinity. Then, the problem (3) is recast to a strongly convex unconstrained minimization problem with the smooth function $f_\varepsilon(x,\alpha)$ and a Newton-Armijo algorithm is proposed to solve it. It is proved that when the smoothing parameter α approaches to infinity, the unique solution of the reformulated problem converges to the unique solution of the original problem [7, Theorem 2.2]. However, the smoothing parameter α is fixed in the proposed algorithm, and in the implementation of this algorithm, α cannot be set large enough.

In this paper, we introduce two smooth support vector machines for ε-insensitive regression. For the first smooth support vector machine, we reformulated ε-SVR to a strongly convex unconstrained optimization problem with one type of smoothing functions $\phi_\varepsilon(x,\alpha)$. Then, we define a new function H_ϕ, which corresponds to the optimality condition of the unconstrained optimization problem. From the solution of $H_\phi(z) = 0$, we can obtain the solution of ε-SVR. For the second smooth support vector machine, we smooth the optimality condition of the strongly convex unconstrained optimization problem of (3) with another type of smooth functions $\psi_\varepsilon(x,\alpha)$. Accordingly we define the function H_ψ, which also possesses the same properties as H_ϕ does. For either $H_\phi = 0$ or $H_\psi = 0$, we consider the smoothing Newton method to solve it. The algorithm is shown to be globally convergent, specifically, the iterative sequence converges to the unique solution to (3). Furthermore, the algorithm is shown to be locally quadratically convergent without any assumptions.

The paper is organized as follows. In Sects. 2 and 3, we introduce two smooth support vector machine reformulations by two types of smoothing functions. In Sect. 4, we propose a smoothing Newton algorithm and study its global and local quadratic convergence. Numerical results and comparisons are reported in Sect. 5. Throughout this paper, $\mathcal{K} := \{1, 2, \ldots\}$, all vectors will be column vectors. For a given vector $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, the plus function $[x]_+$ is defined as

$$([x]_+)_i = \max\{0, x_i\}, \quad i = 1, \ldots, n.$$

For a differentiable function f, we denote by $\nabla f(x)$ and $\nabla^2 f(x)$ the gradient and the Hessian matrix of f at x, respectively. For a differentiable mapping $G : \mathbb{R}^n \to \mathbb{R}^m$, we denote by $G'(x) \in \mathbb{R}^{m\times n}$ the Jacabian of G at x. For a matrix $A \in \mathbb{R}^{m\times n}$, A_i^T is the i-th row of A. A column vector of ones and identity matrix of arbitrary dimension will be denoted by $\mathbf{1}$ and I, respectively. We denote the sign function by

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if } x > 0, \\ [-1, 1] & \text{if } x = 0, \\ -1 & \text{if } x < 0. \end{cases}$$

2 The first smooth support vector machine

As mentioned in [7], it is known that ε-SVR can be reformulated as a strongly convex unconstrained optimization problem (3). Denote $\omega := (x, b) \in \mathbb{R}^{n+1}$, $\bar{A} := (A, \mathbf{1})$ and $\bar{A}_i^T$ is the i-th row of $\bar{A}$, then the smooth support vector regression (3) can be rewritten as

$$\min_{\omega} \frac{1}{2}\omega^T \omega + \frac{C}{2}\sum_{i=1}^{m} \left|\bar{A}_i^T \omega - y_i\right|_{\varepsilon}^2. \tag{5}$$

Note that $|\cdot|_{\varepsilon}^2$ is smooth, but not twice differentiable, which means the objective function is not twice continuously differentiable. Hence, the Newton-type method cannot be applied to solve (5) directly.

In view of this fact, we propose a family of twice continuously differentiable functions $\phi_{\varepsilon}(x, \alpha)$ to replace $|x|_{\varepsilon}^2$. The family of functions $\phi_{\varepsilon}(x, \alpha) : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ is given by

$$\phi_{\varepsilon}(x, \alpha) = \begin{cases} (|x| - \varepsilon)^2 + \frac{1}{3}\alpha^2 & \text{if } |x| - \varepsilon \geq \alpha, \\ \frac{1}{6\alpha}(|x| - \varepsilon + \alpha)^3 & \text{if } ||x| - \varepsilon| < \alpha, \\ 0 & \text{if } |x| - \varepsilon \leq -\alpha, \end{cases} \tag{6}$$

where $0 < \alpha < \varepsilon$ is a smooth parameter. The graphs of $\phi_{\varepsilon}(x, \alpha)$ are depicted in Fig. 1. From this geometric view, it is clear to see that $\phi_{\varepsilon}(x, \alpha)$ is a class of smoothing functions for $|x|_{\varepsilon}^2$.

Besides the geometric approach, we hereat show that $\phi_{\varepsilon}(x, \alpha)$ is a class of smoothing functions for $|x|_{\varepsilon}^2$ by algebraic verification. To this end, we compute the partial derivatives of $\phi_{\varepsilon}(x, \alpha)$ as below:

$$\nabla_x \phi_{\varepsilon}(x, \alpha) = \begin{cases} 2(|x| - \varepsilon)\operatorname{sgn}(x) & \text{if } |x| - \varepsilon \geq \alpha, \\ \frac{1}{2\alpha}(|x| - \varepsilon + \alpha)^2 \operatorname{sgn}(x) & \text{if } \left||x| - \varepsilon\right| < \alpha, \\ 0 & \text{if } |x| - \varepsilon \leq -\alpha. \end{cases} \tag{7}$$

$$\nabla_{xx}^2 \phi_{\varepsilon}(x, \alpha) = \begin{cases} 2 & \text{if } |x| - \varepsilon \geq \alpha \\ \frac{|x| - \varepsilon + \alpha}{\alpha} & \text{if } \left||x| - \varepsilon\right| < \alpha, \\ 0 & \text{if } |x| - \varepsilon \leq -\alpha. \end{cases} \tag{8}$$

$$\nabla_{x\alpha}^2 \phi_{\varepsilon}(x, \alpha) = \begin{cases} 0 & \text{if } |x| - \varepsilon \geq \alpha, \\ \frac{(|x| - \varepsilon + \alpha)(\alpha - |x| + \varepsilon)}{2\alpha^2}\operatorname{sgn}(x) & \text{if } \left||x| - \varepsilon\right| < \alpha, \\ 0 & \text{if } |x| - \varepsilon \leq -\alpha. \end{cases} \tag{9}$$

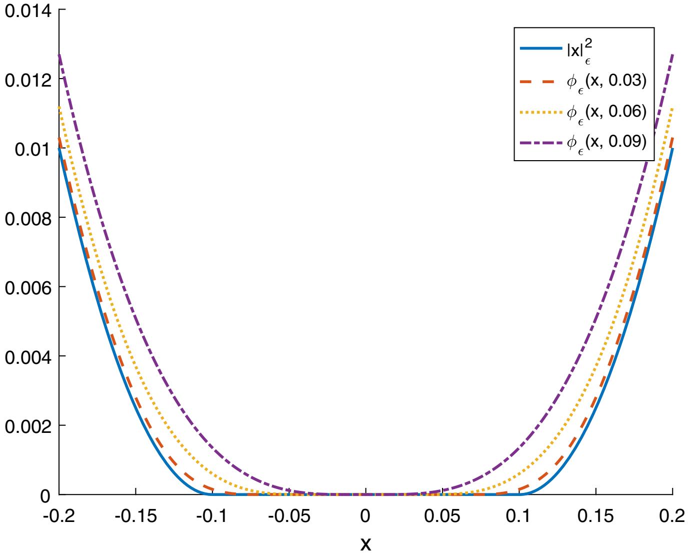

Fig. 1 Graphs of $\phi_\varepsilon(x, \alpha)$ with $\varepsilon = 0.1$ and $\alpha = 0.03, 0.06, 0.09$

With the above, the following lemma shows some basic properties of $\phi_\varepsilon(x, \alpha)$.

Lemma 2.1 *Let $\phi_\varepsilon(x, \alpha)$ be defined as in (6). Then, the following hold.*

(a) *For $0 < \alpha < \varepsilon$, there holds $0 \le \phi_\varepsilon(x, \alpha) - |x|^2_\varepsilon \le \frac{1}{3}\alpha^2$.*
(b) *The function $\phi_\varepsilon(x, \alpha)$ is twice continuously differentiable with respect to x for $0 < \alpha < \varepsilon$.*
(c) $\lim\limits_{\alpha\to 0} \phi_\varepsilon(x, \alpha) = |x|^2_\varepsilon$ *and* $\lim\limits_{\alpha\to 0} \nabla_x\phi_\varepsilon(x, \alpha) = \nabla(|x|^2_\varepsilon)$.

Proof (a) To complete the arguments, we need to discuss four cases.
(i) For $|x| - \varepsilon \ge \alpha$, it is clear that $\phi_\varepsilon(x, \alpha) - |x|^2_\varepsilon = \frac{1}{3}\alpha^2$.
(ii) For $0 < |x| - \varepsilon < \alpha$, i.e., $0 < x - \varepsilon < \alpha$ or $0 < -x - \varepsilon < \alpha$, there have two subcase.
If $0 < x - \varepsilon < \alpha$, letting $f(x) := \phi_\varepsilon(x, \alpha) - |x|^2_\varepsilon = \frac{1}{6\alpha}(x - \varepsilon + \alpha)^3 - (x - \varepsilon)^2$ gives

$$\begin{cases} f'(x) = \frac{(x-\varepsilon+\alpha)^2}{2\alpha} - 2(x - \varepsilon), & \forall x \in (\varepsilon, \varepsilon + \alpha), \\ f''(x) = \frac{x-\varepsilon+\alpha}{\alpha} - 2 < 0, & \forall x \in (\varepsilon, \varepsilon + \alpha). \end{cases}$$

This indicates that $f'(x)$ is monotone decreasing on $(\varepsilon, \varepsilon + \alpha)$, which further implies

$$f'(x) \ge f'(\varepsilon + \alpha) = 0, \quad \forall x \in (\varepsilon, \varepsilon + \alpha).$$

Thus, we obtain that $f(x)$ is monotone increasing on $(\varepsilon, \varepsilon + \alpha)$. With this, we have $f(x) \le f(\varepsilon + \alpha) = \frac{1}{3}\alpha^2$, which yields

$$\phi_\varepsilon(x, \alpha) - |x|^2_\varepsilon \le \frac{1}{3}\alpha^2, \quad \forall x \in (\varepsilon, \varepsilon + \alpha).$$

If $0 < -x - \varepsilon < \alpha$, the arguments are similar as above, and we omit them.
(iii) For $-\alpha < |x| - \varepsilon \leq 0$, it is clear that $\phi_\varepsilon(x, \alpha) - |x|_\varepsilon^2 = \frac{1}{6\alpha}(|x| - \varepsilon + \alpha)^3 \leq \frac{\alpha^3}{6\alpha} \leq \frac{\alpha^2}{3}$.
(iv) For $|x| - \varepsilon \leq -\alpha$, we have $\phi_\varepsilon(x, \alpha) - |x|_\varepsilon^2 = 0$. Then, the desired result follows.
(b) To prove the twice continuous differentiability of $\phi_\varepsilon(x, \alpha)$, we need to check $\phi_\varepsilon(\cdot, \alpha)$, $\nabla_x \phi_\varepsilon(\cdot, \alpha)$ and $\nabla_{xx}^2 \phi_\varepsilon(\cdot, \alpha)$ are all continuous. Since they are piecewise functions, it suffices to check the junction points.

