JOTA manuscript No.
(will be inserted by the editor)

Smoothing Functions for Sparse Optimization: A Unified Framework

Chieu Thanh Nguyen · Jein-Shan Chen

Received: date / Accepted: date

Abstract This paper presents a unified framework for constructing smoothing functions tailored to a broad class of widely used regularizers, including the plus function, the pinball function, the ℓ_0-norm, the ℓ_p-norm for $0 < p \leq 1$, the MCP, and the SCAD. By transforming nonsmooth regularizers into smooth approximations, the proposed framework facilitates the application of efficient optimization algorithms to sparse optimization problems. The framework is systematically derived from continuous approximations of the step function, offering a principled approach to generating smoothing functions across various regularizers. These approximations are, in turn, constructed using polynomial functions and the Dirac delta function.

Keywords Regularizer · Regularized function · Smoothing function.

Mathematics Subject Classification (2000) 49J52 · 90C26 · 90C90 · 94A12

1 Introduction

A general sparse optimization problem, frequently encountered in fields such as machine learning, data science, and artificial intelligence, can be formulated as follows:

$$\min_{x \in \Omega} \; f(x) + \lambda R(x), \tag{1}$$

where $\Omega \subseteq \mathbb{R}^n$ is the feasible set. The function $f : \mathbb{R}^n \to \mathbb{R}$ typically represents a loss or cost function, serving as a data-fitting term, while the regularizer $R : \mathbb{R}^n \to \mathbb{R}$ promotes sparsity and mitigates overfitting. The design of $R(x)$ aims to incorporate structural constraints that guide the solution toward desired properties. The hyperparameter λ governs the trade-off between fidelity to the data and the strength of regularization, thereby striking a balance between model complexity and generalization. In general, Ω may also be a subset of $\mathbb{R}^{m\times n}$ when considering

Communicated by Felipe Lara.

Chieu Thanh Nguyen
Faculty of Information Technology,
Vietnam National University of Agriculture, Hanoi 131000, Vietnam
ntchieu@vnua.edu.vn

Jein-Shan Chen, Corresponding author
Department of Mathematics, National Taiwan Normal University
jschen@math.ntnu.edu.tw

matrix-valued sparse optimization problems, such as low-rank matrix recovery. In this setting, one similarly considers $f, R : \mathbb{R}^{m\times n} \to \mathbb{R}$.

Numerous regularization functions have been introduced in the literature, including the plus function, the pinball function, the ℓ_0-norm, the ℓ_p-norm for $0 < p \leq 1$, the minimax concave penalty (MCP), and the smoothly clipped absolute deviation (SCAD) function. Model (1) typically falls under the class of nonsmooth optimization problems, owing to the nonsmoothness of either the objective function f or the regularizer p, or both. To address this inherent nonsmoothness, smoothing techniques are often employed to approximate nonsmooth functions with smooth surrogates, thereby enabling the application of gradient-based optimization algorithms. These techniques facilitate efficient computation while preserving critical features of the original problem, such as sparsity and, in some cases, convexity.

A substantial body of research has been devoted to reformulating nonsmooth problems as smooth ones in order to leverage the power of well-established optimization methods. For instance, Chen [6] investigated smoothing functions for the ℓ_p-norm ($0 < p \leq 1$) in the context of nonsmooth and nonconvex optimization, extending earlier work by Chen and Mangasarian [5] on smoothing the plus function. Beck and Teboulle [1] proposed the infimal convolution smoothing technique for nonsmooth convex minimization, which encompasses several existing methods, including Nesterov's smoothing scheme [17]. In addition, a wide range of nonsmooth optimization problems have been successfully addressed through smooth approximations, as documented in works such as [3,32,21,2,11,18,19,27,20,28,14], among others. These studies firmly establish the effectiveness of smoothing techniques in handling nonsmooth optimization. Motivated by this, we propose a unified and systematic framework for constructing smoothing functions tailored to various regularizers, thereby facilitating the solution of a broad class of nonsmooth optimization problems.

To proceed, the notion of a smoothing function for a nonsmooth function is formally defined as follows.

Definition 1.1 [20,6] Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. A function $\phi : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}$ is a smoothing function of f if it satisfies the following properties:

(i) $\phi(\cdot, \tau)$ is continuously differentiable for any $\tau > 0$;
(ii) $\lim_{z\to y, \tau\to 0^+} \phi(z, \tau) = f(y)$ for any $y \in \mathbb{R}^n$.

It is well known that two principal techniques have been developed for constructing smoothing functions for the plus function: convolution-based smoothing [11,5,20] and infimal-convolution smoothing [1]. Beyond these, various other methods have also been proposed in the literature [32,26,29]. This paper introduces new smoothing techniques that offer a simpler approach to constructing smoothing functions for the plus function than existing methods. The central idea is to employ continuous functions that approximate the step function. In particular, we propose two distinct approaches for constructing such approximations:

1. **Polynomial-based approximation:** Continuous approximations of the step function are constructed using polynomial functions derived from Bézier curves.
2. **Dirac delta-based approximation:** By leveraging the intrinsic connection between the step function and the Dirac delta function, we derive continuous approximations from classical smooth approximations of the Dirac delta function.

As a brief review of smoothing functions for widely used regularizers, early work on the smoothed ℓ_0-norm appeared in [14,15], where smoothing functions were derived from approximation functions of the Kronecker delta. However, these approximations were not explored in depth, and only a few illustrative examples were provided. Smoothing functions for the MCP and SCAD regularizers were studied in [7], based on smoothing functions of the plus function. In particular, their formulations involve integrals of the smoothed plus function, which can present difficulties in numerical implementation. Specific smoothing techniques for the pinball function have also been proposed, as discussed in [34,31]. Meanwhile, the ℓ_p-norm with $0 < p \leq 1$ has been the subject of various smoothing approaches, including those found in [6,30].

It is important to note that regularizers play a pivotal role in sparse optimization, as many models can be reformulated through the incorporation of suitable regularization terms. In Section 3, we demonstrate how several sparse optimization models can be cast into forms involving regularizers. This reformulation facilitates the use of fast smoothing-based methods such as Newton, quasi-Newton (e.g., BFGS), and conjugate gradient methods to efficiently solve the resulting smooth approximations of originally nonsmooth problems.

In this work, we uncover a fundamental connection between a range of regularizers and the step function. In particular, we construct smoothing functions for the ℓ_0-norm and the plus function by exploiting their intrinsic links to the step function. Furthermore, since the MCP and SCAD regularizers can be expressed in terms of absolute value functions, their smoothing functions can be naturally derived from those of the absolute value function. We also establish a general form for smoothing functions of the pinball function by leveraging its relationship with the plus function. Similarly, smoothing functions for the ℓ_p-norm with $0 < p \leq 1$ are readily obtained using our framework, which is again built upon smooth approximations of the absolute value function.

These findings demonstrate that our approach offers a unified and systematic framework for constructing smoothing functions across a broad class of regularizers. This framework not only simplifies numerical implementation by using C^1 or C^2 smoothing functions, but also ensures desirable theoretical properties, including provable error bounds with respect to the plus function, gradient boundedness, and gradient consistency. These properties are critical in establishing convergence results for smooth approximations, thereby making our technique a powerful and reliable tool for solving nonsmooth optimization problems. The overall structure of the smoothing functions for various regularizers is illustrated in Figure 1.

2 How to Construct Smoothing Functions

2.1 Smoothing approximation of step function

In this section, we construct continuous approximation, potentially smoothing, functions of the step function, derived from polynomial functions and the Dirac delta function.

Definition 2.1 $\sigma(x)$ is called the step function if

$$\sigma(x) = \begin{cases} 1 \text{ if } x > 0, \\ 0 \text{ if } x \leq 0. \end{cases}$$

Since $\sigma(x)$ is discontinuous, Definition 1.1 does not apply directly for defining its smoothing function. Instead, we introduce the notion of a continuous approximation for a discontinuous function, as formally defined below.

Definition 2.2 Let $f : \mathbb{R} \to \mathbb{R}$ be a discontinuous function. A function $\psi : \mathbb{R} \times \mathbb{R}_{++} \to \mathbb{R}$ is referred to as a continuous approximation function of f if it adheres to the following conditions:

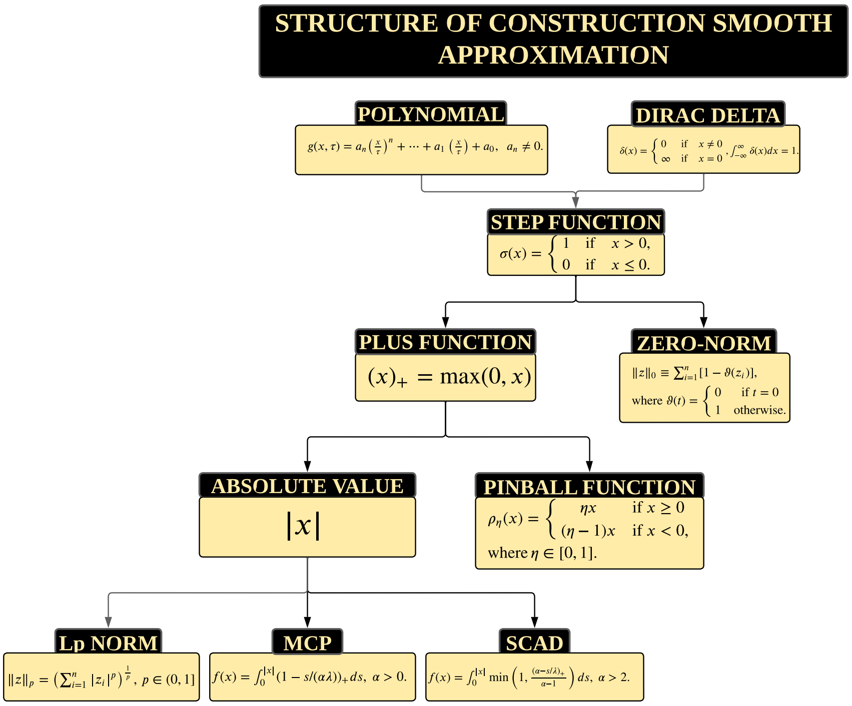


Fig. 1: Structure of Smoothing Functions of Regularizers.

(i) $\psi(\cdot,\tau)$ is continuous for any $\tau > 0$;
(ii) $\lim_{\tau\to 0^+} \psi(x,\tau) = f(x)$ almost everywhere.

Moreover, if $\psi(x,\tau)$ is C^1, then $\psi(x,\tau)$ is a C^1-approximation function of $f(x)$.

Definition 2.3 Let $f : \mathbb{R} \to \mathbb{R}$ be a discontinuous function. A function $\psi : \mathbb{R} \times \mathbb{R}_{++} \to \mathbb{R}$ is referred to as a smoothing function of f if it adheres to the following conditions:

(i) $\psi(\cdot,\tau)$ is continuously differentiable for any $\tau > 0$;
(ii) $\lim_{\tau\to 0^+} \psi(x,\tau) = f(x)$ for any $x \in \mathbb{R}$.

2.1.1 Approximation via polynomial

Inspired by the Bézier curve, specifically, the fact that it always passes through the first and last points of a given set of control points. We are motivated to construct continuous approximation functions for $\sigma(x)$ using Bernstein polynomials. To that end, we begin by briefly introducing the formulation of a Bézier curve.

