Mathematical Programming manuscript No.
(will be inserted by the editor)

Smoothing Methods for Nonsmooth, Nonconvex Minimization

Xiaojun Chen

12 September, 2011, Revised 12 March, 2012, 6 April, 2012

Abstract We consider a class of smoothing methods for minimization problems where the feasible set is convex but the objective function is not convex, not differentiable and perhaps not even locally Lipschitz at the solutions. Such optimization problems arise from wide applications including image restoration, signal reconstruction, variable selection, optimal control, stochastic equilibrium and spherical approximations. In this paper, we focus on smoothing methods for solving such optimization problems, which use the structure of the minimization problems and composition of smoothing functions for the plus function $(x)_+$. Many existing optimization algorithms and codes can be used in the inner iteration of the smoothing methods. We present properties of the smoothing functions and the gradient consistency of subdifferential associated with a smoothing function. Moreover, we describe how to update the smoothing parameter in the outer iteration of the smoothing methods to guarantee convergence of the smoothing methods to a stationary point of the original minimization problem.

Keywords nonsmooth · nonconvex minimization · smoothing methods · regularized least squares · eigenvalue optimization · stochastic variational inequalities

1 Introduction

This paper considers the following nonsmooth, nonconvex optimization problem

$$\min_{x\in X} \quad f(x) \tag{1}$$

Department of Applied Mathematics, The Hong Kong Polytechnic University, Hong Kong, China. E-mail: maxjchen@polyu.edu.hk. This work was supported in part by Hong Kong Research Grant Council grant PolyU5003/10p.

where the feasible set $X \subseteq R^n$ is closed and convex, and the objective function $f : R^n \to R$ is continuous and almost everywhere differentiable in X. Of particular interest is the case where f is not smooth, not convex and perhaps not even locally Lipschitz. Here nonsmoothness refers to nondifferentiability.

Recently, problem (1) has attracted significant attention in engineering and economics. An increasing number of practical problems require minimizing a nonsmooth, nonconvex function on a convex set, including image restoration, signal reconstruction, variable selection, optimal control, stochastic equilibrium problems and spherical approximations [1,8,10,11,13,21,29,30,33,42,44, 61,76,85,101,105]. In many cases, the objective function f is not differentiable at the minimizers, but the constraints are simple, such as box constraints, $X = \{ x \in R^n \,|\ell \le x \le u\}$ for two given vectors $\ell \in \{R \cup \{-\infty\}\}^n$ and $u \in \{R \cup \{\infty\}\}^n$. For example, in some minimization models of image restoration, signal reconstruction and variable selection [8,33,42,85], the objective function is the sum of a quadratic data-fidelity term and a nonconvex, non-Lipschitz regularization term. Sometimes nonnegative constraints are used to reflect the fact that the pixels are nonnegative, and box constraints are used to represent thresholds for some natural phenomena or finance decisions. Moreover, a number of constrained optimization problems can be reformulated as problem (1) by using exact penalty methods. However, many well-known optimization algorithms lack effectiveness and efficiency in dealing with nonsmooth, nonconvex objective functions. Furthermore, for non-Lipschitz continuous functions, the Clarke generalized gradients [34] can not be used directly in the analysis.

Smooth approximations for optimization problems have been studied for decades, including complementarity problems, variational inequalities, second-order cone complementarity problems, semidefinite programming, semi-infinite programming, optimal control, eigenvalue optimization, penalty methods and mathematical programs with equilibrium constraints. Smoothing methods are not only efficient for problems with nonsmooth objective functions, but also for problems with nonsmooth constraints. See [1,2,4–7,14,16–18,22,24–31,33, 39,40,46–50,54,56–60,63,64,66,70,72,73,81,82,88–92,94,97,103,104].

In this paper, we describe a class of smooth approximations which are constructed based on the special structure of nonsmooth, nonconvex optimization problems and the use of smoothing functions for the plus function $(t)_+ := \max(0, t)$. Using the structure of problems and the composition of smoothing functions, we can develop efficient smoothing algorithms for solving many important optimization problems including regularized minimization problems, eigenvalue optimization and stochastic complementarity problems.

In particular, we consider a class of smoothing functions with the following definition.