First, we check that $\phi_\varepsilon(\cdot, \alpha)$ is continuous.
(i) If $|x| - \varepsilon = \alpha$, then $\phi_\varepsilon(x, \alpha) = \frac{4}{3}\alpha^2$, which implies $\phi_\varepsilon(\cdot, \alpha)$ is continuous.
(ii) If $|x| - \varepsilon = -\alpha$, then $\phi_\varepsilon(x, \alpha) = 0$. Hence, $\phi_\varepsilon(\cdot, \alpha)$ is continuous.
Next, we check $\nabla_x \phi_\varepsilon(\cdot, \alpha)$ is continuous.
(i) If $|x| - \varepsilon = \alpha$, then $\nabla_x \phi_\varepsilon(x, \alpha) = 2\alpha \operatorname{sgn}(x)$.
(ii) If $|x| - \varepsilon = -\alpha$, then $\nabla_x \phi_\varepsilon(x, \alpha) = 0$. From the above, it clear to see that $\nabla_x \phi_\varepsilon(\cdot, \alpha)$ is continuous.
Now we show that $\nabla_{xx}^2 \phi_\varepsilon(\cdot, \alpha)$ is continuous.
(i) If $|x| - \varepsilon = \alpha$, $\nabla_{xx}^2 \phi_\varepsilon(x, \alpha) = 2$.
(ii) $|x| - \varepsilon = -\alpha$ then $\nabla_{xx}^2 \phi_\varepsilon(x, \alpha) = 0$. Hence, $\nabla_{xx}^2 \phi_\varepsilon(\cdot, \alpha)$ is continuous.
(c) It is clear that $\lim\limits_{\alpha \to 0} \phi_\varepsilon(x, \alpha) = |x|_\varepsilon^2$ holds by part(a). It remains to verify $\lim\limits_{\alpha \to 0} \nabla_x \phi_\varepsilon(x, \alpha) = \nabla(|x|_\varepsilon^2)$. First, we compute that

$$\nabla(|x|_\varepsilon^2) = \begin{cases} 2(|x| - \varepsilon)\operatorname{sgn}(x) & \text{if } |x| - \varepsilon \geq 0, \\ 0 & \text{if } |x| - \varepsilon < 0. \end{cases} \tag{10}$$

In light of (10), we proceed the arguments by discussing four cases.
(i) For $|x| - \varepsilon \geq \alpha$, we have $\nabla_x \phi_\varepsilon(x, \alpha) - \nabla(|x|_\varepsilon^2) = 0$. Then, the desired result follows.
(ii) For $0 < |x| - \varepsilon < \alpha$, we have

$$\nabla_x \phi_\varepsilon(x, \alpha) - \nabla(|x|_\varepsilon^2) = \frac{1}{2\alpha}(|x| - \varepsilon + \alpha)^2 \operatorname{sgn}(x) - 2(|x| - \varepsilon)\operatorname{sgn}(x)$$

which yields

$$\lim_{\alpha \to 0}(\nabla_x \phi_\varepsilon(x, \alpha) - \nabla(|x|_\varepsilon^2)) = \lim_{\alpha \to 0} \frac{(|x| - \varepsilon + \alpha)^2 - 4\alpha(|x| - \varepsilon)}{2\alpha} \operatorname{sgn}(x).$$

We notice that $|x| \to \varepsilon$ when $\alpha \to 0$, and hence $\frac{(|x| - \varepsilon + \alpha)^2 - 4\alpha(|x| - \varepsilon)}{2\alpha} \to \frac{0}{0}$. Then, applying L'hopital rule yields

$$\lim_{\alpha \to 0} \frac{(|x| - \varepsilon + \alpha)^2 - 4\alpha(|x| - \varepsilon)}{2\alpha} = \lim_{\alpha \to 0}(\alpha - (|x| - \varepsilon)) = 0.$$

This implies $\lim\limits_{\alpha \to 0}(\nabla_x \phi_\varepsilon(x, \alpha) - \nabla(|x|_\varepsilon^2)) = 0$, which is the desired result.

(iii) For $-\alpha < |x| - \varepsilon \le 0$, we have $\nabla_x \phi_\varepsilon(x, \alpha) - \nabla(|x|_\varepsilon^2) = \frac{1}{2\alpha}(|x| - \varepsilon + \alpha)^2 \mathrm{sgn}(x)$. Then, applying L'hopital rule gives

$$\lim_{\alpha \to 0} \frac{(|x| - \varepsilon + \alpha)^2}{2\alpha} = \lim_{\alpha \to 0} (|x| - \varepsilon + \alpha) = 0.$$

Thus, we prove that $\lim\limits_{\alpha \to 0} (\nabla_x \phi_\varepsilon(x, \alpha) - \nabla(|x|_\varepsilon^2)) = 0$ under this case.

(iv) For $|x| - \varepsilon \le -\alpha$, we have $\nabla_x \phi_\varepsilon(x, \alpha) - \nabla(|x|_\varepsilon^2) = 0$. Then, the desired result follows clearly. □

Now, we use the family of smoothing functions ϕ_ε to replace the square of ε-insensitive loss function in (5) to obtain the first smooth support vector regression. In other words, we consider

$$\min_{\omega} F_{\varepsilon,\alpha}(\omega) := \frac{1}{2}\omega^T \omega + \frac{C}{2}\mathbf{1}^T \Phi_\varepsilon \left(\bar{A}\omega - y, \alpha\right). \tag{11}$$

where $\omega := (x, b) \in \mathbb{R}^{n+1}$, and $\Phi_\varepsilon (Ax + \mathbf{1}b - y, \alpha) \in \mathbb{R}^m$ is defined by

$$\Phi_\varepsilon(Ax + \mathbf{1}b - y, \alpha)_i = \phi_\varepsilon (A_i x + b - y_i, \alpha).$$

This is a strongly convex unconstrained optimization with the twice continuously differentiable objective function. Noting $\lim\limits_{\alpha \to 0} \phi_\varepsilon(x, \alpha) = |x|_\varepsilon^2$, we see that

$$\min_{\omega} F_{\varepsilon,0}(\omega) := \lim_{\alpha \to 0} F_{\varepsilon,\alpha}(\omega) = \frac{1}{2}\omega^T \omega + \frac{C}{2}\sum_{i=1}^{m} \left|\bar{A}_i^T \omega - y_i\right|_\varepsilon^2 \tag{12}$$

which is exactly the problem (5).

The following Theorem shows that the unique solution of the smooth problem (11) approaches to the unique solution of the problem (12) as $\alpha \to 0$. Indeed, it plays as the same role as [7, Theorem 2.2].

Theorem 2.1 *Let $F_{\varepsilon,\alpha}(\omega)$ and $F_{\varepsilon,0}(\omega)$ be defined as in (11) and (12), respectively. Then, the following hold.*

(a) *There exists a unique solution $\bar{\omega}_\alpha$ to $\min\limits_{\omega \in \mathbb{R}^{n+1}} F_{\varepsilon,\alpha}(\omega)$ and a unique solution $\bar{\omega}$ to $\min\limits_{\omega \in \mathbb{R}^{n+1}} F_{\varepsilon,0}(\omega)$.*

(b) *For all $0 < \alpha < \varepsilon$, we have the following inequality:*

$$\|\bar{\omega}_\alpha - \bar{\omega}\|^2 \le \frac{1}{6} C m \alpha^2. \tag{13}$$

Moreover, $\bar{\omega}_\alpha$ converges to $\bar{\omega}$ as $\alpha \to 0$ with an upper bound given by (13).

Proof (a) In view of $\phi_\varepsilon(x,\alpha) - |x|_\varepsilon^2 \geq 0$ in Lemma 2.1(a), we see that the level sets

$$L_v(F_{\varepsilon,\alpha}(\omega)) := \left\{\omega \in \mathbb{R}^{n+1} \mid F_{\varepsilon,\alpha}(\omega) \leq v\right\}$$
$$L_v(F_{\varepsilon,0}(\omega)) := \left\{\omega \in \mathbb{R}^{n+1} \mid F_{\varepsilon,0}(\omega) \leq v\right\}$$

satisfy

$$L_v(F_{\varepsilon,\alpha}(\omega)) \subseteq L_v(F_{\varepsilon,0}(\omega)) \subseteq \left\{\omega \in \mathbb{R}^{n+1} \mid \omega^T \omega \leq 2v\right\} \tag{14}$$

for any $v \geq 0$. Hence, we obtain that $L_v(F_{\varepsilon,\alpha}(\omega))$ and $L_v(F_{\varepsilon,0}(\omega))$ are compact (closed and bounded) subsets in $\mathbb{R}^{n+1}$. Then, by the strong convexity of $F_{\varepsilon,0}(\omega)$ and $F_{\varepsilon,\alpha}(\omega)$ with $\alpha > 0$, each of the problems $\min\limits_{\omega\in\mathbb{R}^{n+1}} F_{\varepsilon,\alpha}(\omega)$ and $\min\limits_{\omega\in\mathbb{R}^{n+1}} F_{\varepsilon,0}(\omega)$ has a unique solution.

(b) From the optimality condition and strong convexity of $F_{\varepsilon,0}(\omega)$ and $F_{\varepsilon,\alpha}(\omega)$ with $\alpha > 0$, we know that

$$F_{\varepsilon,0}(\bar{\omega}_\alpha) - F_{\varepsilon,0}(\bar{\omega}) \geq \nabla F_{\varepsilon,0}(\bar{\omega}_\alpha - \bar{\omega}) + \frac{1}{2}\|\bar{\omega}_\alpha - \bar{\omega}\|^2 \geq \frac{1}{2}\|\bar{\omega}_\alpha - \bar{\omega}\|^2, \tag{15}$$
$$F_{\varepsilon,\alpha}(\bar{\omega}) - F_{\varepsilon,\alpha}(\bar{\omega}_\alpha) \geq \nabla F_{\varepsilon,\alpha}(\bar{\omega} - \bar{\omega}_\alpha) + \frac{1}{2}\|\bar{\omega} - \bar{\omega}_\alpha\|^2 \geq \frac{1}{2}\|\bar{\omega} - \bar{\omega}_\alpha\|^2. \tag{16}$$

Note that $F_{\varepsilon,\alpha}(\omega) \geq F_{\varepsilon,0}(\omega)$ because $\phi_\varepsilon(x,\alpha) - |x|_\varepsilon^2 \geq 0$. Then, adding up (15) and (16) along with this fact yield

$$\begin{aligned}
\|\bar{\omega}_\alpha - \bar{\omega}\|^2 &\leq (F_{\varepsilon,\alpha}(\bar{\omega}) - F_{\varepsilon,0}(\bar{\omega})) - (F_{\varepsilon,\alpha}(\bar{\omega}_\alpha) - F_{\varepsilon,0}(\bar{\omega}_\alpha)) \\
&\leq F_{\varepsilon,\alpha}(\bar{\omega}) - F_{\varepsilon,0}(\bar{\omega}) \\
&= \frac{C}{2} 1^T \Phi_\varepsilon(\bar{A}\bar{\omega} - y, \alpha) - \frac{C}{2}\sum_{i=1}^{m} \left|\bar{A}_i^T \bar{\omega} - y_i\right|_\varepsilon^2 \\
&= \frac{C}{2}\sum_{i=1}^{m} \phi_\varepsilon(\bar{A}_i\bar{\omega} - y_i, \alpha) - \frac{C}{2}\sum_{i=1}^{m} \left|\bar{A}_i^T \bar{\omega} - y_i\right|_\varepsilon^2 \\
&\leq \frac{1}{6} C m \alpha^2,
\end{aligned}$$

where the last inequality is due to Lemma 2.1(a). It is clear that $\bar{\omega}_\alpha$ converges to $\bar{\omega}$ as $\alpha \to 0$ with an upper bound given by the above. Then, the proof is complete. □

Next, we focus on the optimality condition of the minimization problem (11), which is indeed sufficient and necessary for (11) and has the form of

$$\nabla_\omega F_{\varepsilon,\alpha}(\omega) = 0.$$

With this, we define a function $H_\phi : \mathbb{R}^{n+2} \to \mathbb{R}^{n+2}$ by

$$H_\phi(z) = \begin{bmatrix} \alpha \\ \nabla_\omega F_{\varepsilon,\alpha}(\omega) \end{bmatrix} = \begin{bmatrix} \alpha \\ \omega + C\sum_{i=1}^m \nabla_x \phi_\varepsilon(\bar{A}_i^T\omega - y_i, \alpha)\bar{A}_i \end{bmatrix} \tag{17}$$

where $z := (\alpha, \omega) \in \mathbb{R}^{n+2}$. From Lemma 2.1 and the strong convexity of $F_{\varepsilon,\alpha}(\omega)$, it is easy to see that if $H_\phi(z) = 0$, then $\alpha = 0$ and ω solves (11); and for any $z \in \mathbb{R}_{++} \times \mathbb{R}^{n+1}$, the function H_ϕ is continuously differentiable. In addition, the Jacobian of H_ϕ can be calculated as below:

$$H'_\phi(z) = \begin{bmatrix} 1 & 0 \\ \nabla^2_{\omega\alpha}F_{\varepsilon,\alpha}(\omega) & \nabla^2_{\omega\omega}F_{\varepsilon,\alpha}(\omega) \end{bmatrix} \tag{18}$$

where

$$\nabla^2_{\omega\alpha}F_{\varepsilon,\alpha}(\omega) = C\sum_{i=1}^m \nabla^2_{x\alpha}\phi_\varepsilon\left(\bar{A}_i^T\omega - y_i, \alpha\right)\bar{A}_i,$$
$$\nabla^2_{\omega\omega}F_{\varepsilon,\alpha}(\omega) = I + C\sum_{i=1}^m \nabla^2_{xx}\phi_\varepsilon\left(\bar{A}_i^T\omega - y_i, \alpha\right)\bar{A}_i\bar{A}_i^T.$$