Definition 2.4 Consider a set of $n+1$ control points $P_0,\ldots,P_n$. A Bézier curve based on these points is defined by

$$B(t) = \sum_{i=0}^{n} \binom{n}{k} (1-t)^{n-i} t^k P_k \quad 0 \le t \le 1,$$

where $\binom{n}{k}$ are the binomial coefficients defined by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

In particular, we will describe two examples using polynomial with degrees 1 and 2 to construct a continuous approximation function for $\sigma(x)$ based on a Bézier curve. Given two points $P_0 = (-\tau, 0)$ and $P_1 = (\tau, 1)$ with $\tau > 0$, a linear Bézier curve is given by

$$B(t) = (1-t)P_0 + tP_1 = (-\tau(1-t) + \tau t, t),\ 0 \le t \le 1.$$

A corresponding parametric equation:

$$\begin{cases} x(t) = -\tau(1-t) + \tau t \\ y(t) = t \end{cases} \iff \begin{cases} \frac{x(t)}{\tau} = -(1-t) + t = 2t - 1 \\ y(t) = t. \end{cases}$$

Hence,

$$y = \frac{1}{2}\left(\frac{x}{\tau}\right) + \frac{1}{2}$$

is a straight line through $P_0 = (-\tau, 0)$ and $P_1 = (\tau, 1)$. Then a corresponding continuous approximation function of $\sigma(x)$ is

$$\psi(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \frac{1}{2}\left(\frac{x}{\tau}\right) + \frac{1}{2} & \text{if } |x| < \tau \\ 1 & \text{if } x > \tau. \end{cases}$$

Given three points $P_0 = (-\tau, 0)$, $P_1 = (0, 0)$, $P_2 = (\tau, 1)$ with $\tau > 0$, a quadratic Bézier curve can be defined as follows

$$\begin{aligned} B(t) &= (1-t)^2 P_0 + 2(1-t)tP_1 + t^2 P_2 \\ &= \left(\tau\left(t^2 - (1-t)^2\right), t^2\right),\ 0 \le t \le 1. \end{aligned}$$

A corresponding parametric equation:

$$\begin{cases} \frac{x(t)}{\tau} = t^2 - (1-t)^2 \\ y(t) = t^2. \end{cases}$$

Hence,

$$y = \frac{1}{4}\left(\frac{x}{\tau}\right)^2 + \frac{1}{2}\left(\frac{x}{\tau}\right) + \frac{1}{4}$$

is a quadratic curve through $P_0 = (-\tau, 0)$, $P_2 = (\tau, 1)$. Then a corresponding continuous approximation function of $\sigma(x)$ is

$$\psi(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \frac{1}{4}\left(\frac{x}{\tau}\right)^2 + \frac{1}{2}\left(\frac{x}{\tau}\right) + \frac{1}{4} & \text{if } |x| \le \tau \\ 1 & \text{if } x > \tau. \end{cases}$$

From the above examples, we observe that if we take $n+1$ points $P_0, \dots, P_n$ with $P_0 = (-\alpha\tau, 0)$, $P_n = (\beta\tau, 1)$ and other points P_i, $i = 1, \dots, n-1$, then there is always a corresponding polynomial degree n which is a curve through two points P_0 and P_n. This leads us to define a form of continuous approximation functions of $\sigma(x)$ as follows. For any $\tau > 0$,

$$\psi(x,\tau) = \begin{cases} 0 & \text{if } x < -\alpha\tau \\ g(x,\tau) & \text{if } -\alpha\tau \le x \le \beta\tau \\ 1 & \text{if } x > \beta\tau, \end{cases} \tag{2}$$

where $g(x,\tau)$ is a polynomial degree n of $\frac{x}{\tau}$, $n \geq 1$, i.e.,

$$g(x,\tau) = a_n \left(\frac{x}{\tau}\right)^n + \cdots + a_1 \left(\frac{x}{\tau}\right) + a_0, \quad a_n \neq 0, \tag{3}$$

satisfying

$$\begin{cases} g(-\alpha\tau,\tau) = 0 \\ g(\beta\tau,\tau) \ \ = 1 \end{cases} \tag{4}$$

with

$$\begin{cases} \alpha \geq 0 \\ \beta > 0 \end{cases} \text{ or } \begin{cases} \alpha > 0 \\ \beta \geq 0. \end{cases} \tag{5}$$

Theorem 2.1 *Suppose that $\psi(x,\tau)$ given by (2). Then $\psi(x,\tau)$ is continuous approximation of $\sigma(x)$. Consequently, $\psi(x,\tau)$ acts as a smoothing function of $\sigma(x)$ if it is C^1 and $a_0 = 0$.*

Proof From (2)-(5), we have that $\psi(\tau,\cdot)$ is continuous for any $\tau > 0$. If $x = 0$, then $\psi(x,\tau) = g(0,\tau) = a_0$ for any $\tau > 0$. Otherwise, if $x \neq 0$, there exists a sufficiently small value τ such that

$$\begin{cases} x < -\tau\alpha \text{ if } x < 0 \\ x > \tau\beta \quad \text{if } x > 0. \end{cases}$$

These imply

$$\lim_{\tau\to 0^+} \psi(x,\tau) = \begin{cases} \sigma(x) \text{ if } x \neq 0 \\ \ a_0 \quad \text{if } x = 0. \end{cases}$$

Thus, $\psi(x,\tau)$ sever as a continuous approximation of $\sigma(x)$. According to Definition 2.3, it is evident that $\psi(x,\tau)$ becomes a smoothing function of $\sigma(x)$ when $\psi(x,\tau)$ is C^1 and $a_0 = 0$. □

We will now construct continuous approximation functions of $\sigma(x)$ by using $\psi(x,\tau)$ as defined in (2). For $n = 1$, $g(x,\tau)$ is a first degree polynomial, that is

$$g(x,\tau) = a_1 \left(\frac{x}{\tau}\right) + a_0.$$

It is from the condition (4) that

$$\begin{cases} g(-\alpha\tau,\tau) = 0 \\ g(\beta\tau,\tau) \ \ = 1 \end{cases} \iff \begin{cases} -\alpha a_1 + a_0 = 0 \\ \beta a_1 + a_0 \ \ = 1 \end{cases} \iff \begin{cases} a_1 = \frac{1}{\alpha+\beta} \\ a_0 = \frac{\alpha}{\alpha+\beta}. \end{cases}$$

For example, we have the following continuous approximation functions of $\sigma(x)$.

$$\psi_1(x,\tau) = \begin{cases} 0 \text{ if } x < 0 \\ \frac{x}{\tau} \text{ if } 0 \leq x \leq \tau \\ 1 \text{ if } x > \tau \end{cases} \text{ with } \alpha = 0 \text{ and } \beta = 1. \tag{6}$$

$$\psi_2(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \frac{x}{2\tau} + \frac{1}{2} & \text{if } |x| \leq \tau \\ 1 & \text{if } x > \tau \end{cases} \text{ with } \alpha = \beta = 1.$$

For $n = 2$, i.e., $g(x,\tau)$ is a polynomial degree 2, we have

$$g(x,\tau) = a_2 \left(\frac{x}{\tau}\right)^2 + a_1 \left(\frac{x}{\tau}\right) + a_0,$$

where a_0, a_1, a_2 satisfy

$$\begin{cases} \alpha^2 a_2 - \alpha a_1 + a_0 = 0 \\ \beta^2 a_2 + \beta a_1 + a_0 = 1. \end{cases}$$

For example, we have the following continuous approximation functions of $\sigma(x)$.

$$\psi_3(x,\tau)=\begin{cases}0 & \text{if } x<0\\ \lambda\left(\frac{x}{\tau}\right)^2+(1-\lambda)\left(\frac{x}{\tau}\right) & \text{if } 0\le x\le\tau\\ 1 & \text{if } x>\tau\end{cases}\quad \text{with } \alpha=0 \text{ and } \beta=1,\ a_2=\lambda\in\mathbb{R}.$$

$$\psi_4(x,\tau)=\begin{cases}0 & \text{if } x<-\tau\\ \lambda\left(\frac{x}{\tau}\right)^2+\frac{1}{2}\left(\frac{x}{\tau}\right)+\frac{1}{2}-\lambda & \text{if } |x|\le\tau\\ 1 & \text{if } x>\tau\end{cases}\quad \text{with } \alpha=\beta=1,\ a_2=\lambda\in\mathbb{R}.$$

Let $\lambda=\frac{1}{3}$, the graphs representing these functions are illustrated in Figure 2.

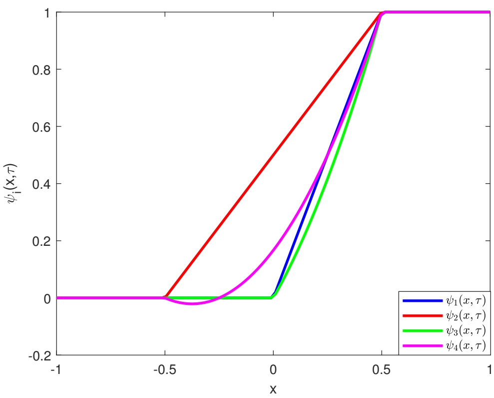


Fig. 2: Graphs of $\psi_i(x,\tau)$ with $\tau=0.5$.

Remark 2.1 For $n=1$ and $n=2$, the function $\psi(x,\tau)$ given by relation (2) serves as a continuous approximation to $\sigma(x)$, but it is not a C^1-approximation. We propose another way to construct a C^1-continuous approximation of $\sigma(x)$ when $n=2$ by modifying the form of $\psi(x,\tau)$. In particular, we consider a form of $\psi(x,\tau)$ as follows.

$$\psi(x,\tau)=\begin{cases}0 & \text{if } x<-\alpha\tau\\ g_1(x,\tau) & \text{if } -\alpha\tau\le x\le 0\\ g_2(x,\tau) & \text{if } 0\le x\le\beta\tau\\ 1 & \text{if } x>\beta\tau,\end{cases}$$

where $\alpha>0,\beta>0$ and

$$g_1(x,\tau)=a_2\left(\frac{x}{\tau}\right)^2+a_1\left(\frac{x}{\tau}\right)+c \quad\text{and}\quad g_2(x,\tau)=b_2\left(\frac{x}{\tau}\right)^2+b_1\left(\frac{x}{\tau}\right)+c.$$

We wish $\psi(x,\tau)$ is C^1. Then the following conditions must hold.

$$\begin{cases}g_1(-\alpha\tau,\tau)=0\\ g_2(\beta\tau,\tau)\quad=1\\ g_1'(-\alpha\tau,\tau)=0\\ g_1'(0,\tau)\quad=g_2'(0,\tau)\\ g_2'(\beta\tau,\tau)\quad=0.\end{cases}$$

Therefore,

$$\begin{cases} a_2 = \frac{1}{\alpha(\alpha+\beta)} \\ b_2 = -\frac{1}{\alpha(\alpha+\beta)} \\ a_1 = b_1 = \frac{2}{\alpha+\beta} \\ c \ = \frac{\alpha}{\alpha+\beta}. \end{cases}$$

Let $\alpha = \beta = 1$, we have a corresponding C^1-approximation function of $\sigma(x)$ as follows.

$$\psi_5(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \frac{1}{2}\left(\frac{x}{\tau}\right)^2 + \frac{x}{\tau} + \frac{1}{2} & \text{if } -\tau \le x \le 0 \\ -\frac{1}{2}\left(\frac{x}{\tau}\right)^2 + \frac{x}{\tau} + \frac{1}{2} & \text{if } 0 \le x \le \tau \\ 1 & \text{if } x > \tau \end{cases}$$

$$= \begin{cases} 0 & \text{if } x < -\tau \\ \frac{(x+\tau)^2}{2\tau^2} & \text{if } -\tau \le x \le 0 \\ 1 - \frac{(x-\tau)^2}{2\tau^2} & \text{if } 0 \le x \le \tau \\ 1 & \text{if } x > \tau. \end{cases} \tag{7}$$

The following theorem establishes C^1-approximation functions of $\sigma(x)$ with polynomial $g(x,\tau)$ degree $n \ge 3$.

Theorem 2.2 *Suppose that $\psi(x,\tau)$ is given by (2). Then for $n \ge 3$, $\psi(x,\tau)$ is a C^1-approximation function of $\sigma(x)$ if coefficients a_i $(i = 0, 1, \ldots, n)$ of polynomial $g(x,\tau)$ satisfy the following system of linear equations*

$$Ax = b, \tag{8}$$

where A is the $4 \times (n+1)$ matrix satisfying

$$A = \begin{cases} \begin{pmatrix} -\alpha^n & \alpha^{n-1} & \ldots & -\alpha & 1 \\ \beta^n & \beta^{n-1} & \ldots & \beta & 1 \\ n\alpha^{n-1} & -(n-1)\alpha^{n-2} & \ldots & 1 & 0 \\ n\beta^{n-1} & (n-1)\beta^{n-2} & \ldots & 1 & 0 \end{pmatrix} & \text{if } n \text{ is odd}, \\ \begin{pmatrix} \alpha^n & -\alpha^{n-1} & \ldots & -\alpha & 1 \\ \beta^n & \beta^{n-1} & \ldots & \beta & 1 \\ -n\alpha^{n-1} & (n-1)\alpha^{n-2} & \ldots & 1 & 0 \\ n\beta^{n-1} & (n-1)\beta^{n-2} & \ldots & 1 & 0 \end{pmatrix} & \text{if } n \text{ is even}, \end{cases} \tag{9}$$

and

$$x = (a_n, \ldots, a_0)^T \in \mathbb{R}^{n+1}, \quad b = (0, 1, 0, 0)^T. \tag{10}$$

Proof Calculating the derivative of $g(x,\tau)$ given by (3), we have

$$g'(x,\tau) = na_n \frac{x^{n-1}}{\tau^n} + \cdots + a_1 \frac{1}{\tau}.$$

Then for $n \ge 3$, $\psi(x,\tau)$ is a C^1-approximation function of $\sigma(x)$ if the following conditions hold:

$$\begin{cases} g(-\alpha\tau,\tau) = 0, \quad g(\beta\tau,\tau) = 1 \\ g'(-\alpha\tau,\tau) = g'(\beta\tau,\tau) = 0. \end{cases}$$

If n is odd, then the coefficients a_i $(i = 0, 1, \ldots, n)$ of polynomial $g(x, \tau)$ are a solution of the following equations

$$\begin{cases} -\alpha^n a_n + \alpha^{n-1} a_{n-1} + \cdots - \alpha a_1 + a_0 &= 0 \\ \beta^n a_n + \beta^{n-1} a_{n-1} + \cdots + \beta a_1 + a_0 &= 1 \\ n\alpha^{n-1} a_n - (n-1)\alpha^{n-2} a_{n-1} + \cdots + a_1 &= 0 \\ n\beta^{n-1} a_n + (n-1)\beta^{n-2} a_{n-1} + \cdots + a_1 &= 0 \end{cases} \Longleftrightarrow Ax = b,$$

where A, x, b are given by (9)-(10). Similar to the case n even. □

Note that the system equations (8) has a unique solution when $n = 3$ due to

$$\det A = -\frac{(\alpha+\beta)(2\alpha^4 + 9\alpha^3\beta + 15\alpha^2\beta^2 + 11\alpha\beta^3 + 3\beta^4)}{2\alpha + 3\beta} \neq 0$$

with α, β satisfying (5). For the case $n > 3$, the system (8) has infinitely many solutions which says that there are a lot of C^1-approximation functions of $\sigma(x)$ if $n > 3$. For example, let $n = 3, 4, 5$ and $\alpha = \beta = 1$, the following C^1-approximation functions of $\sigma(x)$ can be readily obtained.

$$\psi_6(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ -\frac{1}{4}\left(\frac{x}{\tau}\right)^3 + \frac{3}{4}\left(\frac{x}{\tau}\right) + \frac{1}{2} & \text{if } |x| \leq \tau \\ 1 & \text{if } x > \tau. \end{cases} \tag{11}$$

$$\psi_7(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \lambda\left(\frac{x}{\tau}\right)^4 - \frac{1}{4}\left(\frac{x}{\tau}\right)^3 - 2\lambda\left(\frac{x}{\tau}\right)^2 + \frac{3}{4}\left(\frac{x}{\tau}\right) + \frac{1}{2} + \lambda & \text{if } |x| \leq \tau \\ 1 & \text{if } x > \tau \end{cases} \quad \text{for any } \lambda \in \mathbb{R}. \tag{12}$$

$$\psi_8(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \lambda\left(\frac{x}{\tau}\right)^5 + \gamma\left(\frac{x}{\tau}\right)^4 - \left(\frac{1}{4} + 2\lambda\right)\left(\frac{x}{\tau}\right)^3 - 2\gamma\left(\frac{x}{\tau}\right)^2 + \left(\frac{3}{4} + \lambda\right)\left(\frac{x}{\tau}\right) + \frac{1}{2} + \gamma & \text{if } |x| \leq \tau \\ 1 & \text{if } x > \tau \end{cases}$$

for any $\lambda, \gamma \in \mathbb{R}$. Note that $\psi_6(x,\tau)$ and $\psi_8(x,\tau)$ with $\lambda = \frac{3}{16}, \gamma = 0$ are called the Tishler-Zang C^1-approximation functions of $\sigma(x)$ in [23]. Let $\lambda = \frac{1}{2}, \gamma = 0$, the graphs representing these functions are illustrated in Figure 3.

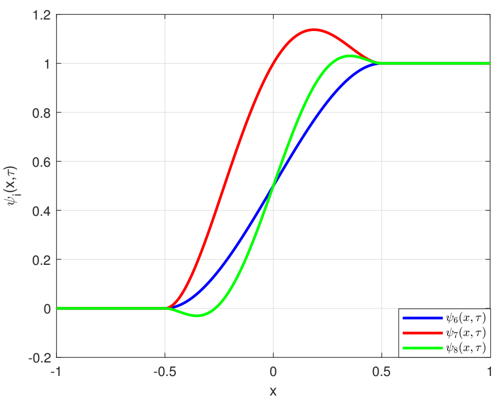


Fig. 3: Graphs of $\psi_i(x,\tau)$ with $\tau = 0.5$.

2.1.2 Approximation via Dirac delta function

In this subsection, we develop continuous approximation functions for $\sigma(x)$ based on the Dirac delta function. Notably, the step function $\sigma(x)$ can be represented as the integral of the Dirac delta function, as follows:

$$\sigma(x) = \int_{-\infty}^{x} \delta(s)ds,$$

where $\delta(s)$ represents the Dirac delta function, which satisfies the following properties:

$$\delta(s) = \begin{cases} 0 & \text{if } s \neq 0 \\ \infty & \text{if } s = 0 \end{cases}, \int_{-\infty}^{\infty} \delta(s)ds = 1.$$

To approximate $\delta(s)$, let $\rho(s)$ be a piecewise continuous density function with a finite number of pieces, satisfying

$$\rho(s) \geq 0, \qquad \int_{-\infty}^{+\infty} \rho(s)ds = 1.$$

Then

$$\hat{t}(x,\tau) := \frac{1}{\tau}\rho\left(\frac{x}{\tau}\right) \approx \delta(x). \tag{13}$$

Theorem 2.3 *Suppose that $\rho(x)$ is a piecewise continuous density function. Then*

$$\psi(x,\tau) = \int_{-\infty}^{x} \hat{t}(s,\tau)ds \tag{14}$$

is a continuous approximation function of $\sigma(x)$, where $\hat{t}(s,\tau)$ is defined as in (13). Moreover, if $\rho(x)$ is continuous, then $\psi(x,\tau)$ is a C^1-approximation function of $\sigma(x)$.

Proof Since $\rho(x)$ is piecewise continuous then $\psi(x,\tau)$ is continuous. On the other hand, we have

$$\lim_{\tau\to 0^+} \psi(x,\tau) = \lim_{\tau\to 0^+} \int_{-\infty}^{\frac{x}{\tau}} \rho(s)ds = \begin{cases} \int_{-\infty}^{0} \rho(s)ds & \text{if } x = 0 \\ \sigma(x) & \text{if } x \neq 0. \end{cases} \tag{15}$$

Hence, $\psi(x,\tau)$ serves as a continuous approximation of $\sigma(x)$. If $\rho(x)$ is continuous, then its derivative $\psi'(x,\tau) = \hat{t}(x,\tau)$ is also continuous. Consequently, $\psi(x,\tau)$ is a C^1-approximation of the step function $\sigma(x)$. □

To illustrate, we present several examples of ρ, a continuous density function, as follows:

$$\begin{aligned}
\rho_1(s) &= \frac{1}{2(s^2+1)^{3/2}}, \\
\rho_2(s) &= \frac{e^{-s}}{(1+e^{-s})^2}, \\
\rho_3(s) &= \frac{1}{\pi(1+s^2)}, \\
\rho_4(s) &= \frac{e^{-s^2}}{\sqrt{\pi}}, \\
\rho_5(s) &= \frac{1}{2}\left(1-\tanh^2(s)\right), \\
\rho_6(s) &= \frac{2}{\pi(e^s+e^{-s})}.
\end{aligned}$$

Applying the above theorem, we obtain the corresponding C^1-approximation functions of $\sigma(x)$ as below:

$$\psi_1(x,\tau) = \frac{1}{2}\left(\frac{x}{\sqrt{x^2+\tau^2}} + 1\right), \tag{16}$$

$$\psi_2(x,\tau) = \frac{e^{\frac{x}{\tau}}}{e^{\frac{x}{\tau}}+1}, \tag{17}$$

$$\psi_3(x,\tau) = \frac{1}{\pi}\left(\arctan\left(\frac{x}{\tau}\right) + \frac{\pi}{2}\right), \tag{18}$$

$$\psi_4(x,\tau) = \frac{1}{2}\left(\mathrm{erf}(\frac{x}{\tau}) + 1\right), \tag{19}$$

$$\psi_5(x,\tau) = \frac{1}{2}\left(\tanh(\frac{x}{\tau}) + 1\right), \tag{20}$$

$$\psi_6(x,\tau) = \frac{2}{\pi}\arctan(e^{\frac{x}{\tau}}), \tag{21}$$

where the error function is defined by

$$\mathrm{erf}(t) = \frac{2}{\sqrt{\pi}}\int_0^t e^{-u^2}du \quad \forall t \in \mathrm{I\!R}.$$

Moreover, let ρ be the piecewise continuous density function given by

$$\rho_7(s) = \begin{cases} 1 \text{ if } 0 \le s \le 1 \\ 0 \text{ otherwise,} \end{cases}$$

$$\rho_8(s) = \begin{cases} \frac{\pi}{4}\cos\left(\frac{\pi s}{2}\right) & \text{if } |s| \le 1 \\ 0 & \text{if } |s| > 1, \end{cases}$$

$$\rho_9(s) = \begin{cases} \frac{3}{4}\left(1-s^2\right) & \text{if } |s| \le 1 \\ 0 & \text{if } |s| > 1. \end{cases}$$

Then, the corresponding continuous approximation functions of $\sigma(x)$ are

$$\psi_7(x,\tau) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x}{\tau} & \text{if } 0 \le x \le \tau \\ 1 & \text{if } x > \tau, \end{cases}$$

$$\psi_8(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \frac{1}{2}\left(\sin\left(\frac{\pi}{2}\frac{x}{\tau}\right) + 1\right) & \text{if } |x| \le \tau \\ 1 & \text{if } x > \tau, \end{cases} \tag{22}$$

$$\psi_9(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ -\frac{x^3}{4\tau^3} + \frac{3}{4}\frac{x}{\tau} + \frac{1}{2} & \text{if } |x| \le \tau \\ 1 & \text{if } x > \tau. \end{cases}$$

The graphs representing these functions are depicted in Figure 4.

We observe that when piecewise continuous density functions are used, the resulting continuous approximation functions $\psi(x,\tau)$ for $\sigma(x)$ share the same structure as the form given in (2), as seen in examples such as $\psi_7(x,\tau)$ and $\psi_9(x,\tau)$. Additionally, $\psi_8(x,\tau)$ also exhibits a piecewise continuous form analogous to that of $\psi(x,\tau)$ in (2), where the function $g(x,\tau) := \frac{1}{2}\left(\sin\left(\frac{\pi}{2}\frac{x}{\tau}\right) + 1\right)$ satisfies the condition (4). Based on these observations, we propose the following general form of continuous approximation functions $\psi(x,\tau)$ for the step function $\sigma(x)$:

$$\psi(x,\tau) = \begin{cases} 0 & \text{if } x < -\alpha\tau \\ g(x,\tau) & \text{if } -\alpha\tau \le x \le \beta\tau \\ 1 & \text{if } x > \beta\tau, \end{cases} \tag{23}$$

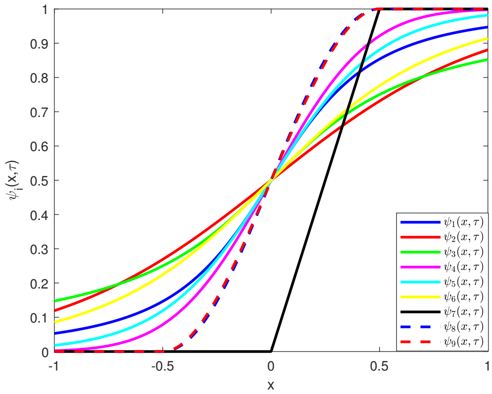


Fig. 4: Graphs of $\psi_i(x,\tau)$ with $\tau = 0.5$.

where $g : \mathbb{R} \to \mathbb{R}$ is continuous and $g(x,\tau) := g(\frac{x}{\tau})$ satisfies

$$\begin{cases} g(-\alpha\tau,\tau) = 0 \\ g(\beta\tau,\tau) \ \ = 1 \end{cases} \tag{24}$$

with

$$\begin{cases} \alpha \geq 0 \\ \beta > 0 \end{cases} \text{ or } \begin{cases} \alpha > 0 \\ \beta \geq 0. \end{cases}$$

2.2 Smoothing approximation of plus function

Definition 2.5 The mapping $(x)_+ : \mathbb{R} \to \mathbb{R}_+$ is called the plus function if

$$(x)_+ = \max(0, x).$$

Note that there are close connections between $(x)_+$ and $\sigma(x)$, as described below:

$$(x)_+ = \int_{-\infty}^{x} \sigma(t)dt \quad \text{and} \quad (x)_+ = x\sigma(x).$$

To proceed, we utilize the identity $(x)_+ = \int_{-\infty}^{x} \sigma(t)dt$ to construct smoothing functions for $(x)_+$, as formally stated in the following theorem.