From (8), we can see $\nabla^2_{xx}\phi_\varepsilon(x,\alpha) \geq 0$, which implies $C\sum_{i=1}^m \nabla^2_{xx}\phi_\varepsilon(\bar{A}_i^T\omega - y_i, \alpha)\bar{A}_i\bar{A}_i^T$ is positive semidefinite. Hence, $\nabla^2_{\omega\omega}F_{\varepsilon,\alpha}(\omega)$ is positive definite. This helps us to prove that $H'_\phi(z)$ is invertible at any $z \in \mathbb{R}_{++} \times \mathbb{R}^{n+1}$. In fact, if there exists a vector $d := (d_1, d_2) \in \mathbb{R} \times \mathbb{R}^{n+1}$ such that $H'_\phi(z)d = 0$, then we have

$$\begin{bmatrix} d_1 \\ d_1\nabla^2_{\omega\alpha}F_{\varepsilon,\alpha}(\omega) + \nabla^2_{\omega\omega}F_{\varepsilon,\alpha}(\omega)d_2 \end{bmatrix} = 0.$$

This implies that $d = 0$, and hence $H'_\phi(z)$ is invertible at any $z \in \mathbb{R}_{++} \times \mathbb{R}^{n+1}$.

3 The second smooth support vector machine

In this section, we consider another type of smoothing functions $\psi_{\varepsilon,p}(x,\alpha) : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$, which is defined by

$$\psi_{\varepsilon,p}(x,\alpha) = \begin{cases} 0 & \text{if } 0 \leq |x| \leq \varepsilon - \alpha, \\ \frac{\alpha}{p-1}\left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^p & \text{if } \varepsilon - \alpha < |x| < \varepsilon + \frac{\alpha}{p-1}, \\ |x| - \varepsilon & \text{if } |x| \geq \varepsilon + \frac{\alpha}{p-1}. \end{cases} \tag{19}$$

here $p \geq 2$. The graphs of $\psi_{\varepsilon,p}(x,\alpha)$ are depicted in Fig. 2, which clearly verify that $\psi_{\varepsilon,p}(x,\alpha)$ is a family of smoothing functions for $|x|_\varepsilon$.

As in Lemma 3.1, we verify that $\psi_{\varepsilon,p}(x,\alpha)$ is a family of smoothing functions for $|x|_\varepsilon$, hence, $\psi^2_{\varepsilon,p}(x,\alpha)$ is also a family of smoothing functions for $|x|^2_\varepsilon$. Then, we can

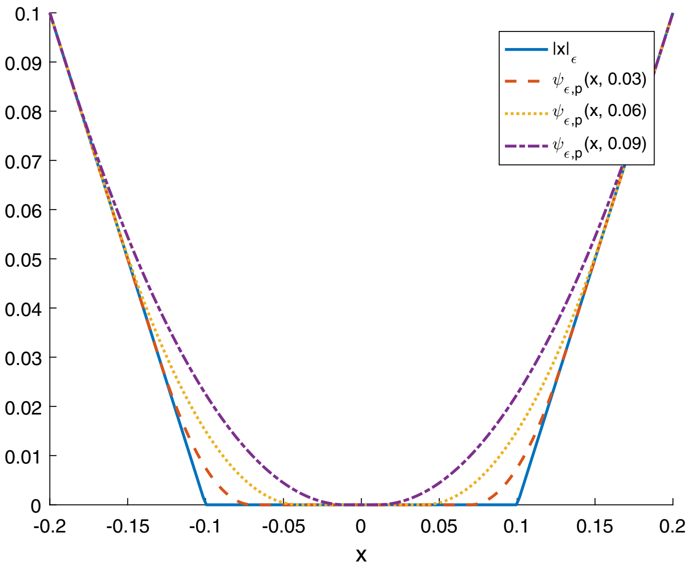

Fig. 2 Graphs of $\psi_{\varepsilon,p}(x,\alpha)$ with $\varepsilon = 0.1$, $\alpha = 0.03, 0.06, 0.09$ and $p = 2$

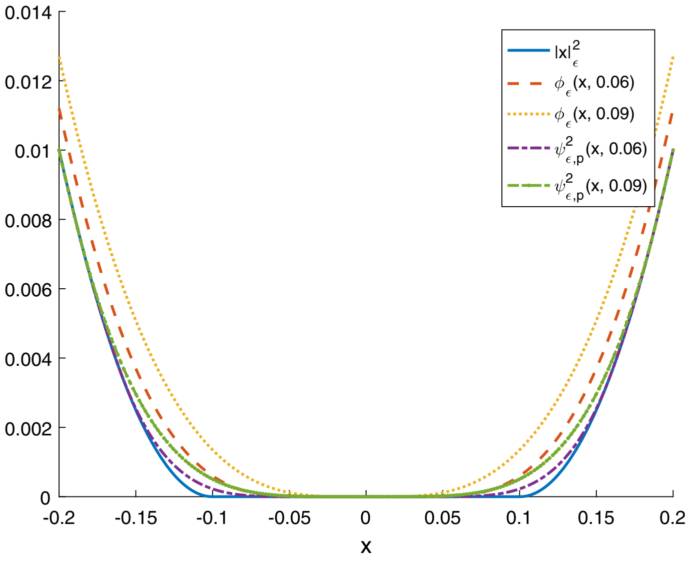

Fig. 3 Graphs of $|x|^2_\varepsilon$, $\phi_\varepsilon(x,\alpha)$ and $\psi^2_{\varepsilon,p}(x,\alpha)$ with $\varepsilon = 0.1$, $\alpha = 0.06, 0.09$ and $p = 2$

employ $\psi^2_{\varepsilon,p}$ to replace the square of ε-insensitive loss function in (5) as the same way done in Sect. 2. The graphs of $\psi^2_{\varepsilon,p}(x,\alpha)$ with comparison to $\phi_\varepsilon(x,\alpha)$ are depicted in Fig. 3. In fact, there is a relation between $\psi^2_{\varepsilon,p}(x,\alpha)$ and $\phi_\varepsilon(x,\alpha)$ shown as in Proposition 3.1.

In other words, we obtain an alternative strongly convex unconstrained optimization for (5):

$$\min_{\omega} \frac{1}{2}\omega^T \omega + \frac{C}{2}\sum_{i=1}^{m} \psi_{\varepsilon,p}^2\left(\bar{A}_i^T \omega - y_i, \alpha\right). \tag{20}$$

However, the smooth function $\psi_{\varepsilon,p}^2(x,\alpha)$ is not twice differentiable with respect x, and hence the objective function of (20) is not twice differentiable although it smooth. Then, we still cannot apply Newton-type method to solve (20). To conquer this, we take another smoothing technique. Before presenting the idea of this smoothing technique, the following two lemmas regarding properties of $\psi_{\varepsilon,p}(x,\alpha)$ are needed. To this end, we also compute the partial derivative of $\psi_{\varepsilon,p}(x,\alpha)$ as below:

$$\nabla_x \psi_{\varepsilon,p}(x,\alpha) = \begin{cases} 0 & \text{if } 0 \le |x| \le \varepsilon - \alpha, \\ \operatorname{sgn}(x)\left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^{p-1} & \text{if } \varepsilon - \alpha < |x| < \varepsilon + \frac{\alpha}{p-1}, \\ \operatorname{sgn}(x) & \text{if } |x| \ge \varepsilon + \frac{\alpha}{p-1}. \end{cases}$$

$$\nabla_\alpha \psi_{\varepsilon,p}(x,\alpha) = \begin{cases} 0 & \text{if } 0 \le |x| \le \varepsilon - \alpha, \\ \frac{(\varepsilon-|x|)(p-1)+\alpha}{p\alpha}\left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^{p-1} & \text{if } \varepsilon - \alpha < |x| < \varepsilon + \frac{\alpha}{p-1}, \\ 0 & \text{if } |x| \ge \varepsilon + \frac{\alpha}{p-1}. \end{cases}$$

Lemma 3.1 *Let $\psi_{\varepsilon,p}(x,\alpha)$ be defined as in (19). Then, we have*

(a) *$\psi_{\varepsilon,p}(x,\alpha)$ is smooth with respect to x for any $p \ge 2$;*
(b) *$\lim_{\alpha\to 0} \psi_{\varepsilon,p}(x,\alpha) = |x|_\varepsilon$ for any $p \ge 2$.*

Proof (a) To prove the result, we need to check both $\psi_{\varepsilon,p}(\cdot,\alpha)$ and $\nabla_x\psi_{\varepsilon,p}(\cdot,\alpha)$ are continuous.
(i) If $|x| = \varepsilon - \alpha$, then $\psi_{\varepsilon,p}(x,\alpha) = 0$.
(ii) If $|x| = \varepsilon + \frac{\alpha}{p-1}$, then $\psi_{\varepsilon,p}(x,\alpha) = \frac{\alpha}{p-1}$. Form (i) and (ii), it is clear to see $\psi_{\varepsilon,p}(\cdot,\alpha)$ is continuous.
Moreover, (i) If $|x| = \varepsilon - \alpha$, then $\nabla_x\psi_{\varepsilon,p}(x,\alpha) = 0$.
(ii) If $|x| = \varepsilon + \frac{\alpha}{p-1}$, then $\nabla_x\psi_{\varepsilon,p}(x,\alpha) = \operatorname{sgn}(x)$. In view of (i) and (ii), we see that $\nabla_x\psi_{\varepsilon,p}(\cdot,\alpha)$ is continuous.
(b) To proceed, we discuss four cases.
(1) If $0 \le |x| \le \varepsilon - \alpha$, then $\psi_{\varepsilon,p}(x,\alpha) - |x|_\varepsilon = 0$. Then, the desired result follows.
(2) If $\varepsilon - \alpha \le |x| \le \varepsilon$, then $\psi_{\varepsilon,p}(x,\alpha) - |x|_\varepsilon = \frac{\alpha}{p-1}\left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^p$. Hence,

$$\begin{aligned} &\lim_{\alpha\to 0}\left(\psi_{\varepsilon,p}(x,\alpha) - |x|_\varepsilon\right) \\ &\quad = \lim_{\alpha\to 0}\left(\frac{\alpha}{p-1}\right)\left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^p \\ &\quad = \lim_{\alpha\to 0}\left(\frac{\alpha}{p-1}\right)\lim_{\alpha\to 0}\left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^p. \end{aligned}$$

It is clear that the first limit is zero, so we only need to show that the second limit is bounded. To this end, we rewrite it as

$$\lim_{\alpha\to 0}\left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^p=\lim_{\alpha\to 0}\left(\frac{p-1}{p}\right)^p\left[\frac{|x|-\varepsilon+\alpha}{\alpha}\right]^p.$$

We notice that $|x|\to\varepsilon$ when $\alpha\to 0$ so that $\frac{|x|-\varepsilon+\alpha}{\alpha}\to\frac{0}{0}$. Therefore, by applying L'hopital's rule, we obtain

$$\lim_{\alpha\to 0}\left[\frac{|x|-\varepsilon+\alpha}{\alpha}\right]=1$$

which implies that $\lim_{\alpha\to 0}\left(\psi_{\varepsilon,p}(x,\alpha)-|x|_\varepsilon\right)=0$ under this case.
(3) If $\varepsilon\le|x|\le\varepsilon+\frac{\alpha}{p-1}$, then

$$\psi_{\varepsilon,p}(x,\alpha)-|x|_\varepsilon=\frac{\alpha}{p-1}\left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^p-(|x|-\varepsilon).$$

We have shown in case (2) that

$$\lim_{\alpha\to 0}\frac{\alpha}{p-1}\left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^p=0.$$

It is also obvious that $\lim_{\alpha\to 0}(|x|-\varepsilon)=0$. Hence, we obtain $\lim_{\alpha\to 0}\big(\psi_{\varepsilon,p}(x,\alpha)-|x|_\varepsilon\big)=0$ under this case.
(4) If $|x|\ge\varepsilon+\frac{\alpha}{p-1}$, the desired result follows since it is clear that $\psi_{\varepsilon,p}(x,\alpha)-|x|_\varepsilon=0$. From all the above, the proof is complete. □

Lemma 3.2 *Let $\psi_{\varepsilon,p}(x,\alpha)$ be defined as in (19). Then, we have*

(a) $\psi_{\varepsilon,p}(x,\alpha)\mathrm{sgn}(x)$ *is smooth with respect to x for any* $p\ge 2$*;*
(b) $\lim\limits_{\alpha\to 0}\psi_{\varepsilon,p}(x,\alpha)\mathrm{sgn}(x)=|x|_\varepsilon\mathrm{sgn}(x)$ *for any* $p\ge 2$.