Theorem 2.4 *Suppose that $\psi(x,\tau)$ is defined as in (23)-(24). Let*

$$\phi^{\text{plus}}(x,\tau) = \int_{-\infty}^{x} \psi(t,\tau)dt.$$

(i) For fixed $\tau > 0$

$$-\tau \left(\int_{-\alpha}^{\beta} |g(t)|dt + \beta \right) \leq \phi^{\text{plus}}(x,\tau) - (x)_+ \leq \tau \int_{-\alpha}^{\beta} |g(t)|dt.$$

(ii) $\phi^{\text{plus}}(x,\tau)$ *is a smoothing function of* $(x)_+$.
(iii) For any x, $\left(\phi^{\text{plus}}(x,\tau)\right)'_x$ *is bounded.*
(iv) If $g(x,\tau)$ *is increasing on* $[-\alpha\tau,\beta\tau]$ *then*

$$\left\{\lim_{x\to\bar{x},\tau\to 0^+}\left(\phi^{\text{plus}}(x,\tau)\right)'_x\right\}\subset\partial(\bar{x})_+=\begin{cases}1 & \text{if } \bar{x}>0\\ 0 & \text{if } \bar{x}<0\\ [0,1] & \text{if } \bar{x}=0.\end{cases}$$

This relation is called the gradient consistency property.

Proof (i) For $x<-\alpha\tau$, it is clear that $\phi^{\text{plus}}(x,\tau)-(x)_+=0$. If $x>\beta\tau$ then

$$\phi^{\text{plus}}(x,\tau)-(x)_+=\int_{-\alpha\tau}^{\beta\tau}g(t,\tau)dt+\int_{\beta\tau}^{x}1dt-x=\tau\int_{-\alpha}^{\beta}g(t)dt-\beta\tau.$$

Since g is continuous and $-|g|\le g\le|g|$, we have

$$-\tau\left(\int_{-\alpha}^{\beta}|g(t)|dt+\beta\right)\le\phi^{\text{plus}}(x,\tau)-(x)_+\le\tau\int_{-\alpha}^{\beta}|g(t)|dt.$$

If $0\le x\le\beta\tau$, then

$$\phi^{\text{plus}}(x,\tau)-(x)_+=\int_{-\alpha\tau}^{x}g(t,\tau)dt-x.$$

Hence

$$-\tau\left(\int_{-\alpha}^{\beta}|g(t)|dt+\beta\right)\le\int_{-\alpha\tau}^{x}g(t,\tau)dt-\beta\tau\le x\le\phi^{\text{plus}}(x,\tau)-(x)_+\le\int_{-\alpha\tau}^{x}g(t,\tau)dt\le\tau\int_{-\alpha}^{\beta}|g(t)|dt.$$

If $-\alpha\tau\le x\le 0$, then

$$\phi^{\text{plus}}(x,\tau)-(x)_+=\int_{-\alpha\tau}^{x}g(t,\tau)dt.$$

Hence

$$-\tau\int_{-\alpha}^{\beta}|g(t)|dt\le\phi^{\text{plus}}(x,\tau)-(x)_+\le\tau\int_{-\alpha}^{\beta}|g(t)|dt. \tag{25}$$

From all above, we obtain the desired inequality.

(ii) Since ψ is continuous, ϕ^{plus} is continuously differentiable. From the inequality (25) we have that

$$\lim_{x\to\bar{x},\tau\to 0^+}\phi^{\text{plus}}(x,\tau)=(\bar{x})_+,$$

which says $\phi^{\text{plus}}(x,\tau)$ is a smoothing function of $(x)_+$.

(iii) Since g is continuous on $[-\alpha\tau,\beta\tau]$, g is bounded on $[-\alpha\tau,\beta\tau]$. Hence ψ is also bounded. Moreover, it is clear that $\left(\phi^{\text{plus}}(x,\tau)\right)'_x=\psi(x,\tau)$.

(iv) Since g is increasing on $[-\alpha\tau,\beta\tau]$ and using the condition (24), it implies that $0\le g\le 1$ which yields $0\le\left(\phi^{\text{plus}}(x,\tau)\right)'_x\le 1$. Thus, we have

$$\left\{\lim_{x\to\bar{x},\tau\to 0^+}\left(\phi^{\text{plus}}(x,\tau)\right)'_x\right\}\subset\partial(\bar{x})_+.$$

□

Let $\psi(x,\tau)$ be as defined in (6)-(7). Applying Theorem 2.4, we obtain the corresponding smoothing functions for $(x)_+$ as follows.

$$\phi_1^{\text{plus}}(x,\tau) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x^2}{2\tau} & \text{if } 0 \le x \le \tau \\ x - \frac{\tau}{2} & \text{if } x > \tau. \end{cases}$$

$$\phi_2^{\text{plus}}(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \frac{x^2}{4\tau} + \frac{x}{2} + \frac{\tau}{4} & \text{if } |x| \le \tau \\ x & \text{if } x > \tau. \end{cases} \tag{26}$$

$$\phi_3^{\text{plus}}(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \frac{(x+\tau)^3}{6\tau^2} & \text{if } -\tau \le x \le 0 \\ x - \frac{(x-\tau)^3}{6\tau^2} & \text{if } 0 \le x \le \tau \\ x & \text{if } x > \tau \end{cases}$$

Note that $\phi_1^{\text{plus}}(x,\tau)$ corresponds to the Pinar-Zenios smoothing function [18] for $(x)_+$, $\phi_2^{\text{plus}}(x,\tau)$ is the Zang smoothing function [32], and $\phi_3^{\text{plus}}(x,\tau)$ is a C^2 function known as the Ye-Liu-Zhou-Liu smoothing function [28]. The graphs of these smoothing functions are depicted in Figure 5.

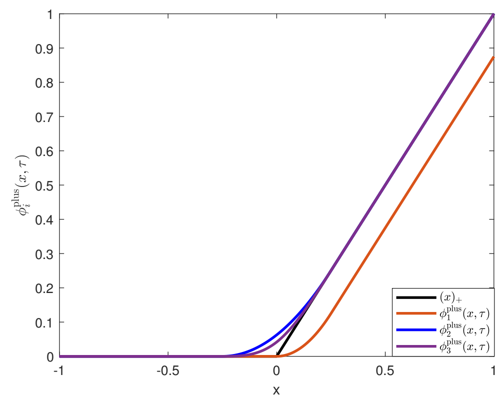


Fig. 5: Graphs of $(x)_+$ and $\phi_i^{\text{plus}}(x,\tau)$ with $\tau = 0.25$.

Next, we use the relation $(x)_+ = x\sigma(x)$ to generate smoothing functions for $(x)_+$ by the following theorems.

Theorem 2.5 *Suppose that $\psi(x,\tau)$ given by (23)-(24) is a C^1-approximation function of $\sigma(x)$. Let*

$$\phi^{\text{plus}}(x,\tau) = x\psi(x,\tau).$$

Then

(i) For fixed $\tau > 0$, there exists $M > 0$ satisfying

$$\left|\phi^{\text{plus}}(x,\tau) - (x)_+\right| \le M\tau.$$

(ii) $\phi^{\text{plus}}(x,\tau)$ is a smoothing functions of $(x)_+$.

(iii) For any x, $\left(\phi^{\text{plus}}(x,\tau)\right)'_x$ is bounded
(iv) For any x,

$$\left\{\lim_{x\to\bar{x},\tau\to 0^+}\left(\phi^{\text{plus}}(x,\tau)\right)'_x\right\}\subset\partial(\bar{x})_+,\quad \forall\,\bar{x}\neq 0.$$

Proof (i) If $x<-\alpha\tau$ or $x>\beta\tau$, then $\phi^{\text{plus}}(x,\tau)-(x)_+=0$. Since g is bounded on $[-\alpha\tau,\beta\tau]$ then there exists $M_0>0$ such that

$$|g(x,\tau)|\leq M_0.$$

If $-\alpha\tau\leq x\leq 0$

$$\left|\phi^{\text{plus}}(x,\tau)-(x)_+\right|=|xg(x,\tau)|\leq|x||g(x,\tau)|\leq\alpha\tau M_0.$$

If $0\leq x\leq\beta\tau$

$$\left|\phi^{\text{plus}}(x,\tau)-(x)_+\right|=|x\left(g(x,\tau)-1\right)|\leq|x||g(x,\tau)-1|\leq\beta\tau(M_0+1).$$

Set $M=\max\{\alpha M_0,\beta(M_0+1)\}$. We have

$$\left|\phi^{\text{plus}}(x,\tau)-(x)_+\right|\leq M\tau.$$

(ii) The result is obtained immediately from (i).

(iii) First, we observe that

$$\left(\phi^{\text{plus}}(x,\tau)\right)'_x=\psi(x,\tau)+x\psi'(x,\tau).$$

Since $\psi(x,\tau)$ is C^1-approximation function of $\sigma(x)$, $\psi'(x,\tau)$ is bounded. Hence, $\left(\phi^{\text{plus}}(x,\tau)\right)'_x$ is also bounded.

(iv) If $\bar{x}\neq 0$ then for sufficiently small τ, we have either $x>\beta\tau$ or $x<-\alpha\tau$. This implies that

$$\lim_{x\to\bar{x},\tau\to 0^+}\left(\phi^{\text{plus}}(x,\tau)\right)'_x=\begin{cases}1 \text{ if } \bar{x}>0\\ 0 \text{ if } \bar{x}<0\end{cases}.$$

Hence, it follows

$$\left\{\lim_{x\to\bar{x},\tau\to 0^+}\left(\phi^{\text{plus}}(x,\tau)\right)'_x\right\}\subset\partial(\bar{x})_+,\quad \forall\,\bar{x}\neq 0.$$

□

Note that the argument of Theorem 2.5 is also true when we define $\phi^{\text{plus}}(x,\tau)=(x+\alpha\tau)\psi(x,\tau)$ or $\phi^{\text{plus}}(x,\tau)=(x-\beta\tau)\psi(x,\tau)$. Let $\psi(x,\tau)$ be given in (11). We obtain the following smoothing functions of $(x)_+$.

$$\phi_4^{\text{plus}}(x,\tau)=\begin{cases}0 & \text{if } x<-\tau\\ -\frac{1}{4}\frac{x^4}{\tau^3}+\frac{3}{4}\frac{x^2}{\tau}+\frac{x}{2} & \text{if } |x|\leq\tau\\ x & \text{if } x>\tau.\end{cases}$$

$$\phi_5^{\text{plus}}(x,\tau)=\begin{cases}0 & \text{if } x<-\tau\\ -\frac{1}{4}\frac{x^4}{\tau^3}-\frac{1}{4}\frac{x^3}{\tau^2}+\frac{3}{4}\frac{x^2}{\tau}+\frac{5}{4}x+\frac{\tau}{2} & \text{if } |x|\leq\tau\\ x+\tau & \text{if } x>\tau.\end{cases}$$

$$\phi_6^{\text{plus}}(x,\tau)=\begin{cases}0 & \text{if } x<-\tau\\ -\frac{1}{4}\frac{x^4}{\tau^3}+\frac{1}{4}\frac{x^3}{\tau^2}+\frac{3}{4}\frac{x^2}{\tau}-\frac{1}{4}x-\frac{\tau}{2} & \text{if } |x|\leq\tau\\ x-\tau & \text{if } x>\tau.\end{cases}$$

The graphs representing these functions are illustrated in Figure 6.

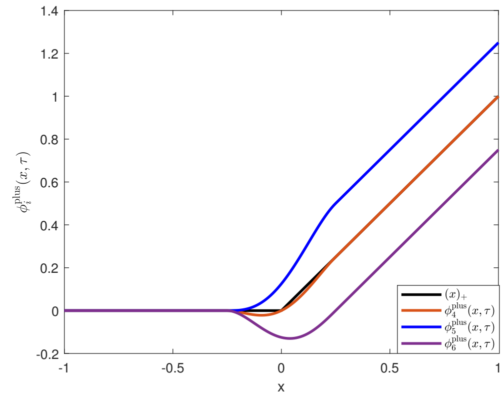


Fig. 6: Graphs of $(x)_+$ and $\phi_i^{\text{plus}}(x,\tau)$ with $\tau = 0.25$.

Theorem 2.6 *Suppose that $\psi(x,\tau)$ given by (23)-(24) is a C^1-approximation function of $\sigma(x)$. Let*

$$\phi^{\text{plus}}(x,\tau) = x\psi(x,\tau) - \int_{-\infty}^{x} t\psi'(t,\tau)dt.$$

Then, the following hold.

(i) For fixed $\tau > 0$, there exists $M > 0$ such that

$$\left|\phi^{\text{plus}}(x,\tau) - (x)_+\right| \le \tau \left(M + \int_{-\alpha}^{\beta} |tg'(t)|dt \right).$$

(ii) $\phi^{\text{plus}}(x,\tau)$ is a smoothing function of $(x)_+$.
(iii) For any x, $\left(\phi^{\text{plus}}(x,\tau)\right)'_x$ is bounded.
(iv) If $g(x,\tau)$ is increasing on $[-\alpha\tau, \beta\tau]$, then

$$\left\{ \lim_{x\to\bar{x},\tau\to 0^+} \left(\phi^{\text{plus}}(x,\tau)\right)'_x \right\} \subset \partial(\bar{x})_+, \quad \forall \bar{x}.$$

Proof (i) If $-\alpha\tau \le x \le \beta\tau$, we know

$$\int_{-\infty}^{x} t\psi'(t,\tau)dt = \int_{-\alpha\tau}^{x} t\psi'(t,\tau)dt = \tau^2 \int_{-\alpha}^{\frac{x}{\tau}} tg'(t)dt.$$

Using the Triangle Inequality for Integrals, we have

$$0 \le \left| \tau^2 \int_{-\alpha}^{\frac{x}{\tau}} tg'(t)dt \right| \le \tau^2 \int_{-\alpha}^{\frac{x}{\tau}} |tg'(t)|dt \le \tau^2 \int_{-\alpha}^{\beta} |tg'(t)|dt \le \tau \int_{-\alpha}^{\beta} |tg'(t)|dt.$$

which together with the proof of part (i) of Theorem 2.4, establishes the existence of $M > 0$ such that

$$\left|\phi^{\text{plus}}(x,\tau) - (x)_+\right| \le M\tau + \tau \int_{-\alpha}^{\beta} |tg'(t)|dt = \tau \left(M + \int_{-\alpha}^{\beta} |tg'(t)|dt \right).$$

(ii) It is easily derived from part (i).

(iii) Since $\left(\phi^{\text{plus}}(x,\tau)\right)'_x = \psi(x,\tau)$ and $\psi(x,\tau)$ is bounded, it is clear that $\left(\phi^{\text{plus}}(x,\tau)\right)'_x$ is also bounded.

(iv) The proof is the same as part (iv) of Theorem 2.4. □

In addition, we consider $\psi(x,\tau)$ given in (11), (12) with $\lambda = \frac{1}{10}$ and (22), which provide the below smoothing functions of $(x)_+$.

$$\phi_7^{\text{plus}}(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ -\frac{1}{16}\frac{x^4}{\tau^3} + \frac{3}{8}\frac{x^2}{\tau} + \frac{x}{2} + \frac{3\tau}{16} & \text{if } |x| \le \tau \\ x & \text{if } x > \tau. \end{cases}$$

$$\phi_8^{\text{plus}}(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \frac{1}{50}\frac{x^5}{\tau^4} - \frac{1}{16}\frac{x^4}{\tau^3} - \frac{1}{15}\frac{x^3}{\tau^2} + \frac{3}{8}\frac{x^2}{\tau} + \frac{3}{5}x + \frac{289}{1200}\tau & \text{if } |x| \le \tau \\ x + \frac{7}{85}\tau & \text{if } x > \tau. \end{cases}$$

$$\phi_9^{\text{plus}}(x,\tau) = \begin{cases} 0 & \text{if } x < -\tau \\ \frac{x}{2} - \tau\left(\frac{1}{\pi}\cos\left(\frac{\pi}{2\tau}x\right) - \frac{1}{2}\right) & \text{if } |x| \le \tau \\ x & \text{if } x > \tau. \end{cases}$$

The graphs representing these functions are illustrated in Figure 7.

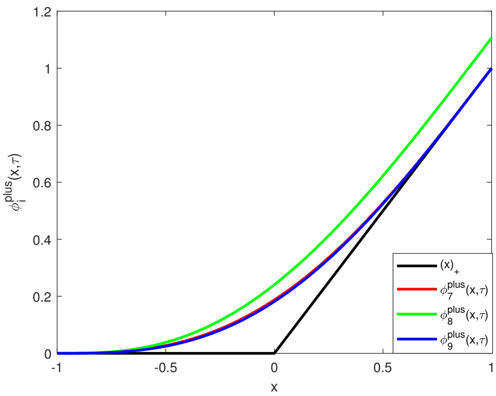


Fig. 7: Graphs of $(x)_+$ and $\phi_i^{\text{plus}}(x,\tau)$ with $\tau = 1$.

Theorem 2.7 *Suppose that $\psi(x,\tau)$ is given by (14) and $\rho(x)$ is continuous density function. Then*

(i) $\phi^{\text{plus}}(x,\tau) = x\psi(x,\tau)$ *is a smoothing function of* $(x)_+$.

(ii) For fixed $\tau > 0$, $|x|\min\left\{-\int_{-\infty}^{0}\rho(s)ds, \int_{-\infty}^{0}\rho(s)ds - 1\right\} \le \phi^{\text{plus}}(x,\tau) - (x)_+ \le 0.$

(iii) If $x\rho(x)$ *is bounded then* $\left(\phi^{\text{plus}}(x,\tau)\right)'_x$ *is bounded.*

(iv) For any x, $\left\{\lim_{x\to\bar{x},\tau\to 0^+}\left(\phi^{\text{plus}}(x,\tau)\right)'_x\right\} \subset \partial(\bar{x})_+, \quad \forall\, \bar{x} \ne 0.$

Proof (i) First, we need to show that

$$\lim_{x\to\bar{x},\tau\to 0^+}\phi^{\text{plus}}(x,\tau) = (\bar{x})_+.$$

Indeed, for $\bar{x} \neq 0$, we see that

$$\lim_{x \to \bar{x}, \tau \to 0^+} \phi^{\text{plus}}(x, \tau) = \lim_{x \to \bar{x}, \tau \to 0^+} x \int_{-\infty}^{x} \hat{t}(s, \tau) ds = \lim_{x \to \bar{x}, \tau \to 0^+} x \int_{-\infty}^{\frac{x}{\tau}} \rho(s) ds = (\bar{x})_+.$$

If $\bar{x} = 0$, by the definition of density function, we obtain

$$0 \leq \int_{-\infty}^{x} \hat{t}(s, \tau) ds = \int_{-\infty}^{\frac{x}{\tau}} \rho(s) ds \leq 1$$

which leads to

$$0 \leq |x \int_{-\infty}^{x} \hat{t}(s, \tau) ds| \leq |x|.$$

This further implies that

$$\lim_{x \to 0, \tau \to 0^+} \phi^{\text{plus}}(x, \tau) = 0.$$

On the other hand, we have

$$\left(\phi^{\text{plus}}(x, \tau)\right)'_x = \int_{-\infty}^{x} \hat{t}(s, \tau) ds + x \hat{t}(x, \tau).$$

Since $\int_{-\infty}^{x} \hat{t}(s, \tau) ds$ and $\hat{t}(x, \tau)$ are continuous, $\left(\phi^{\text{plus}}(x, \tau)\right)'_x$ is continuous which means that $\phi^{\text{plus}}(x, \tau)$ is C^1. Thus $\phi^{\text{plus}}(x, \tau)$ is a smoothing function of $(x)_+$.

(ii) For $x > 0$, there hold

$$x \left(\int_{-\infty}^{0} \rho(s) ds - 1 \right) \leq x \int_{-\infty}^{x} \hat{t}(s, \tau) ds - (x)_+ = x \int_{-\infty}^{\frac{x}{\tau}} \rho(s) ds - x \leq 0.$$

For $x \leq 0$, we have

$$x \int_{-\infty}^{0} \rho(s) ds \leq x \int_{-\infty}^{x} \hat{t}(s, \tau) ds - (x)_+ = x \int_{-\infty}^{\frac{x}{\tau}} \rho(s) ds \leq 0.$$

Hence, we obtain the desired result.

(iii) Since $x\rho(x)$ is bounded, it follows that $x\hat{t}(x, \tau)$ is also bounded. This together with the boundedness of $\int_{-\infty}^{x} \hat{t}(s, \tau) ds$ implies that $\left(\phi^{\text{plus}}(x, \tau)\right)'_x$ is bounded.

(iv) If $\bar{x} \neq 0$ then

$$\lim_{x \to \bar{x}, \tau \to 0^+} x \hat{t}(x, \tau) = \lim_{x \to \bar{x}, \tau \to 0^+} x . \frac{1}{\tau} \rho\left(\frac{x}{\tau}\right) = 0$$

which leads to

$$\lim_{x \to \bar{x}, \tau \to 0^+} \left(\phi^{\text{plus}}(x, \tau)\right)'_x = \begin{cases} 1 \text{ if } \bar{x} > 0 \\ 0 \text{ if } \bar{x} < 0 \end{cases}.$$

Then, the proof is complete. □

Moreover, consider $\psi(x,\tau)$ given by (16)-(21), and applying Theorem 2.7, we have the following corresponding smoothing functions of $(x)_+$:

$$\phi_{10}^{\text{plus}}(x,\tau) = \frac{x}{2}\left(\frac{x}{\sqrt{x^2+\tau^2}}+1\right), \tag{27}$$

$$\phi_{11}^{\text{plus}}(x,\tau) = \frac{xe^{\frac{x}{\tau}}}{e^{\frac{x}{\tau}}+1}, \tag{28}$$

$$\phi_{12}^{\text{plus}}(x,\tau) = \frac{x}{\pi}\left(\arctan\left(\frac{x}{\tau}\right)+\frac{\pi}{2}\right), \tag{29}$$

$$\phi_{13}^{\text{plus}}(x,\tau) = \frac{x}{2}\left(\text{erf}(\frac{x}{\tau})+1\right), \tag{30}$$

$$\phi_{14}^{\text{plus}}(x,\tau) = \frac{x}{2}\left(\tanh(\frac{x}{\tau})+1\right), \tag{31}$$

$$\phi_{15}^{\text{plus}}(x,\tau) = \frac{2x}{\pi}\arctan(e^{\frac{x}{\tau}}), \tag{32}$$

where the error function is defined by

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-u^2}du \quad \forall x \in \mathbb{R}.$$

The graphs representing these functions are illustrated in Figure 8.

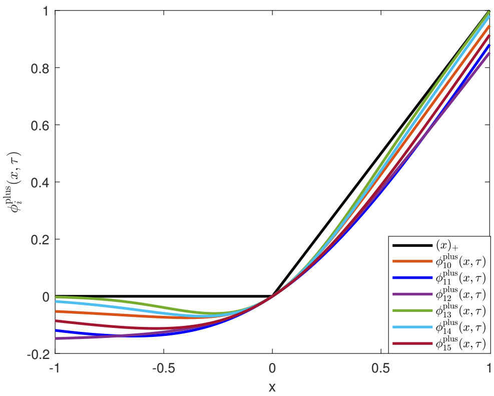


Fig. 8: Graphs of $(x)_+$ and $\phi_i^{\text{plus}}(x,\tau)$ with $\tau = 0.5$.