Proof (a) First, we observe that $\psi_{\varepsilon,p}(x,\alpha)\mathrm{sgn}(x)$ can be written as

$$\psi_{\varepsilon,p}(x,\alpha)\mathrm{sgn}(x)=\begin{cases}0 & \text{if } 0\le|x|\le\varepsilon-\alpha,\\ \frac{\alpha}{p-1}\left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^p\mathrm{sgn}(x) & \text{if } \varepsilon-\alpha<|x|<\varepsilon+\frac{\alpha}{p-1},\\ (|x|-\varepsilon)\mathrm{sgn}(x) & \text{if } |x|\ge\varepsilon+\frac{\alpha}{p-1}.\end{cases}$$

Note that $\mathrm{sgn}(x)$ is continuous at $x\neq 0$ and $\psi_{\varepsilon,p}(x,\alpha)=0$ at $x=0$, then applying Lemma 3.1(a) yields $\psi_{\varepsilon,p}(x,\alpha)\mathrm{sgn}(x)$ is continuous. Furthermore, by simple calculations, we have

$$\begin{aligned}\nabla_x(\psi_{\varepsilon,p}(x,\alpha)\mathrm{sgn}(x)) &= \nabla_x\psi_{\varepsilon,p}(x,\alpha)\mathrm{sgn}(x)\\ &= \begin{cases}0 & \text{if } 0 \le |x| \le \varepsilon-\alpha,\\ \left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^{p-1} & \text{if } \varepsilon-\alpha < |x| < \varepsilon+\frac{\alpha}{p-1},\\ 1 & \text{if } |x| \ge \varepsilon+\frac{\alpha}{p-1}.\end{cases}\end{aligned} \tag{21}$$

Mimicking the arguments as in Lemma 3.1(a), we can verify that $\nabla_x(\psi_{\varepsilon,p}(x,\alpha)\mathrm{sgn}(x))$ is continuous. Thus, the desired result follows.
(b) By Lemma 3.1(b), it is easy to see that $\lim\limits_{\alpha\to 0}\psi_{\varepsilon,p}(x,\alpha)\mathrm{sgn}(x) = |x|_\varepsilon\mathrm{sgn}(x)$. Then, the desired result follows. □

Note that $|x|_\varepsilon^2$ is smooth with

$$\nabla(|x|_\varepsilon^2) = 2|x|_\varepsilon\mathrm{sgn}(x) = \begin{cases}2(|x|-\varepsilon)\mathrm{sgn}(x) & \text{if } |x| > \varepsilon,\\ 0 & \text{if } |x| \le \varepsilon.\end{cases}$$

being continuous (but not differentiable). Then, we consider the optimality condition of (12), that is

$$\nabla_\omega F_{\varepsilon,0}(\omega) = \omega + C\sum_{i=1}^m |\bar{A}_i^T\omega - y_i|_\varepsilon\mathrm{sgn}\left(\bar{A}_i^T\omega - y_i\right)\bar{A}_i = 0, \tag{22}$$

which is indeed sufficient and necessary for (5). Hence, solving (22) is equivalent to solving (5).

Using the family of smoothing functions $\psi_{\varepsilon,p}$ to replace ε-loss function of (22) leads to a system of smooth equations. More specifically, we define a function $H_\psi : \mathbb{R}^{n+2} \to \mathbb{R}^{n+2}$ by

$$H_\psi(z) = H_\psi(\alpha,\omega) = \begin{bmatrix}\alpha\\ \omega + C\sum_{i=1}^m \psi_\varepsilon\left(\bar{A}_i^T\omega - y_i,\alpha\right)\mathrm{sgn}\left(\bar{A}_i^T\omega - y_i\right)\bar{A}_i\end{bmatrix}$$

where $z := (\alpha,\omega) \in \mathbb{R}^{n+2}$. From Lemma 3.1, it is easy to see that if $H_\psi(z) = 0$, then $\alpha = 0$ and ω is the solution of the equations (22), i.e., the solution of (12). Moreover, for any $z \in \mathbb{R}_{++} \times \mathbb{R}^{n+1}$, the function H_ψ is continuously differentiable with

$$H'_\psi(z) = \begin{bmatrix}1 & 0\\ E(\omega) & I + D(\omega)\end{bmatrix} \tag{23}$$

where

$$\begin{aligned}E(\omega) &= C\sum_{i=1}^m \nabla_\alpha\psi_\varepsilon\left(\bar{A}_i^T\omega - y_i,\alpha\right)\mathrm{sgn}\left(\bar{A}_i^T\omega - y_i\right)\bar{A}_i,\\ D(\omega) &= C\sum_{i=1}^m \nabla_x\psi_\varepsilon\left(\bar{A}_i^T\omega - y_i,\alpha\right)\mathrm{sgn}\left(\bar{A}_i^T\omega - y_i\right)\bar{A}_i\bar{A}_i^T.\end{aligned}$$

Because $\nabla_x \psi_\varepsilon(\bar{A}_i^T \omega - y_i, \alpha)\mathrm{sgn}(\bar{A}_i^T \omega - y_i)$ is nonnegative for any $\alpha > 0$ from (21), we see that $I + D(x)$ is positive definite at any $z \in \mathbb{R}_{++} \times \mathbb{R}^{n+1}$. Following the similar arguments as in Sect. 2, we obtain that $H'_\psi(z)$ is invertible at any $z \in \mathbb{R}_{++} \times \mathbb{R}^{n+1}$.

Proposition 3.1 *Let $\phi_\varepsilon(x, \alpha)$ be defined as in (6) and $\psi_{\varepsilon,p}(x, \alpha)$ be defined as in (19). Then, the following hold.*

(a) *For $p \geq 2$, we have $\phi_\varepsilon(x, \alpha) \geq \psi^2_{\varepsilon,p}(x, \alpha) \geq |x|^2_\varepsilon$.*
(b) *For $p \geq q \geq 2$, we have $\psi_{\varepsilon,q}(x, \alpha) \geq \psi_{\varepsilon,p}(x, \alpha)$.*

Proof (a) First, we show that $\phi_\varepsilon(x, \alpha) \geq \psi^2_{\varepsilon,p}(x, \alpha)$ holds. To proceed, we discuss four cases.
(i) If $|x| \leq \varepsilon - \alpha$, then $\phi_\varepsilon(x, \alpha) = 0 = \psi^2_{\varepsilon,p}(x, \alpha)$.
(ii) If $\varepsilon - \alpha < |x| < \varepsilon + \frac{\alpha}{p-1}$, then $|x| \leq \varepsilon + \frac{\alpha}{p-1}$ which is equivalent to $\frac{1}{|x|-\varepsilon+\alpha} \geq \frac{p-1}{\alpha p}$. Thus, we have

$$\frac{\phi_\varepsilon(x, \alpha)}{\psi^2_{\varepsilon,p}(x, \alpha)} = \frac{\alpha^{2p-3} p^{2p}}{6(p-1)^{2p-2}(|x| - \varepsilon + \alpha)^{2p-3}} \geq \frac{p^3}{6(p-1)} \geq 1,$$

which implies $\phi_\varepsilon(x, \alpha) \geq \psi^2_{\varepsilon,p}(x, \alpha)$.
(iii) For $\varepsilon + \frac{\alpha}{p-1} \leq |x| < \varepsilon + \alpha$, letting $t := |x| - \varepsilon \in [\frac{\alpha}{p-1}, \alpha)$ yields

$$\phi_\varepsilon(x, \alpha) - \psi^2_{\varepsilon,p}(x, \alpha) = \frac{1}{6\alpha}(t + \alpha)^3 - t^2 = t\left(\frac{1}{6\alpha}t^2 - \frac{1}{2}t + \frac{1}{2}\alpha\right) + \frac{1}{6}\alpha^2 \geq 0.$$

here the last inequality follows from the fact that discriminant of $\frac{1}{6\alpha}t^2 - \frac{1}{2}t + \frac{1}{2}\alpha$ is less than 0 and $\frac{1}{6\alpha} > 0$. Then, $\phi_\varepsilon(x, \alpha) - \psi^2_{\varepsilon,p}(x, \alpha) > 0$.
(iv) If $|x| \geq \varepsilon + \alpha$, then it is clear that $\phi_\varepsilon(x, \alpha) = (|x| - \varepsilon)^2 + \frac{1}{3}\alpha^2 \geq (|x| - \varepsilon)^2 = \psi^2_{\varepsilon,p}$.
Now we show that the other part $\psi_{\varepsilon,p}(x, \alpha) \geq |x|_\varepsilon$, which is equivalent to verifying $\psi^2_{\varepsilon,p}(x, \alpha) \geq |x|^2_\varepsilon$. Again, we discuss four cases.
(i) If $|x| \leq \varepsilon - \alpha$, then $\psi_{\varepsilon,p}(x, \alpha) = 0 = |x|_\varepsilon$.
(ii) If $\varepsilon - \alpha < |x| \leq \varepsilon$, then $\varepsilon - \alpha < |x|$ which says $|x| - \varepsilon + \alpha > 0$. Thus, we have $\psi_{\varepsilon,p}(x, \alpha) \geq 0 = |x|_\varepsilon$.
(iii) For $\varepsilon < |x| < \varepsilon + \frac{\alpha}{p-1}$, we let $t := |x| - \varepsilon \in (0, \frac{\alpha}{p-1})$ and define a function as

$$f(t) = \frac{\alpha}{p-1}\left(\frac{(p-1)(t+\alpha)}{p\alpha}\right)^p - t,$$

which is a function on $\left[0, \frac{\alpha}{p-1}\right]$. Note that $f(|x| - \varepsilon) = \psi_{\varepsilon,p}(x, \alpha) - |x|_\varepsilon$ for $|x| \in (\varepsilon, \varepsilon + \frac{\alpha}{p-1})$ and observe that

$$f'(t) = \left(\frac{(p-1)(t+\alpha)}{p\alpha}\right)^{p-1} - 1 \leq \left(\frac{(p-1)(\frac{\alpha}{p-1} + \alpha)}{p\alpha}\right)^{p-1} - 1 = 0.$$

This means $f(t)$ is monotone decreasing on $(0, \frac{\alpha}{p-1})$. Since $f(\frac{\alpha}{p-1}) = 0$, we have $f(t) \geq 0$ for $t \in (0, \frac{\alpha}{p-1})$, which implies $\psi_{\varepsilon,p}(x,\alpha) \geq |x|_\varepsilon$ for $|x| \in (\varepsilon, \varepsilon + \frac{\alpha}{p-1})$.
(iv) If $|x| \geq \varepsilon + \frac{\alpha}{p-1}$, then it is clear that $\psi_{\varepsilon,p}(x,\alpha) = |x| - \varepsilon = |x|_\varepsilon$.
(b) For $p \geq q \geq 2$, it is obvious to see that

$$\psi_{\varepsilon,q}(x,\alpha) = \psi_{\varepsilon,p}(x,\alpha) \quad \text{for} \;\; |x| \in [0, \varepsilon - \alpha] \cup \left[\varepsilon + \frac{\alpha}{q-1}, +\infty\right).$$

If $|x| \in [\varepsilon + \frac{\alpha}{p-1}, \varepsilon + \frac{\alpha}{q-1})$, then $\psi_{\varepsilon,p}(x,\alpha) = |x|_\varepsilon \leq \psi_{\varepsilon,q}(x,\alpha)$ from the above. Thus, we only need to prove the case of $|x| \in (\varepsilon - \alpha, \varepsilon + \frac{\alpha}{p-1})$.