Theorem 2.8 *Suppose that $\psi(x,\tau)$ is given by (14) and $\rho(x)$ is continuous density function. Let*

$$\phi^{\text{plus}}(x,\tau) = x\psi(x,\tau) - \int_{-\infty}^{x} s\cdot\hat{t}(s,\tau)ds = \int_{-\infty}^{x}(x-s)\hat{t}(s,\tau)ds.$$

Then, the following hold.

(i) For fixed $\tau > 0$, $-\tau\max\left\{\int_{-\infty}^{+\infty} s\rho(s)ds, 0\right\} \le \phi^{\text{plus}}(x,\tau) - (x)_+ \le \tau\int_{-\infty}^{0}|s|\rho(s)ds.$

(ii) $\phi^{\text{plus}}(x,\tau)$ is a smoothing function of $(x)_+$.

(iii) For any x, $\left(\phi^{\text{plus}}(x,\tau)\right)'_x \in [0,1]$.

(iv) For any x, $\left\{ \lim_{x\to\bar{x},\tau\to 0^+} \left(\phi^{\text{plus}}(x,\tau)\right)'_x \right\} \subset \partial(\bar{x})_+, \quad \forall\, \bar{x}.$

Proof The proof follows directly from [5, Proposition 2.2]. □

To close this section, we point out that $\phi^{\text{plus}}(x,\tau)$ in Theorem 2.8 can be rewritten as

$$\phi^{\text{plus}}(x,\tau) = \int_{-\infty}^{+\infty} (x-s)_+ \hat{t}(s,\tau)ds,$$

which was previously investigated by Chen and Mangasarian [5].

2.3 Smoothing approximation of ℓ_0-norm

Definition 2.6 The l_0-norm function is defined as

$$\|z\|_0 \equiv \#\{i : z_i \neq 0\}.$$

which is equivalent to

$$\|z\|_0 = \sum_{i=1}^{n} \vartheta(z_i),$$

where $\vartheta : \mathbb{R} \to \mathbb{R}_+$ is defined by

$$\vartheta(t) = \begin{cases} 0 & \text{if } t = 0 \\ 1 & \text{otherwise.} \end{cases}$$

First, we observe that

$$\vartheta(t) = (\text{sgn}(t))^2,$$

where

$$\text{sgn}(t) = \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{if } t = 0 \\ -1 & \text{if } t < 0 \end{cases}$$

with

$$\text{sgn}(t) = \sigma(t) - \sigma(-t).$$

In other words, the ℓ_0-norm can be expressed via sgn function as follows:

$$\|z\|_0 \equiv \sum_{i=1}^{n} (\text{sgn}(z_i))^2 = \sum_{i=1}^{n} (\sigma(z_i) - \sigma(-z_i))^2.$$

Using C^1-approximation functions of $\sigma(x)$, we are able to generate smoothing functions for $\text{sgn}(x)$ as the following theorem.

Theorem 2.9 *Suppose that $\psi(x,\tau)$ given by (23) is a C^1-approximation function of $\sigma(x)$ then*

$$\bar{\psi}(x,\tau) = \psi(x,\tau) - \psi(-x,\tau)$$

is a smoothing function of $\text{sgn}(x)$.

Proof Since $\psi(x,\tau)$ and $\psi(-x,\tau)$ are C^1, so is $\bar{\psi}(x,\tau)$. On the other hand, we have that

$$\lim_{\tau\to 0^+} \psi(x,\tau) = \begin{cases} \sigma(x) & \text{if } x \neq 0 \\ g(0,\tau) & \text{if } x = 0, \end{cases}$$

and

$$\lim_{\tau\to 0^+} \psi(-x,\tau) = \begin{cases} \sigma(-x) & \text{if } x \neq 0 \\ g(0,\tau) & \text{if } x = 0, \end{cases}$$

which lead to $\lim_{\tau\to 0^+} \bar{\psi}(x,\tau) = \text{sgn}(x)$ for all $x \in \mathbb{R}$. Thus $\bar{\psi}(x,\tau)$ is a smoothing function of $\text{sgn}(x)$. □

From the above theorem, if $\bar{\psi}(x,\tau)$ is a smoothing function of $\text{sgn}(x)$, then

$$\phi^0(z,\tau) = \sum_{i=1}^{n} (\bar{\psi}(z_i,\tau))^2$$

is a smoothing function of $\|z\|_0$. For example, let $\psi(x,\tau)$ be as in (7), we have

$$\bar{\psi}(x,\tau) = \begin{cases} -1 & \text{if } x < -\tau \\ \frac{(x+\tau)^2}{\tau^2} - 1 & \text{if } -\tau \leq x \leq 0 \\ 1 - \frac{(x-\tau)^2}{\tau^2} & \text{if } 0 \leq x \leq \tau \\ 1 & \text{if } x > \tau \end{cases} = \text{sgn}(x) \begin{cases} 1 & \text{if } |x| > \tau \\ 1 - \frac{(x-\tau)^2}{\tau^2} & \text{if } |x| \leq \tau. \end{cases}$$

Then

$$\phi^0(x,\tau) = \begin{cases} 1 & \text{if } |x| > \tau \\ \left(1 - \frac{(x-\tau)^2}{\tau^2}\right)^2 & \text{if } |x| \leq \tau \end{cases}$$

is a smoothing function of the ℓ_0-norm with $n = 1$. The graphs of the ℓ_0-norm and $F(x,\tau)$ are depicted in Figure 9.

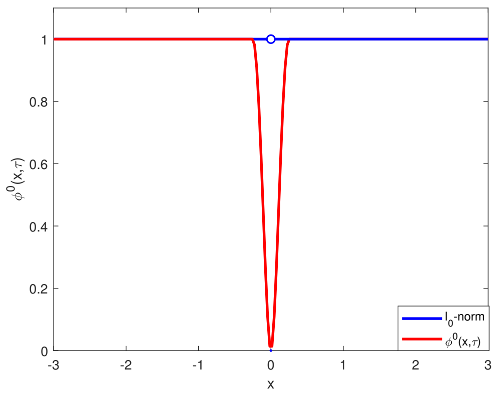


Fig. 9: Graphs of the ℓ_0-norm with $n = 1$ and $\phi^0(x,\tau)$ with $\tau = 0.5$.

Smoothing functions for $\text{sgn}(x)$ can be constructed using density functions, as formally stated in the following theorem.

Theorem 2.10 *Let $\rho(x)$ be a continuous density function satisfying the symmetry condition $\rho(x) = \rho(-x)$. Then a smoothing function of* $\mathrm{sgn}(x)$ *is defined by*

$$\bar{\psi}(x,\tau) = \int_{-\infty}^{x} \hat{t}(s,\tau)ds - \int_{-\infty}^{-x} \hat{t}(-s,\tau)ds = 2\psi(x,\tau) - 1.$$

where $\hat{t}(s,\tau)$ is defined as in (13) *and $\psi(x,\tau)$ is given by* (14).

Proof By symmetric of $\rho(x)$, it follows that

$$\bar{\psi}(x,\tau) = \int_{-\infty}^{x} \hat{t}(s,\tau)ds - \int_{-\infty}^{-x} \hat{t}(-s,\tau)ds = 2\int_{-\infty}^{x} \hat{t}(s,\tau)ds - 1 = 2\psi(x,\tau) - 1.$$

According to (15), we obtain

$$\lim_{\tau\to 0^+} \bar{\psi}(x,\tau) = \lim_{\tau\to 0^+} 2\psi(x,\tau) - 1 = \begin{cases} 0 & \text{if } x = 0 \\ \mathrm{sgn}(x) & \text{if } x \neq 0. \end{cases}$$

On the other hand, we have that

$$\bar{\psi}'(x,\tau) = 2\hat{t}(x,\tau)$$

is a continuous function. Thus, $\bar{\psi}(x,\tau)$ is a smoothing function of $\mathrm{sgn}(x)$. □

To sum up, the corresponding smoothing function of $\|z\|_0$ is

$$\phi^0(z,\tau) = \sum_{i=1}^{n} \left(2\int_{-\infty}^{z_i} \hat{t}(s,\tau)ds - 1\right)^2.$$

Using $\psi_i(x,\tau)$, $i = 1,\ldots,6$ given by (16)-(21) and notice that the corresponding density function of $\psi_i(x,\tau)$, $i = 1,\ldots,6$ is continuous and satisfies $\rho(x) = \rho(-x)$. Then, we obtain the corresponding smoothing functions of the ℓ_0-norm with $n = 1$ as following:

$$\begin{aligned}
\phi_1^0(x,\tau) &= \left(\frac{x}{\sqrt{x^2+\tau^2}}\right)^2, \\
\phi_2^0(x,\tau) &= \left(\frac{e^{\frac{x}{\tau}} - 1}{e^{\frac{x}{\tau}} + 1}\right)^2, \\
\phi_3^0(x,\tau) &= \left(\frac{2}{\pi}\arctan\left(\frac{x}{\tau}\right)\right)^2, \\
\phi_4^0(x,\tau) &= \left(\mathrm{erf}(\frac{x}{\tau})\right)^2, \\
\phi_5^0(x,\tau) &= \left(\tanh(\frac{x}{\tau})\right)^2, \\
\phi_6^0(x,\tau) &= \left(\frac{4}{\pi}\arctan(e^{\frac{x}{\tau}}) - 1\right)^2.
\end{aligned}$$

The graphs of these smoothing functions are depicted in Figure 10.

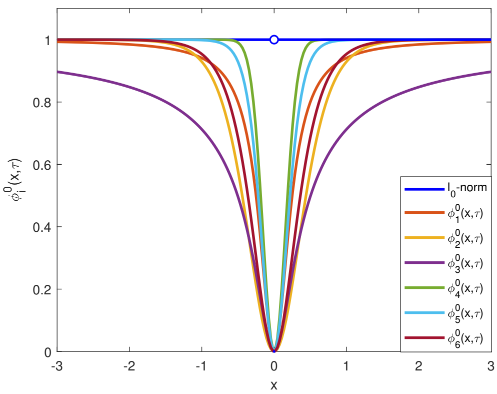


Fig. 10: Graphs of the ℓ_0-norm with $n = 1$ and $\phi_i^0(x, \tau)$ with $\tau = 0.25$.

2.4 Smoothing approximations for other regularizers

2.4.1 Approximations for ℓ_p-norm ($0 < p \leq 1$), MCP, SCAD

It is well known that the absolute value function can be represented in terms of

$$|x| = (x)_+ + (-x)_+ = 2(x)_+ - x.$$

Then

$$\phi^{\text{abs}}(x, \tau) = \phi^{\text{plus}}(x, \tau) + \phi^{\text{plus}}(-x, \tau) = 2\phi^{\text{plus}}(x, \tau) - x$$

is a smoothing function of $|x|$, where $\phi^{\text{plus}}(x, \tau)$ is a smoothing function of $(x)_+$. For example, let $\phi^{\text{plus}}(x, \tau)$ be given by (26), then we have the corresponding smoothing function $|x|$

$$\phi_1^{\text{abs}}(x, \tau) = \begin{cases} -x & \text{if } x < -\tau \\ \frac{x^2}{2\tau} + \frac{\tau}{2} & \text{if } |x| \leq \tau \\ x & \text{if } x > \tau. \end{cases}$$

Using the smoothing functions $\phi^{\text{plus}}(x, \tau)$ given by (27)-(32), we have the following corresponding smoothing functions of $|x|$:

$$\begin{aligned}
\phi_2^{\text{abs}}(x, \tau) &= \frac{x^2}{\sqrt{x^2 + \tau^2}}, \\
\phi_3^{\text{abs}}(x, \tau) &= x\frac{e^{\frac{x}{\tau}} - 1}{e^{\frac{x}{\tau}} + 1}, \\
\phi_4^{\text{abs}}(x, \tau) &= \frac{2x}{\pi} \arctan\left(\frac{x}{\tau}\right), \\
\phi_5^{\text{abs}}(x, \tau) &= x\text{erf}\left(\frac{x}{\tau}\right), \\
\phi_6^{\text{abs}}(x, \tau) &= x\tanh(\frac{x}{\tau}), \\
\phi_7^{\text{abs}}(x, \tau) &= \frac{4x}{\pi} \arctan(e^{\frac{x}{\tau}}) - x.
\end{aligned}$$

Definition 2.7 The ℓ_p-norm function is defined by

$$\|z\|_p = \left(\sum_{i=1}^{n} |z_i|^p \right)^{\frac{1}{p}}$$

for any $z \in \mathbb{R}^n$, and $p \in (0, 1]$.

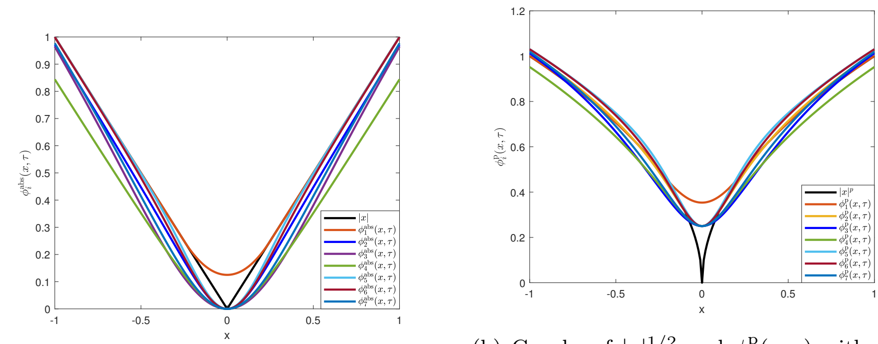
We can utilize smoothing functions of $|x|$ to derive smoothing functions for ℓ_1-norm and positive smoothing functions of $|x|$ to obtain smoothing functions of the ℓ_p-norm ($p \in (0,1)$). For example, since $\phi_1^{\text{abs}}(x,\tau) > 0$, the corresponding smoothing function of $|x|^p$ ($p \in (0,1)$) is:

$$\phi_1^{\text{p}}(x,\tau) = \begin{cases} \left(\frac{x^2}{2\tau} + \frac{\tau}{2} \right)^p & \text{if } |x| \leq \tau \\ |x|^p & \text{if } |x| > \tau. \end{cases}$$

Note that $\phi_i^{\text{abs}}(x,\tau) \geq 0$, $i = 2, \ldots, 7$, hence smoothing functions for $|x|^p$ ($p \in (0,1)$) are generated by modifying $\phi_i^{\text{abs}}(x,\tau)$ as follows.

$$\phi_i^{\text{p}}(x,\tau) = \left(\phi_i^{\text{abs}}(x,\tau) + \tau^2 \right)^p, \quad i = 2, \ldots, 7.$$

The graphs of smoothing functions of $|x|$ and $|x|^p$ are depicted in Figure 11.



(a) Graphs of $|x|$ and $\phi_i^{\text{abs}}(x,\tau)$ with $\tau = 0.25$.

(b) Graphs of $|x|^{1/2}$ and $\phi_i^{\text{p}}(x,\tau)$ with $\tau = 0.25$.

Fig. 11: Graph of smoothing functions of $|x|$ and $|x|^{1/2}$.

Note that smoothing of the absolute value can also be applied to matrix norms. For instance, consider the induced 1-norm $\|X\|_1 = \max\limits_{1 \leq j \leq n} \sum\limits_{i}^{m} |x_{ij}|$. A smooth approximation of $\|X\|_1$ is then given by

$$\phi(X,\tau) = \frac{1}{\tau} \log \left(\sum_{j=1}^{n} e^{\tau \sum\limits_{i=1}^{m} \phi^{\text{abs}}(x_{ij},\tau)} \right),$$

where $\phi^{\text{abs}}(x_{ij},\tau)$ is smoothing function of $|x_{ij}|$.

Definition 2.8 [33] The minimax concave penalty (MCP) function is defined by

$$f^{\text{mcp}}(x) = \int_0^{|x|} (1 - s/(\alpha\lambda))_+ ds, \ \alpha > 0$$
$$= \begin{cases} \frac{1}{2}\alpha\lambda & \text{if } x > \alpha\lambda \\ |x| - \frac{x^2}{2\alpha\lambda} & \text{if } \alpha\lambda \le x \le \alpha\lambda \\ \frac{1}{2}\alpha\lambda & \text{if } x < \alpha\lambda. \end{cases}$$

MCP function $f^{\text{mcp}}(x)$ can be rewritten as

$$f^{\text{mcp}}(x) = \frac{\alpha\lambda}{2} + |x| + \frac{|x - \alpha\lambda|\,(x - \alpha\lambda)}{4\alpha\lambda} - \frac{|x + \alpha\lambda|\,(x + \alpha\lambda)}{4\alpha\lambda}.$$

Then a smoothing function of MCP is

$$\phi^{\text{mcp}}(x,\tau) = \frac{\alpha\lambda}{2} + \phi^{\text{abs}}(x,\tau) + \frac{\phi^{\text{abs}}(x - \alpha\lambda,\tau)(x - \alpha\lambda)}{4\alpha\lambda} - \frac{\phi^{\text{abs}}(x + \alpha\lambda,\tau)(x + \alpha\lambda)}{4\alpha\lambda},$$

where $\phi^{\text{abs}}(x,\tau)$ is a smoothing function of $|x|$. Using the preceding smoothing functions of $|x|$ given by (33)-(33), the graphs of smoothing functions of the MCP are depicted in Figure 12.

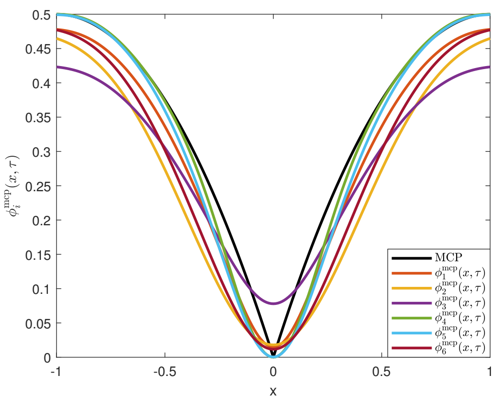


Fig. 12: Graphs of the MCP with $\alpha = \lambda = 1$ and $\phi_i^{\text{mcp}}(x,\tau)$ with $\tau = 0.25$.

Definition 2.9 [8] The smoothly clipped absolute deviation (SCAD) function is defined by

$$f^{\text{scad}}(x) = \int_0^{|x|} \min\left(1, \frac{(\alpha - s/\lambda)_+}{\alpha - 1}\right) ds, \ \alpha > 2$$
$$= \begin{cases} \frac{\lambda(\alpha+1)}{2} & \text{if } |x| \ge \alpha\lambda \\ \frac{2\alpha\lambda|x| - x^2 - \lambda^2}{2\lambda(\alpha-1)} & \text{if } \lambda < |x| < \alpha\lambda \\ |x| & \text{if } |x| \le \lambda. \end{cases}$$

SCAD function $f^{\text{scad}}(x)$ can be rewritten as

$$f^{\text{scad}}(x) = \frac{\lambda(\alpha+1)}{2} + |x| + \frac{(x - \alpha\lambda)\,|x - \alpha\lambda| - (x + \alpha\lambda)\,|x + \alpha\lambda|}{4\lambda(\alpha - 1)}$$
$$+ \frac{(x + \lambda)\,|x + \lambda| - (x - \lambda)\,|x - \lambda|}{4\lambda(\alpha - 1)}.$$

Then a smoothing function of SCAD is

$$\phi^{\text{scad}}(x,\tau) = \frac{\lambda(\alpha+1)}{2} + \phi^{\text{abs}}(x,\tau) + \frac{(x-\alpha\lambda)\phi^{\text{abs}}(x-\alpha\lambda,\tau) - (x+\alpha\lambda)\phi^{\text{abs}}(x+\alpha\lambda,\tau)}{4\lambda(\alpha-1)}$$
$$+ \frac{(x+\lambda)\phi^{\text{abs}}(x+\lambda,\tau) - (x-\lambda)\phi^{\text{abs}}(x-\lambda,\tau)}{4\lambda(\alpha-1)},$$

where $\phi^{\text{abs}}(x,\tau)$ is a smoothing function of $|x|$. Using the preceding smoothing functions of $|x|$ given by (33)-(33), the graphs of smoothing functions of the SCAD are depicted in Figure 13.

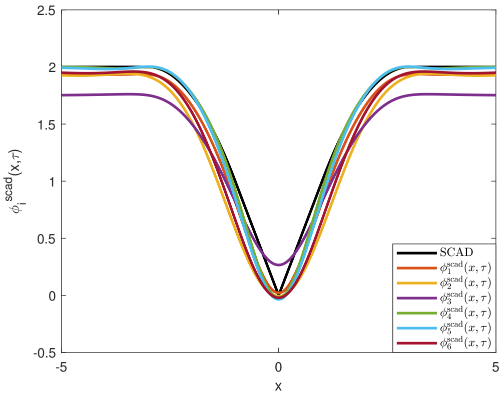


Fig. 13: Graphs of the SCAD with $\alpha = 3, \lambda = 1$ and $\phi_i^{\text{scad}}(x,\tau)$ with $\tau = 0.8$.

2.4.2 Smoothing approximation for pinball function

Definition 2.10 The pinball function is defined by

$$\rho_\eta(x) = \begin{cases} \eta x & \text{if } x \geq 0 \\ (\eta-1)x & \text{if } x < 0, \end{cases} \tag{33}$$

where $\eta \in [0,1]$.

Note that

$$\rho_\eta(x) = \begin{cases} (x)_+ & \text{if } \eta = 1 \\ (-x)_+ & \text{if } \eta = 0, \end{cases}$$

and the pinball function (33) with $\eta \in (0,1)$ is called the quantile loss function [10].

The pinball function can be rewritten as

$$\rho_\eta(x) = \eta(x)_+ - (\eta-1)(-x)_+.$$

Then a smoothing function of $\rho_\eta(x)$ is defined by

$$\phi^{\text{pin}}(x,\tau) = \eta\phi^{\text{plus}}(x,\tau) - (\eta-1)\phi^{\text{plus}}(-x,\tau),$$

where $\phi^{\text{plus}}(x,\tau), \phi^{\text{plus}}(-x,\tau)$ are smoothing functions of $(x)_+, (-x)_+$, respectively.

Definition 2.11 [9] The pinball function with ϵ zone is defined by

$$\rho_\eta^\epsilon(x) = \max\{x-\epsilon, 0, -\eta x - \epsilon\},$$

where $0 \le \eta \le 1$, $\epsilon \ge 0$.

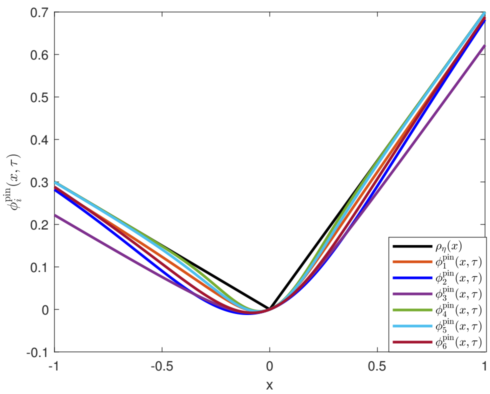
The pinball function with ϵ zone can be rewritten as

$$\rho_\eta^\epsilon(x) = (x-\epsilon)_+ + (-\eta x - \epsilon)_+.$$

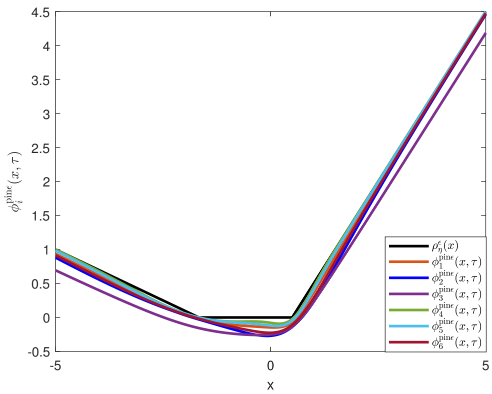
Then a smoothing function of $\rho_\eta^\epsilon(x)$ is defined by

$$\phi^{\text{pin}\epsilon}(x,\tau) = \phi^{\text{plus}}(x-\epsilon,\tau) + \phi^{\text{plus}}(-\eta x-\epsilon,\tau),$$

where $\phi^{\text{plus}}(x-\epsilon,\tau)$, $\phi^{\text{plus}}(-\eta x-\epsilon,\tau)$ are smoothing functions of $(x-\epsilon)_+$, $(-\eta x-\epsilon)_+$, respectively. Let $\phi^{\text{plus}}(x,\tau)$ be defined as in (27)-(32), the graphs of smoothing functions of the pinball and the pinball with ϵ zone functions are depicted in Figure 14.