Consider $|x| \in (\varepsilon - \alpha, \varepsilon + \frac{\alpha}{p-1})$ and $t := |x| - \varepsilon + \alpha$, we observe that $\frac{\alpha}{t} \geq \frac{p-1}{p}$. Then, we verify that

$$\begin{aligned}
\frac{\psi_{\varepsilon,q}(x,\alpha)}{\psi_{\varepsilon,p}(x,\alpha)} &= \frac{(q-1)^{q-1}p^p}{(p-1)^{p-1}q^q} \cdot \left(\frac{\alpha}{t}\right)^{p-q} \\
&\geq \frac{(q-1)^{q-1}p^p}{(p-1)^{p-1}q^q} \cdot \left(\frac{p-1}{p}\right)^{p-q} \\
&= \left(\frac{p}{q}\right)^q \cdot \left(\frac{q-1}{p-1}\right)^{q-1} \\
&= \frac{\left(1 + \frac{p-q}{q}\right)^q}{\left(1 + \frac{p-q}{q-1}\right)^{q-1}} \\
&\geq 1,
\end{aligned}$$

where the last inequality is due to $(1 + \frac{p-q}{x})^x$ being increasing for $x > 0$. Thus, the proof is complete. □

4 A smoothing Newton algorithm

In Sects. 2 and 3, we construct two systems of smooth equations: $H_\phi(z) = 0$ and $H_\psi(z) = 0$. We briefly describe the difference between $H_\phi(z) = 0$ and $H_\psi(z) = 0$. In general, the way we come up with $H_\phi(z) = 0$ and $H_\psi(z) = 0$ is a bit different. For achieving $H_\phi(z) = 0$, we first use the twice continuously differentiable functions $\phi_\varepsilon(x,\alpha)$ to replace $|x|_\varepsilon^2$ in problem (5), and then write out its KKT condition. To the contrast, for achieving $H_\psi(z) = 0$, we write out the KKT condition of problem (5) first, then we use the smoothing functions $\psi_{\varepsilon,p}(x,\alpha)$ to replace ε-loss function of (22) therein. For convenience, we denote $\tilde{H}(z) \in \{H_\phi(z), H_\psi(z)\}$. In other words, $\tilde{H}(z)$ possesses the property that if $\tilde{H}(z) = 0$, then $\alpha = 0$ and ω solves (12). In view of this, we apply some Newton-type methods to solve the system of smooth equations $\tilde{H}(z) = 0$ at each iteration and letting $\alpha \to 0$ so that the solution to the problem (12) can be found.

Algorithm 4.1 *(A smoothing Newton method)*

Step 0 *Choose* $\delta \in (0,1)$, $\sigma \in (0,\frac{1}{2})$, *and* $\alpha_0 > 0$. *Take* $\tau \in (0,1)$ *such that* $\tau\alpha_0 < 1$. *Let* $\omega_0 \in \mathbb{R}^{n+1}$ *be an arbitrary vector. Set* $z^0 := (\alpha_0, \omega_0)$. *Set* $e^0 := (1, 0, \ldots, 0) \in \mathbb{R}^{n+2}$.

Step 1 *If* $\|\tilde{H}(z^k)\| = 0$, *stop.*

Step 2 *Define function* Γ, β *by*

$$\Gamma(z) := \|\tilde{H}(z^k)\|^2 \quad \textit{and} \quad \beta(z) := \tau \min\{1, \Gamma(z)\}. \tag{24}$$

Compute $\triangle z^k := (\triangle\alpha_k, \triangle x^k)$ *by*

$$\tilde{H}\left(z^k\right) + \tilde{H}'\left(z^k\right)\triangle z^k = \alpha_0\beta\left(z^k\right)e^0.$$

Step 3 *Let* θ_k *be the maximum of the values* $1, \delta, \delta^2, \cdots$ *such that*

$$\Gamma\left(z^k + \lambda_k \Delta z^k\right) \leq \left[1 - 2\sigma\left(1 - \gamma\alpha_0\right)\theta_k\right]\Gamma\left(z^k\right). \tag{25}$$

Step 4 *Set* $z^{k+1} := z^k + \theta_k \Delta z^k$, *and* $k := k+1$, *Go to step 1.*

Proposition 4.1 *Suppose that the sequence* $\{z^k\}$ *is generated by Algorithm 4.1. Then, the following results hold.*

(a) $\{\Gamma(z^k)\}$ *is monotonically decreasing.*
(b) $\{\tilde{H}(z^k)\}$ *and* $\{\beta(z^k)\}$ *are monotonically decreasing.*
(c) *Let* $\mathcal{N}(\tau) := \{z \in \mathbb{R}_+ \times \mathbb{R}^{n+1} : \alpha_0\beta(z) \leq \alpha\}$, *then* $z^k \in \mathcal{N}(\tau)$ *for any* $k \in \mathcal{K}$ *and* $0 < \alpha_{k+1} \leq \alpha_k$.
(d) *The algorithm is well defined.*

Proof Since the proof is much similar to [6, Remark 2.1], we omit it here. □

Lemma 4.1 *Let* $\bar{\lambda} := \max\left\{\lambda_i(\sum_{i=1}^m \bar{A}_i \bar{A}_i{}^T)\right\}$. *Then, for any* $z \in \mathbb{R}_{++} \times \mathbb{R}^{n+1}$, *we have*

(a) $1 \leq \lambda_i(H'_\phi(z)) \leq 1 + 2\bar{\lambda}$, $i = 1, \cdots, n+2$;
(b) $1 \leq \lambda_i(H'_\psi(z)) \leq 1 + \bar{\lambda}$, $i = 1, \cdots, n+2$.

Proof (a) $H'_\phi(z)$ is continuously differentiable at any $z \in \mathbb{R}_{++} \times \mathbb{R}^{n+1}$, and by (18), it is easy to see that $\{1, \lambda_1(\nabla^2_{\omega\omega}F_{\varepsilon,\alpha}(\omega)), \cdots, \lambda_{n+1}(\nabla^2_{\omega\omega}F_{\varepsilon,\alpha}(\omega))\}$ are eigenvalues of $H'_\phi(z)$. From the representation of $\nabla^2_{xx}\phi_\varepsilon$ in (8), we have $0 \leq \nabla^2_{xx}\phi_\varepsilon(\bar{A}_i^T\omega - y_i, \alpha) \leq 2$. As $\nabla^2_{\omega\omega}F_{\varepsilon,\alpha}(\omega) = I + \sum_{i=1}^m \nabla^2_{xx}\phi_\varepsilon(\bar{A}_i^T\omega - y_i, \alpha)\bar{A}_i\bar{A}_i^T$, then

$$1 \leq \lambda_i\left(\nabla^2_{\omega\omega}F_{\varepsilon,\alpha}(\omega)\right) \leq 1 + 2\bar{\lambda}(i = 1, \cdots, n+1). \tag{26}$$

Thus the result (i) holds.

(b) Note that

$$\nabla_x \psi_{\varepsilon,p}(x,\alpha)\text{sgn}(x) = \begin{cases} 0 & 0 \leq |x| \leq \varepsilon - \alpha, \\ \left[\frac{(p-1)(|x|-\varepsilon+\alpha)}{p\alpha}\right]^{p-1} & \varepsilon - \alpha < |x| < \varepsilon + \frac{\alpha}{p-1}, \\ 1 & |x| \geq \varepsilon + \frac{\alpha}{p-1}, \end{cases}$$

which says $0 \leq \nabla_x \psi_{\varepsilon,p}(x,\alpha) \leq 1$. Then, following the similar arguments as in part(a), the result of part(b) cab be proved. □

Proposition 4.2 *$\{\tilde{H}(\alpha,\omega)\}$ is coercive for any fixed $\alpha > 0$, i.e., $\lim_{\|\omega\|\to+\infty} \|\tilde{H}(\alpha,\omega)\| = +\infty$.*

Proof We first claim that $\{H_\phi(\alpha,\omega)\}$ is coercive for any fixed $\alpha > 0$. By the definition of $H_\phi(\alpha,\omega)$ in (17), $\|H_\phi(\alpha,\omega)\|^2 = \alpha^2 + \|\nabla_\omega F_{\varepsilon,\alpha}(\omega)\|^2$. Then for any fixed $\alpha > 0$,

$$\lim_{\|\omega\|\to+\infty} \|H_\phi(\alpha,\omega)\| = +\infty \;\Leftrightarrow\; \lim_{\|\omega\|\to+\infty} \|\nabla_\omega F_{\varepsilon,\alpha}(\omega)\| = +\infty.$$

By (26), we have $\|\nabla^2_{\omega\omega} F_{\varepsilon,\alpha}(x,b)\| \geq 1$. For any $\omega_0 \in \mathbb{R}^{n+1}$,

$$\begin{aligned} \|\nabla_\omega F_{\varepsilon,\alpha}(\omega)\| + \|\nabla_\omega F_{\varepsilon,\alpha}(\omega_0)\| &\geq \|\nabla_\omega F_{\varepsilon,\alpha}(\omega) - \nabla_\omega F_{\varepsilon,\alpha}(\omega_0)\| \\ &= \|\nabla^2_{\omega\omega} F_{\varepsilon,\alpha}(\hat{\omega})(\omega - \omega_0)\| \\ &\geq \|\omega - \omega_0\|, \end{aligned}$$

where $\hat{\omega}$ between ω_0 and ω, then $\lim_{\|\omega\|\to+\infty} \|\nabla_\omega F_{\varepsilon,\alpha}(\omega)\| = +\infty$.

By a similar proof, we can get $\{H_\psi(\alpha,\omega)\}$ is coercive for any fixed $\alpha > 0$.

From the above, $\tilde{H}(\alpha,\omega) \in \{H_\phi(\alpha,\omega), H_\psi(\alpha,\omega)\}$ is coercive for any fixed $\alpha > 0$. □

Lemma 4.2 *Let $\Omega \subseteq \mathbb{R}^{n+1}$ be a compact set and $\Gamma(\alpha,\omega)$ be defined as in (24). Then, for every $\varsigma > 0$, there exists a $\bar{\alpha} > 0$ such that*

$$|\Gamma(\alpha,\omega) - \Gamma(0,\omega)| \leq \varsigma$$

for all $\omega \in \Omega$ and all $\alpha \in [0,\bar{\alpha}]$.

Proof The function $\Gamma(\alpha,\omega)$ defined as in (24) is continuous on the compact set $[0,\bar{\alpha}] \times \Omega$. The lemma is then an immediate consequence of the fact that every continuous function on a compact set is uniformly continuous there. □

Lemma 4.3 (Mountain Pass Theorem [12, Theorem 9.2.7]) *Suppose that $g : \mathbb{R}^m \to \mathbb{R}$ is a continuously differentiable and coercive function. Let $\Omega \subset \mathbb{R}^m$ be a nonempty and compact set and ξ be the minimum value of g on the boundary of Ω, i.e., $\xi := \min_{y\in\partial\Omega} g(y)$. Assume that there exist points $a \in \Omega$ and $b \notin \Omega$ such that $g(a) < \xi$ and $g(b) < \xi$. Then, there exists a point $c \in \mathbb{R}^m$ such that $\nabla g(c) = 0$ and $g(c) \geq \xi$.*

Theorem 4.1 *Suppose the sequence $\{z^k\}$ is generated by Algorithm 4.1. Then, the sequence $\{z^k\}$ is bounded, and $\omega^k = (x^k, b^k)$ converges to the unique solution $\omega^{sol} = (x^{sol}, b^{sol})$ of problem (12).*

Proof (a) We first show that the sequence $\{z^k\}$ is bounded. It is clear from Proposition 4.1(c) that the sequence $\{\alpha_k\}$ is bounded. In the following two cases, by assuming that $\{\omega^k\}$ is unbounded, we will derive a contradiction. By passing to subsequence if necessary, we assume $\|\omega^k\| \to +\infty$ as $k \to +\infty$. Then, we discuss two cases.

(i) If $\alpha_* = \lim\limits_{k\to+\infty} \alpha_k > 0$, applying Proposition 4.1(b) yields that $\left\{\tilde{H}(z^k)\right\}$ is bounded. In addition, by Proposition 4.2, we have

$$\lim_{k\to+\infty} \tilde{H}(\alpha_*, \omega^k) = \lim_{\|\omega^k\|\to+\infty} \tilde{H}(\alpha_*, \omega^k) = +\infty. \tag{27}$$

Hence, a contradiction is reached.

(ii) If $\alpha_* = \lim\limits_{k\to+\infty} \alpha_k = 0$, by assuming that $\{\omega^k\}$ is unbounded, there exists a compact set $\Omega \subset \mathbb{R}^n$ with

$$\omega^{sol} \notin \Omega \tag{28}$$

for all k sufficiently large. Since

$$\bar{m} := \min_{\omega\in\partial\Omega} \Gamma(0, \omega) > 0,$$

we can apply Lemma 4.2 with $\varsigma := \bar{m}/4$ and conclude that

$$\Gamma\left(\alpha_k, \omega^{sol}\right) \leq \frac{1}{4}\bar{m} \tag{29}$$

and

$$\min_{\omega\in\partial\Omega} \Gamma(\alpha_k, \omega) \geq \frac{3}{4}\bar{m}$$

for all k sufficiently large. Since $\alpha_k \to 0$ in this case, combining (24) and Proposition 4.1(c) gives

$$\Gamma\left(\alpha_k, \omega^k\right) = \beta\left(\alpha_k, \omega^k\right) \leq \alpha_k/\alpha_0.$$

Hence,

$$\Gamma\left(\alpha_k, \omega^k\right) \leq \frac{1}{4}\bar{m} \tag{30}$$

for all k sufficiently large. Now let us fix an index k such that (29) and (30) hold. Applying the Mountain Pass Theorem 4.3 with $a := \omega^{sol}$ and $b := \omega^k$, we obtain the existence of a vector $c \in \mathbb{R}^{n+1}$ such that

$$\nabla_\omega \Gamma(\alpha_k, c) = 0 \quad \text{and} \quad \Gamma(\alpha_k, c) \geq \frac{3}{4}\bar{m} > 0. \tag{31}$$

To derive a contradiction, we need to show that c is a global minimizer of $\Gamma(\alpha_k, \omega)$. Since $\Gamma(\alpha_k, \omega) \geq \alpha^2$, it is sufficient to show $\Gamma(\alpha_k, c) = \alpha^2$. We discuss this in two cases:

- If $\tilde{H} = H_p$, then

  $$\nabla_\omega \Gamma(\alpha_k, c) = 2\nabla^2_{\omega\omega} F_{\varepsilon,\alpha_k}(c)\bar{H}_p(\alpha_k, c)$$

  where $\bar{H}_p$ is the last $n+1$ components of H_ϕ, i.e., $\bar{H}_p = H_p(2 : n+2)$. Then, using (31) and the fact that $\nabla^2_{\omega\omega} F_{\varepsilon,\alpha_k}(c)$ is invertible for $\alpha_k > 0$, we have $\bar{H}_p(\alpha_k, c) = 0$. Furthermore,

  $$\Gamma(\alpha_k, c) = \|H(\alpha_k, c)\|^2 = \alpha^2.$$

- If $\tilde{H} = H_\psi$, then

  $$\nabla_\omega \Gamma(\alpha_k, c) = 2(I + D(\omega))\bar{H}_\psi(\alpha_k, c)$$

  where $I + D(\omega)$ is given by (23) and $\bar{H}_\psi$ is the last $n + 1$ components of H_ψ. Since $I + D(\omega)$ is invertible for $\alpha_k > 0$, we obtain that $\Gamma(\alpha_k, c) = \alpha^2$ by the same way as in the above case.

(b) From Proposition 4.1, we know that sequences $\{\tilde{H}(z^k)\}$ and $\{\Gamma(z^k)\}$ are nonnegative and monotone decreasing, and hence they are convergent. In addition, by using the first result of this theorem, we obtain that the sequence $\{z^k\}$ is bounded. Passing to subsequence if necessary, we may assume that there exists a point $z^* = (\alpha_*, \omega^*)\mathbb{R}_{++} \times \mathbb{R}^{n+1}$ such that $\lim_{k\to+\infty} z^k = z^*$, and hence,

$$\lim_{k\to+\infty} \|H(z^k)\| = \|H(z^*)\| \quad \text{and} \quad \lim_{k\to+\infty} \Gamma(z^k) = \Gamma(z^*).$$

For $H(z^*) = 0$, by a simple continuity discussion, we obtain that ω^* is a solution to problem (12). For the case of $H(z^*) > 0$, and hence $\alpha^* > 0$, we will derive a contradiction. First, by the assumption that $H(z^*) > 0$, we have $\lim_{k\to+\infty} \theta_k = 0$. Thus, for any sufficiently large k, the stepsize $\hat{\theta}_k := \theta_k/\delta$ does not satisfy the line search criterion (25), i.e.,

$$\Gamma\left(z^k + \hat{\theta}_k \triangle z^k\right) > \left[1 - 2\sigma(1 - \gamma\alpha_0)\hat{\theta}_k\right]\Gamma\left(z^k\right),$$

which implies that

$$\frac{\Gamma\left(z^k + \hat{\theta}_k \triangle z^k\right) - \Gamma(z^k)}{\hat{\theta}_k} > -2\sigma(1 - \gamma\alpha_0)\Gamma\left(z^k\right).$$

Since $\alpha_* > 0$, it follows that $\Gamma(z^k)$ is continuously differentiable at z^*. Letting $k \to +\infty$ in the above inequality gives

$$\begin{aligned}
&-2\sigma(1-\gamma\alpha_0)\Gamma(z^*)\\
&\quad \le 2\tilde{H}\left(z^*\right)^T \tilde{H}'\left(z^*\right)\triangle z^* = 2\tilde{H}\left(z^*\right)^T\left(-\tilde{H}\left(z^*\right)+\alpha_0\beta\left(z^*\right)e^0\right)\\
&\quad = -2\tilde{H}\left(z^*\right)^T \tilde{H}\left(z^*\right) + 2\alpha_0\beta\left(z^*\right)\tilde{H}\left(z^*\right)^T e^0\\
&\quad \le 2\left(-1+\gamma\alpha_0\right)\Gamma\left(z^*\right).
\end{aligned}$$

This indicates that $-1+\gamma\alpha_0+\sigma(1-\gamma\alpha_0) \ge 0$, which contradicts the fact that $\gamma\alpha_0 < 1$. Thus, there should be $\tilde{H}(z^*) = 0$.

Because the unique solution to problem (12) is ω^{sol}, we have $z^* = (0, \omega^{sol})$ and the whole sequence $\{z^k\}$ converge to z^*, that is,

$$\lim_{k\to+\infty} z^* = \left(0, \omega^{sol}\right).$$

Then, the proof is complete. □

In the following, we discuss the local convergence of Algorithm 4.1. To this end, we need the concept of semismoothness, which was originally introduced by Mifflin [10] for functionals and was further extended to the setting of vector-valued functions by Qi and Sun [14]. A locally Lipschitz continuous function $F : \mathbb{R}^n \to \mathbb{R}^m$, which has the generalized Jacobian $\partial F(x)$ in the sense of Clarke [2], is said to be semismooth (respectively, strongly semismooth) at $x \in \mathbb{R}^n$, if F is directionally differentiable at x and

$$F(x+h) - F(x) - Vh = o(\|h\|) \quad (= O(\|h\|^2), \text{respectively})$$

holds for any $V \in \partial F(x+h)$.

Lemma 4.4 (a) *Suppose that the sequence $\{z^k\}$ is generated by Algorithm 4.1. Then,* $\left\|\tilde{H}'(z^k)^{-1}\right\| \le 1$.

(b) *$\tilde{H}(z)$ is strongly semismooth at any $z = (\alpha, \omega) \in \mathbb{R}^{n+2}$.*

Proof (a) By Proposition 4.1 (c), we know that $\alpha_k > 0$ for any $k \in \mathcal{K}$. This together with Lemma 4.1 leads to the desired result.

(b) We only provide the proof for the case of $\tilde{H}(\alpha, \omega) = H_\phi(\alpha, \omega)$. For the other case of $\tilde{H}(\alpha, \omega) = H_\psi(\alpha, \omega)$, the proof is similar and is omitted. First, we observe that $H'_\phi(z)$ is continuously differentiable and Lipschitz continuous at $z \in \mathbb{R}_{++} \times \mathbb{R}^{n+1}$ by Lemma 4.1(a). Thus, $H_\phi(z)$ is strongly semismooth at $z \in \mathbb{R}_{++} \times \mathbb{R}^{n+1}$. It remains to verify that $H_\phi(z)$ is strongly semismooth at $z \in \{0\} \times \mathbb{R}^{n+1}$. To see this, we recall that

$$\nabla_x \phi_\varepsilon(x, 0) = \begin{cases} 2(|x|-\varepsilon)\mathrm{sgn}(x), & |x|-\varepsilon \ge 0;\\ 0, & |x|-\varepsilon \le -\alpha. \end{cases}$$

It is a piecewise linear function, and hence $\nabla_x \phi_\varepsilon(x, 0)$ is a strongly semismooth function. In summary, $H_\phi(z)$ is strongly semismooth at $z \in \{0\} \times \mathbb{R}^{n+1}$. □

Theorem 4.2 *Suppose that $z^* = (\mu_*, x^*)$ is an accumulation point of $\{z^k\}$ generated by Algorithm 4.1. Then, we have*

(a) $\|z^{k+1} - z^*\| = O(\|z^k - z^*\|^2)$;
(b) $\alpha_{k+1} = O(\alpha_k^2)$.

Proof The proof can be done by using Lemma 4.4 and following the similar arguments in [6, Theorem 3.2]. □

5 Smooth support vector machines with nonlinear kernel

In this section, we talk about the nonlinear kernel which is traditionally applied to smooth support vector machines. More specifically, we employ the kernel technique that has been used extensively in kernel-based learning algorithm. For convenience, we denote the kernel function by $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. There are many popular choices of K, for example, the linear kernel

$$K(A_i, A_j) = A_i^T A_j$$

where A_i means the i-th input data, and the radial basis function (rbf) kernel

$$K(A_i, A_j) = \exp(-\gamma \|A_i - A_j\|^2)$$

where $\gamma > 0$ is a constant. Other types of nonlinear kernels are the polynomial kernel

$$K(A_i, A_j) = \left(\gamma A_i^T A_j + s\right)^{\text{deg}},$$

and the sigmoid kernel

$$K(A_i, A_j) = \tanh\left(\gamma A_i^T A_j + s\right).$$

here, as mentioned in [15], $\gamma = 1/n$, $s = 0$, and deg $= 3$. Also, a reduced kernel is proposed recently [5,7].

For our numerical implementation, we define m dimension row vector $K(A_i, A^T) : \mathbb{R}^{n\times 1} \times \mathbb{R}^{n\times m} \to \mathbb{R}^{1\times m}$ with

$$K\left(A_i, A^T\right)_j = K\left(A_i, A^j\right), \quad j = 1, \cdots, m. \tag{32}$$

In addition, we denote $\bar{B} = [B, 1]$, $B = [B_1, \cdots, B_m]^T$ with

$$B_i = K\left(A_i, A^T\right)^T, \quad i = 1, \cdots, m. \tag{33}$$

Then, (5) and (11) can be recast as

$$\begin{aligned} &\min_{(u,b)\in \mathbb{R}^{m+1}} \tfrac{1}{2}\left(u^T u + b^2\right) + \tfrac{C}{2}\textstyle\sum_{i=1}^m |K(A_i, A^T)u + b - y_i|_\varepsilon^2 \\ &\quad = \min_{(u,b)\in \mathbb{R}^{m+1}} \tfrac{1}{2}\left(u^T u + b^2\right) + \tfrac{C}{2}\textstyle\sum_{i=1}^m |B_i^T u + b - y_i|_\varepsilon^2, \end{aligned} \tag{34}$$

$$\min_{\upsilon\in \mathbb{R}^{m+1}} \hat{F}_{\varepsilon,\alpha}(\upsilon) := \min_{\upsilon\in \mathbb{R}^{m+1}} \frac{1}{2}\upsilon^T\upsilon + \frac{C}{2}1^T\Phi_\varepsilon(\bar{B}\upsilon - y, \alpha). \tag{35}$$

where $\upsilon = (u, b) \in \mathbb{R}^{m+1}$. The problems (34) and (35) have exactly the same forms as the problems (5) and (11) except that A is replaced by B and the dimension of variables is replaced by $m+1$. Thus, we can directly apply our methods for ε-insesitive nonlinear support vector regression.