(a) Graphs of the pinball function and $\phi_i^{\text{pin}}(x,\tau)$ with $\eta = 0.7$, $\tau = 0.25$.



(b) Graphs of the pinball function with ϵ zone and $\phi_i^{\text{pin}\epsilon}(x,\tau)$ with $\eta = 0.3$, $\epsilon = 0.5$, $\tau = 0.5$.

Fig. 14: Smoothing functions of the pinball and the pinball with ϵ zone functions.

3 Applications to Sparse optimizations and regularization

Sparse optimization encompasses techniques that encourage solutions with many zero, proving especially valuable in fields such as signal processing, machine learning, and statistical modeling. Regularization methods such as the ℓ_p-norm, MCP, SCAD, and nuclear norm, introduce penalty terms that promote sparsity, thereby enhancing the tractability of optimization problems. In this section, we present several key applications of model (1) in sparse optimization, which represent fundamental problems in signal processing and machine learning. These examples highlight the pivotal role of regularization in efficiently recovering structured solutions from high-dimensional data.

(i) Sparse regression

In the classical linear regression problem, given a design matrix $X \in \mathbb{R}^{m\times n}$ and a response vector $y \in \mathbb{R}^m$, the goal is to find a coefficient vector $\beta \in \mathbb{R}^n$ that minimizes the squared error, which yields

$$\min_{\beta} \|y - X\beta\|^2.$$

However, in high-dimensional settings with large datasets, many features are often irrelevant, implying that several components of β should be zero—that is, β is sparse. To promote such sparsity, a natural formulation is the following optimization model:

$$\begin{array}{rl} \min & \frac{1}{2}\|y - X\beta\|^2 \\ \text{subject to} & \|\beta\|_0 \leq s, \end{array}$$

where $\|\beta\|_0$ denotes the number of nonzero elements of β and s is the desired sparsity level (i.e., the maximum number of predictors). Alternatively, a widely used approach is to consider regularized formulations of the form

$$\min_{\beta\in\mathbb{R}^n} \frac{1}{2}\|y - X\beta\|^2 + \lambda R(\beta), \tag{34}$$

where R is a regularizer that induces sparsity, such as the ℓ_p (with $0 \leq p \leq 1$), MCP and SCAD regularizers [4, 13].

(ii) Compressive sensing

Compressive sensing seeks to efficiently acquire and reconstruct sparse signals by solving underdetermined linear systems of the form $Ax = b$, where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ with $m \ll n$. This problem has a wide range of applications, including magnetic resonance imaging (MRI), radar, sampling theory, and machine learning. Mathematically, the objective is to solve the following sparsity-constrained problem:

$$\min_{x\in\mathbb{R}^n} \|x\|_0 \quad \text{subject to} \quad Ax = b.$$

However, owing to the nonconvexity and NP-hardness of this formulation [16], an alternative approximate and relaxed models are commonly employed, such as:

$$\min_{x\in\mathbb{R}^n} R(x) \quad \text{subject to} \quad Ax = b,$$

where $R(x)$ is a sparsity-inducing regularizer that approximates ℓ_0. An even more flexible formulation is given by the unconstrained problem

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2 + \lambda R(x),$$

which is mathematically similar to sparse regression (34).

(iii) Support vector machine

The Support Vector Machine (SVM) is a foundational model in machine learning, particularly well-suited for classification tasks. Given a training dataset $\{(x_i, y_i)\}_{i=1}^m$, where $x_i \in \mathbb{R}^n$ are feature vectors and $y_i \in \{-1, 1\}$ are class labels, the objective of SVM is to determine a hyperplane that best separates the two classes while maximizing the margin between them. This leads to the following optimization formulation:

$$\begin{array}{rl} \min_{\omega,b} & \frac{1}{2}\|\omega\|^2 \\ \text{subject to} & y_i(\omega^T x_i + b) \geq 1, \quad \forall i = 1, 2, ..., m. \end{array}$$

Here, ω denotes the normal vector to the hyperplane, and b is the bias term. In [9], a variant of the SVM model was proposed using the pinball loss function, leading to the following formulation:

$$\min_{\omega,b} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m} L_\eta(1 - y_i(\omega^T x_i + b)),$$

where L_η is pinball function given by

$$L_\eta(u) = \begin{cases} u & \text{if } u \geq 0 \\ -\eta u & \text{if } u < 0, \quad \eta \in (0,1). \end{cases}$$

The examples above illustrate only a subset of the many applications that arise across various applied fields. Additional notable examples include image inpainting [12], sparse signal separation [22], and the estimation of large sparse (inverse) covariance matrices [25]. For a comprehensive survey of applications and models involving regularization techniques, we refer the reader to [24]. A corresponding smooth approximation of the nonsmooth optimization model (1) can be formulated as follows:

$$\min_{x\in\Omega} f(x) + \lambda R_\tau(x), \tag{35}$$

where $R_\tau(x)$ is a smoothing function of $R(x)$.

The boundedness of $\nabla R_\tau(x)$ as provided in Section 2, together the assumption that f is coercive and continuously differentiable, ensures that the overall gradient

$$\nabla F_\tau(x) = \nabla f(x) + \lambda \nabla R_\tau(x)$$

cannot become arbitrarily large. This property is crucial for the stability of gradient-based algorithms (e.g., gradient descent, BFGS, L-BFGS, and so on), as it guarantees that step sizes can be chosen safely and prevents numerical instability when applying continuation methods with decreasing τ. Moreover, the gradient consistency property of $R_\tau(x)$ ensures that any accumulation point generated by the smoothing gradient method is a Clarke stationary point of the original nonsmooth problem (1) (see in [6, Theorem 3]). These properties of the smoothing function can be powerful for establishing the theoretical results and designing algorithms for solving nonsmooth optimization problems.

Futhermore, as established in the literature, the effectiveness of smoothing approaches for nonsmooth problems hinges on both the choice of smoothing functions and the solution methods employed. This underscores the importance of developing a systematic framework for smoothing regularizers. Such a framework facilitates the identification of suitable smoothing functions and supports the design of robust algorithms for solving the smooth approximation problem (35).

4 Conclusion

This paper presents a systematic framework for constructing smoothing functions for a broad class of widely used regularizers, including the plus function, the pinball function, the absolute value function, the ℓ_p-norm ($p \in [0,1]$), the MCP, and the SCAD. Our approach demonstrates that numerous well-known smoothing functions for the plus function arise naturally emerge within this framework. The central idea is built upon continuous approximations of the step function, which provide a foundational structure for generating smoothing functions for various regularizers. These approximations are derived using polynomial functions—motivated by Bézier curves and the Dirac delta function.

As part of future research, we plan to further explore these smoothing functions to identify those most effective for solving the sparse optimization problems discussed in Section 3.

Acknowledgments

The author is sincerely grateful to the anonymous reviewers for their valuable comments and suggestions, which have substantially improved the quality of this paper. The 2nd author's work is supported by National Science and Technology Council (NSTC 114-2115-M-003 -003 -MY2, NSTC 113-2124-M-003-002); Higher Education Sprout Project - Center for Optimal Intelligent Data Analytics and Prediction. We sincerely thank Dr. Jan Harold Alcantara for his valuable comments on the paper.

Declarations

- **Conflict of interest:** The authors declared that there is no conflict of interest.
- **Data Availability Statements:** Not applicable.

References

1. Beck, A., and Teboulle, M.: Smoothing and first order methods: A unified framework. SIAM J. Optim. 22, (2012)
2. Ben-Tal, A., and Teboulle, M.: A smoothing technique for nondifferentiable optimization problems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–11, (1989)
3. Bertsekas, D. P.: Nondifferentiable optimization via approximation. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 1–25, (1975)
4. Breheny, P., and Huang, J.: Coordinate descent algorithm for nonconvex penalized regression, with application to biological feature selection. Ann. Appl. Stat. 5, 232–253, (2011)
5. Chen, C., and Mangasarian, O.: A class of smoothing functions for nonlinear and mixed complementarity problems. Comput. Optim. Appl. 5, 97–138, (1996)
6. Chen, X.: Smoothing methods for nonsmooth, nonconvex minimization. Math. Programming 134, (2012)
7. Chen, X., Niu, L., and Yuan, Y.X.: Optimality conditions and a smoothing trust region newton method for nonlipschitz optimization. SIAM J. Optim. 23, (2013)
8. Fan, J., and Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348–1360, (2001)
9. Huang, X., Shi, L., and Suykens, J.: Support vector machine classifier with pinball loss. IEEE Trans. Pattern Anal. Mach. Intell. 36, 984–997, (2014)
10. Koenker, R.: Quantile Regression. Econometric Society Monographs. Cambridge University Press, (2005)
11. Kreimer, J., and Rubinstein, R.: Nondifferentiable optimization via smooth approximation: General analytical approach. Ann. Oper. Res. 39, 97–119, (1992)
12. Mallat, S., and Yu, G.: Super-resolution with sparse mixing estimators. IEEE Trans. Image Process. 19, 2889 – 2900, (2010)
13. Marjanovic, G., and Solo, V.: Lq sparsity penalized linear regression with cyclic descent. IEEE Trans. Image Process. 62, (2014).
14. Mohimani, H., Babaie-Zadeh, M., and Jutten, C.: Fast sparse representation based on smoothed ℓ_0 norm. Lecture Notes in Computer Science, vol 4666. Springer, Berlin, Heidelberg, (2007)
15. Mohimani, H., Babaie-Zadeh, M., and Jutten, C.: A fast approach for overcomplete sparse decomposition based on smoothed ℓ_0 norm. IEEE Trans. Image Process. 57, 289 – 301, (2009)
16. Natarajan, B. K.: Sparse approximate solutions to linear systems. SIAM J. Comput. 24, 227–234, (1995)
17. Nesterov, Y.: Smooth minimization of non-smooth functions. Math. Programming Ser. A 103, 127–153, (2005)
18. Pinar, M., and Zenios, S.: On smoothing exact penalty functions for convex constrained optimization. SIAM J. Optim. 4, 486–511, (1994)
19. Qi, L., and Chen, X.: A globally convergent successive approximation method for severely nonsmooth equations. SIAM J. Control Optim. 33, 402–418, (1996)
20. Qi, L., and Sun, D.: Smoothing functions and smoothing newton method for complementarity and variational inequality problems. J. Optim. Theory Appl. 113, 121–147, (2002)
21. Rubinstein, R.: Smoothed functionals in stochastic optimization. Math. Oper. Res. 8, 26–33, (1983)

22. Studer, C., and Baraniuk, R.: Stable restoration and separation of approximately sparse signals. Appl. Comput. Harmon. Anal. 37 (2011)
23. Tishler, A., and Zang, I.: A switching regression method using inequality conditions. J. Econom. 11, 259–274, (1979)
24. Wen, F., Chu, L., Liu, P., and Qiu, R.: A survey on nonconvex regularization based sparse and low-rank recovery in signal processing, statistics, and machine learning. IEEE Access 06, 69883–69906, (2018)
25. Wen, F., Yang, Y., Liu, P., and Qiu, R.: Positive definite estimation of large covariance matrix using generalized nonconvex penalties. IEEE Access 4 (2016).
26. Wu, H., Zhang, P., and Lin, G.-H.: Smoothing approximations for some piecewise smooth functions. J. Oper. Res. Soc. China 3 (2015)
27. Yang, X.: Smoothing approximations to nonsmooth optimization problems. J. Austral. Math. Soc. Ser. B 36, 274–285, (1995)
28. Ye, F., Liu, H., Zhou, S., and Liu, S.: A smoothing trust-region Newton-CG method for minimax problem. Appl. Math. Comput. 199, 581–589, (2008)
29. Yilmaz, N., and Sahiner, A.: New smoothing approximations to piecewise smooth functions and applications. Numer. Funct. Anal. Optim. 40, 1–22, (2019)
30. Yilmaz, N., and Sahiner, A.: On a new smoothing technique for non-smooth, non-convex optimization. Numer. Algebra Control Optim. 10, 317–330, (2020)
31. Yu, L., Yang, Z., and Tang, L.: Quantile estimators with orthogonal pinball loss function. Int. J. Forecast.37, 1–17, (2018)
32. Zang, I.: A smoothing-out technique for min-max optimization. Math. Programming 19, 61–77, (1980)
33. Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38(2), 894–942 (2010)
34. Zheng, S.: Gradient descent algorithms for quantile regression with smooth approximation. Int. J. Mach. Learn. Cybern. 2, 191–207, (2011)