6 Numerical results

In this section, we report the numerical results of Algorithm 4.1 for solving the SSVR (3). All numerical experiments are carried out in MATLAB R2015a 64-bit running on a PC with Intel Xeon E3 1231 v3 @ 3.40GHz CPU, 16.0GB DDR3 RAM and Windows 10 64-bit operating system.

In our numerical experiments, the stopping criteria for Algorithm 4.1 is $\|\tilde{H}(z^k)\| < 1\mathrm{e}-6$. We also stop programs when the number of total iterations is more than 20. Throughout the computational experiments, the following parameters are used:

$$\varepsilon = 0.1, \quad C = 100, \quad \delta = 0.3, \quad \sigma = 0.03, \quad \tau = 0.3.$$

In each fitting, we randomly choose ω_0 and all components are generated independently from uniform distribution over $[-1, 1]$.

Table 1 presents eight benchmark datasets that we implement. The first seven are “abalone”, “bodyfat”,“housing”, “mg”, “mpg”, “pyrim”, “space ga”, and “triazines”, which come from LIBSVM Data: Regression. The last dataset is “Friedman #1” regression problem appeared in [4]. The input features $x = (x_1, x_2, \cdots, x_{10})$ are generated

Table 1 Datasets

	#Samples	#Features
abalone	4177	8
bodyfat	252	14
housing	506	13
mg	1385	6
mpg	392	7
pyrim	74	27
space ga	3107	6
triazines	186	60
Friedman1	2500	10

Table 2 Numerical results of $\phi_\varepsilon(x, \alpha)$ when $\alpha_0 = 1e-1, 1e-3, 1e-5$ and $1e-7$

α_0	Number of iterations				Computing time (in s)			
	1e−1	1e−3	1e−5	1e−7	1e−1	1e−3	1e−5	1e−7
abalone	20	14.75	5.05	4.1	2.2113	1.4125	0.39054	0.32613
bodyfat	16.7	9.05	8.2	7.7	0.01102	0.0061242	0.0053697	0.00502
housing	16.35	4.85	4.85	3.65	0.022907	0.0061034	0.0062035	0.0046132
mg	20	14.9	11.2	11.7	0.18556	0.12624	0.080191	0.090198
mpg	16.45	4.6	4.25	3.25	0.019591	0.004651	0.0037961	0.0033604
pyrim	20	17.15	14.85	12.75	0.016807	0.015491	0.012851	0.010785
space ga	20	12.75	8.2	7.6	1.159	0.53963	0.31917	0.29891
triazines	20	18.2	13.7	15.2	0.043735	0.041883	0.031706	0.036713
Friedman1	20	8.55	5.15	4.15	2.5858	1.0182	0.59597	0.49187

Table 3 Numerical results of $\psi_{\varepsilon,p}(x, \alpha)$ when $\alpha_0 = 1e-1, 1e-3, 1e-5$ and $1e-7$

α_0	Number of iterations				Computing time (in s)			
	1e−1	1e−3	1e−5	1e−7	1e−1	1e−3	1e−5	1e−7
abalone	20	12.1	4.9	4.05	1.9415	1.11	0.37749	0.32293
bodyfat	15.3	8.45	7.85	7.7	0.012084	0.0061134	0.0054757	0.0051973
housing	16.7	4.65	4.2	3.4	0.025137	0.0057001	0.0057374	0.0041698
mg	20	13.55	9.7	10.9	0.17853	0.10258	0.069653	0.078864
mpg	16.05	4.15	4.3	3.1	0.019766	0.0039988	0.0039729	0.0029461
pyrim	20	16.1	15.2	13.55	0.015815	0.014137	0.013878	0.011996
space ga	20	10.6	7.65	7.6	1.0904	0.42223	0.29601	0.29527
triazines	20	17.55	15.25	12.75	0.043612	0.039669	0.034606	0.028883
Friedman1	20	9.1	5.4	4.15	2.6285	1.1139	0.62455	0.48496

independently from uniform distribution over [0, 1]. The output target function is defined by

$$y(x) = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + N(0, 1)$$

where $N(0, 1)$ is the normally distributed noise with mean 0 and variance 1. For consistency, we linearly scale the last dataset to $[-1, 1]$. All datasets are transformed to RBF kernel space with $\gamma = 10$. To speed up our smoothing method, if $m/10 > n$, we use reduced kernel matrix [9] $K(A, A^T) \in \mathbb{R}^{m\times m}$ to $K(A, \tilde{A}^T) \in \mathbb{R}^{m\times \tilde{m}}$ where $\tilde{m} = \lceil m/10 \rceil$. We run each test case 20 times and average the number of iterations and computing time.

To compare the performance, we consider the the performance profile which is introduced in [3] as a means. In other words, we regard Algorithm 4.1 corresponding to a smoothing function $\phi_\varepsilon(x, \alpha)$ or $\psi_{\varepsilon,p}(x, \alpha)$ with specific parameters as a solver,

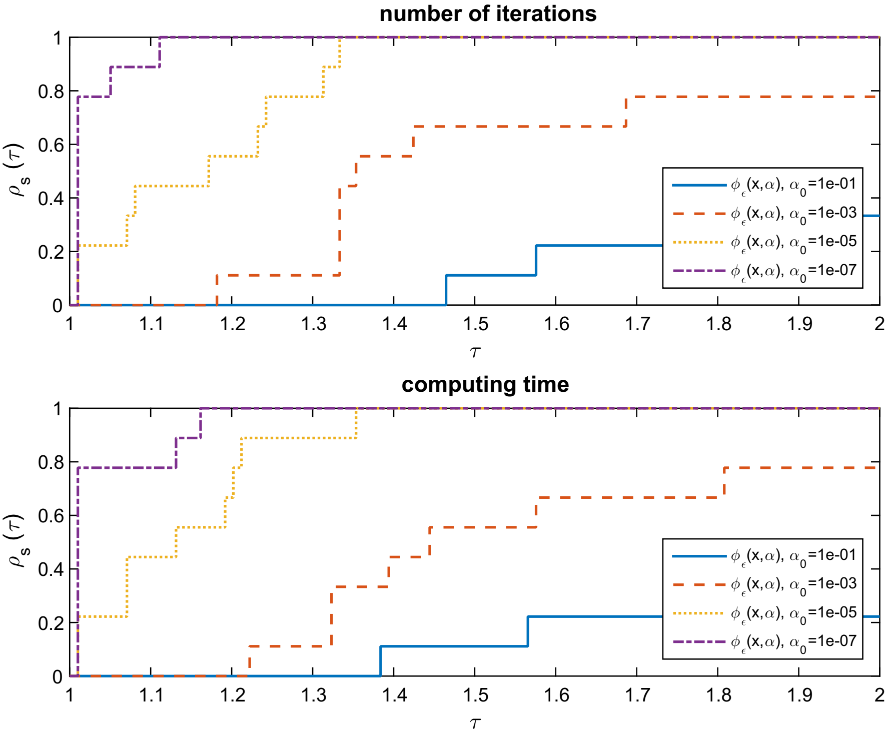

Fig. 4 Performance profile of $\phi_\varepsilon(x, \alpha)$ when $\alpha_0 = 1e - 1$, $1e - 3$, $1e - 5$ and $1e - 7$

and assume that there are n_s solvers and n_p test problems from the test set P which is the datasets mentioned above. We are interested in using the iteration number as performance measure for Algorithm 4.1 with different settings. For each problem p and a solver s, let

$$f_{p,s} = \text{iteration number required to solve problem } p \text{ by solver } s.$$

We employ the performance ratio

$$r_{p,s} = \frac{f_{p,s}}{\min\{f_{p,s} \,|\, s \in S\}}$$

where S is the datasets. We assume that a parameter $r_{p,s} \le r_M$ for all p, s are chosen, and $r_{p,s} = r_M$ if and only if solver s does not solve problem p. In order to obtain an overall assessment for each solver, we define

$$\rho_s(\tau) := \frac{1}{n_p} \text{size} \left\{ p \in P \,|\, r_{p,s} \le \tau \right\},$$

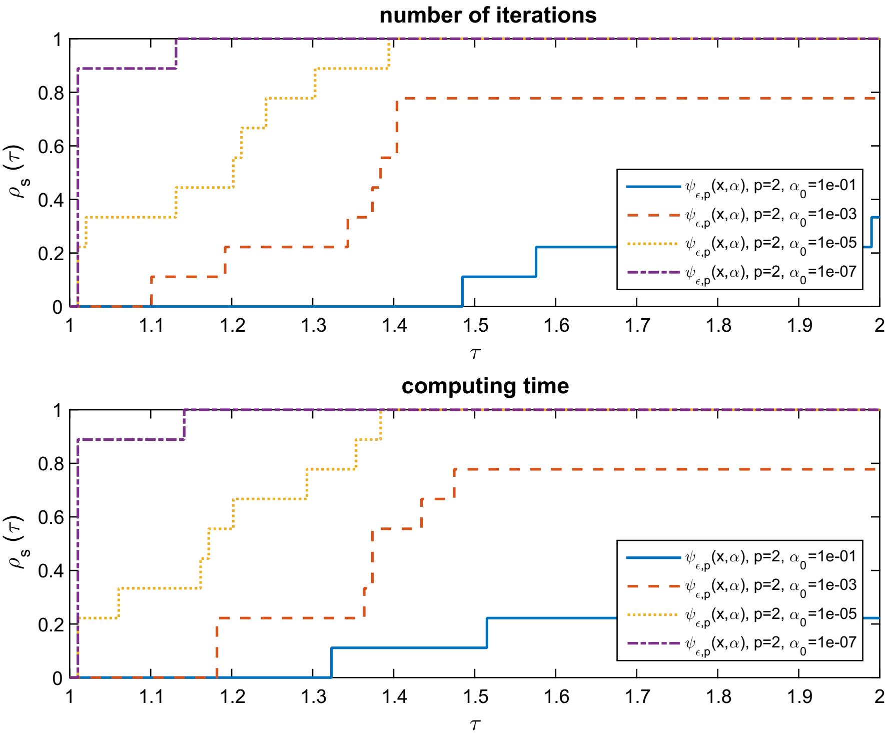

Fig. 5 Performance profile of $\psi_{\varepsilon,p}$ when $\alpha_0 = 1e-1$, $1e-3$, $1e-5$ and $1e-7$

which is called the performance profile of the number of iteration for solver s. Then, $\rho_s(\tau)$ is the probability for solver $s \in S$ that a performance ratio $f_{p,s}$ is within a factor $\tau \in \mathbb{R}$ of the best possible ratio.

We summarize all the comparison results as below.

1. First, we compare the initial values $\alpha_0 = 1e-1$, $1e-3$, $1e-5$ and $1e-7$ for $\phi_\varepsilon(x,\alpha)$ and $\psi_{\varepsilon,p}(x,\alpha)$. The values of p is fixed as 2. The numerical results are listed in Tables 2 and 3. Figures 4 and 5 are the performance profile of iteration numbers and computing times of $\phi_\varepsilon(x,\alpha)$ and $\psi_{\varepsilon,p}(x,\alpha)$. Both figures show that the case of $\alpha_0 = 1e-1$ is the worst, while the case of $\alpha_0 = 1e-7$ performs well.
2. Second, we compare $p = 2, 5, 10, 100$ for $\psi_{\varepsilon,p}(x,\alpha)$. The values of α_0 is fixed as $1e-2$. The numerical results are listed in Table 4. From Fig. 6, we see that $p = 100$ outperforms other values of p.
3. Third, we compare smoothing $\phi_\varepsilon(x,\alpha)$ and $\psi_{\varepsilon,p}(x,\alpha)$ for $p = 2$ and 100. The values of α_0 is fixed as $1e-5$. The numerical results are listed in Table 5. Figure 7 shows that $\psi_{\varepsilon,p}(x,\alpha)$ with $p = 2$ or $p = 100$ performs better than $\phi_\varepsilon(x,\alpha)$.
4. Finally, we compare smoothing $\phi_\varepsilon(x,\alpha)$ and $\psi_{\varepsilon,p}(x,\alpha)$ for $p = 3$ with LIBSVM, which is the most powerful and successful public software for support vector classification, regression, and distribution estimation. The values of α_0 is fixed as $1e-5$. The numerical comaprisons are listed in Table 6. From Fig. 8, we see that our

Table 4 Numerical results of $\psi_{\varepsilon,p}(x,\alpha)$ when $p = 2, 5, 10$ and 100

p	Number of iterations				Computing time (in s)			
	2	5	10	100	1e−1	1e−3	1e−5	1e−7
abalone	19.05	19.3	11.85	5.8	1.9285	1.738	0.9632	0.45631
bodyfat	9.6	9	9.2	8.75	0.007148	0.0069386	0.0073081	0.0063591
housing	7.25	5	4.75	4.5	0.0097576	0.005612	0.0057854	0.0054264
mg	15.05	12.6	11.6	11.5	0.11152	0.092907	0.08445	0.086844
mpg	5.8	5.35	4.45	4.15	0.006003	0.0052232	0.0043621	0.0042389
pyrim	17.75	17.65	17.8	15.05	0.016143	0.015634	0.016047	0.013043
space ga	18.15	12.8	11.35	8.8	0.79167	0.53068	0.4546	0.34446
triazines	19.9	17.75	15.4	14.15	0.0439	0.039515	0.034067	0.03288
Friedman1	11.2	9.7	8.1	5.45	1.3007	1.1126	0.9258	0.63419

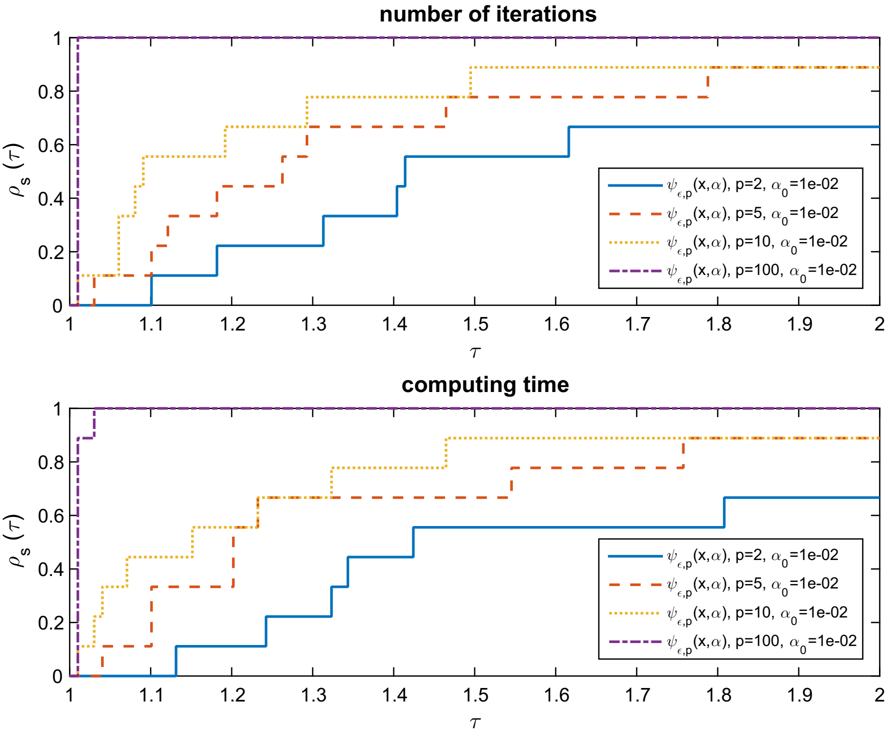

Fig. 6 Performance profile of $\psi_{\varepsilon,p}(x,\alpha)$ when $p = 2, 5, 10, 100$

Table 5 Numerical results of $\phi_\varepsilon(x, \alpha)$ and $\psi_{\varepsilon,p}(x, \alpha)$

	Number of iterations			Computing time (in s)		
	$\phi(x, \alpha)$	$\psi_2(x, \alpha)$	$\psi_{100}(x, \alpha)$	$\phi(x, \alpha)$	$\psi_2(x, \alpha)$	$\psi_{100}(x, \alpha)$
abalone	5.25	5	5	0.41548	0.38232	0.38231
bodyfat	8.35	7.85	7.7	0.0055656	0.0052997	0.0051566
housing	4.75	4.45	4.6	0.0060158	0.0054452	0.0051749
mg	11.05	11.05	10	0.078979	0.080433	0.071441
mpg	4.25	4.1	4.15	0.0041321	0.0040231	0.0039307
pyrim	15.25	14.8	12.5	0.013206	0.013228	0.011089
space ga	8.4	7.6	7.65	0.33463	0.29161	0.29746
triazines	13.1	14.05	15.05	0.029611	0.031629	0.033943
Friedman1	5.25	5.05	5.05	0.59893	0.57985	0.57964

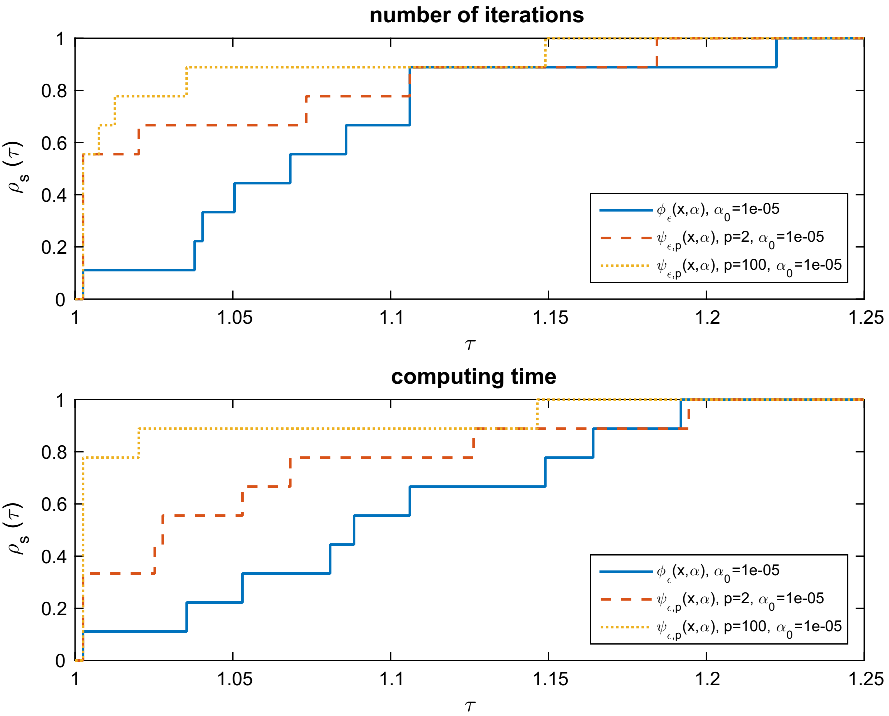

Fig. 7 Performance profile of $\phi_\varepsilon(x, \alpha)$ and $\psi_{\varepsilon,p}(x, \alpha)$ with $p = 2, 100$

Table 6 Numerical results of $\phi_\varepsilon(x,\alpha)$, $\psi_{\varepsilon,p}(x,\alpha)$ and LIBSVM

	Computing time (in s)		
	SSVR FIRST TYPE	SSVR SECOND TYPE with $p = 3$	LIBSVM
abalone	0.38946	0.38636	4.8866
bodyfat	0.0054786	0.0050574	8.8947e−05
housing	0.0052277	0.0054718	0.065959
mg	0.080511	0.083267	0.45638
mpg	0.0037805	0.0041339	0.042289
pyrim	0.013043	0.012822	0.00033134
space ga	0.31607	0.29679	3.1808
triazines	0.033445	0.034838	0.003017
Friedman1	0.59054	0.57585	3.628

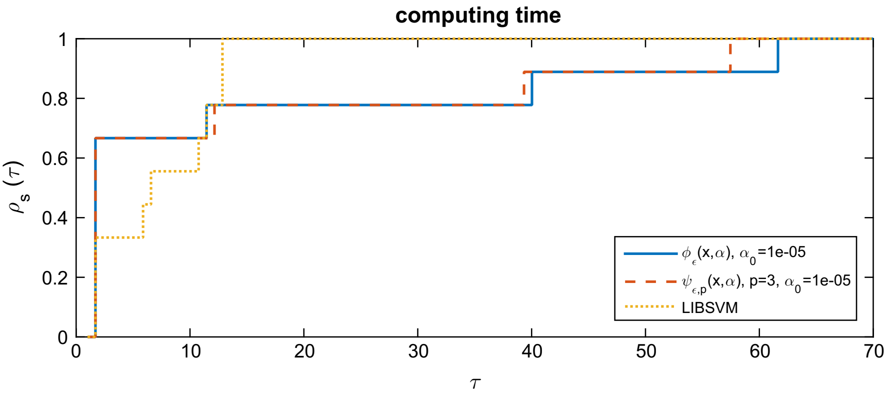

Fig. 8 Performance profile of $\phi_\varepsilon(x,\alpha)$, $\psi_{\varepsilon,p}(x,\alpha)$ with $p = 3$ and LIBSVM

smoothing method performs better than LIBSVM for some datasets. In particular, from Table 6, we see that our method performs much better in large datasets. We point out that we use reduced kernel and do not compare the training/testing set correctness.

References

1. Basak, D., Pal, S., Patranabts, D.C.: Support vector regression. Neural Inf. Process. Lett. Rev. **11**, 203–224 (2007)
2. Clarke, F.H.: Opimization and Nonsmooth Analysis. Wiley, New York (1983)
3. Dolan, E., Moré, J.: Benchmarking optimization software with performance profiles. Math. Program. **91**, 201–213 (2002)
4. Friedman, J.H.: Multivariate adaptive regression splines. Ann. Stat. **19**, 1–67 (1991)
5. Huang, C.-M., Lee, Y.-J.: Reduced support vector machines: a statistical theory. IEEE Trans. Neural Netw. **18**, 1–13 (2007)

6. Huang, Z.-H., Zhang, Y., Wu, W.: A smoothing-type algorithm for solving system of inequalities. J. Comput. Appl. Math. **220**, 355–363 (2008)
7. Lee, Y.-J., Hsieh, W.-F., Huang, C.-M.: ε-SSVR: a smooth support vector machine for ε-insensitive regression. IEEE Trans. Knowl. Eng. **17**, 678–685 (2005)
8. Lee, Y.-J., Mangasarian, O.L.: SSVM: a smooth support vector machine for classification. Comput. Optim. Appl. **20**, 5–22 (2001)
9. Lee, Y.-J., Mangasarian, O.L.: RSVM: Reduced support vector machines. In: Proceedings of the 2001 SIAM International Conference on Data Mining, https://doi.org/10.1137/1.9781611972719.13, (2001)
10. Mifflin, R.: Semismooth and semiconvex functions in constrained optimization. SIAM J. Control Optim. **15**, 957–972 (1977)
11. Musicant, D.R., Feinberg, A.: Active set support vector regression. IEEE Trans. Neural Netw. **15**, 268–275 (2004)
12. Palais, R.S., Terng, C.-L.: Critical Point Theory and Submanifold Geometry, Lecture Notes in Mathematics, vol. 1353, Springer, Berlin (1988)
13. Platt, J.: Sequential minimal optimization: a fast algorithm for training support vector machines. In: Advances in Kernel Methods, Support Vector Learning, vol. 208, pp. 1–21, MIT Press, Boston. (1998)
14. Qi, L.-Q., Sun, J.: A nonsmooth version of Newton's method. Math. Program. **58**, 353–367 (1993)
15. Tseng, P., Yun, S.: A coordinate gradient descent method for linearly constrained smooth optimization and support vector machines training. Comput. Optim. Appl. **47**, 179–206 (2010)
16. Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer, New York (1982)
17. Vapnik, V.: The Natrure of Statistical Theory. Springer, New York (1995)
18. Vapnik, V., Golowith, S., Smola, A.: Support vector method for function approximation, regression estimation, and signal processing. Neural Inf. Process. Syst. **9**, 281–287 (1997)
19. Yuan, Y.-B., Huang, T.-Z.: A polynomial smooth support vector machine for classification. Adv. Data Mining Appl. **3584**, 157–164 (2005)