Definition 1 Let $f : R^n \to R$ be a continuous function. We call $\tilde{f} : R^n \times R_+ \to R$ a smoothing function of f, if $\tilde{f}(\cdot, \mu)$ is continuously differentiable in R^n for any fixed $\mu > 0$, and for any $x \in R^n$,

$$\lim_{z \to x, \mu \downarrow 0} \tilde{f}(z, \mu) = f(x).$$

Based on this definition, we can construct a smoothing method using $\tilde{f}$ and $\nabla_x \tilde{f}$ as follows.

(i) Initial step: Define a parametric smooth function $\tilde{f} : R^n \times R_+ \to R$ to approximate f.

(ii) Inner iteration: Use an algorithm to find an approximate solution (no need to be an exact solution) of the smooth optimization problem

$$\min_{x \in X} \ \tilde{f}(x, \mu_k) \tag{2}$$

for a fixed $\mu_k > 0$.

(iii) Outer iteration: Update μ_k to guarantee the convergence of the smoothing method to a local minimizer or a stationary point of the nonsmooth problem (1).

The advantage of smoothing methods is that we solve optimization problems with continuously differentiable functions for which there are rich theory and powerful solution methods [86], and we can guarantee to find a local minimizer or stationary point of the original nonsmooth problem by updating the smoothing parameter. The efficiency of smoothing methods depends on the smooth approximation function, the solution method for the smooth optimization problem (2) and the updating scheme for the smoothing parameter μ_k. For example, if f is level-bounded and the smoothing function satisfies

$$f(x) \le \tilde{f}(x, \mu) \qquad \forall\, x \in R^n, \quad \forall\, \mu > 0 \tag{3}$$

then problem (2) has a solution for any fixed $\mu_k > 0$. In section 3, we show that a class of smoothing functions satisfy (3).

This paper is organized as follows. In section 2, we describe three motivational problems for the study of smoothing methods for (1): regularized minimization problems, eigenvalue optimization problems and stochastic equilibrium problems. In section 3, we demonstrate how to define smoothing functions for the three motivational problems by using the structure of the problems and the composition of smoothing functions for the plus function $(t)_+$. In section 4, we summarize properties of smoothing functions and the subdifferential associated with a smoothing function. In particular, we consider the relation between the Clarke subdifferential

$$\partial f(x) = \mathrm{con}\{v \,|\, \nabla f(z) \to v,\ f \text{ is differentiable at } z,\ z \to x\}$$

and the subdifferential associated with a smoothing function

$$G_{\tilde{f}}(x) = \mathrm{con}\{v \,|\, \nabla_x \tilde{f}(x^k, \mu_k) \to v,\ \text{for } x^k \to x,\ \mu_k \downarrow 0\},$$

where "con" denotes the convex hull. According to Theorem 9.61 and (b) of Corollary 8.47 in the book of Rockafellar and Wets [94], if f is locally Lipschitz, then $G_{\tilde{f}}(x)$ is nonempty and bounded, and $\partial f(x) \subseteq G_{\tilde{f}}(x)$. We show the gradient consistency

$$\partial f(x) = G_{\tilde{f}}(x) \tag{4}$$

holds for locally Lipschitz functions in the three motivational problems and a class of smoothing functions. Moreover, we show the affine-scaled gradient consistency for a class of non-Lipschitz functions. In section 5, we use a simple smoothing gradient method to illustrate how to update the smoothing parameter in the algorithms to guarantee the convergence of smoothing methods. In section 6, we illustrate numerical implementation of smoothing methods using the smoothing SQP algorithm [6] for solving ℓ_2-$\ell_p(0 < p < 1)$ minimization.

2 Motivational Problems

In this section, we describe three classes of nonsmooth, nonconvex minimization problems which are of the form (1) and can be solved efficiently by the smoothing methods studied in this paper.

2.1 Regularized minimization problems

$$\begin{array}{ll} \min & \Theta(x) + \lambda\Phi(x) \\ \text{s.t.} & Ax = b, \quad \ell \le x \le u, \end{array} \tag{5}$$

where $\Theta : R^n \to R_+$ is a convex function, $\Phi : R^n \to R_+$ is a continuous function, $A \in R^{m\times n}$, $b \in R^m$, $\ell \in \{R \cup \{-\infty\}\}^n$, $u \in \{R \cup \{\infty\}\}^n$, $(\ell \le u)$ are given matrices and vectors. Problem (5) is a nonlinear programming problem with linear constraints, which includes the unconstrained optimization problem as a special case. The matrix A has often simple structure. For instance $A = (1, \ldots, 1) \in R^{1\times n}$ and $b = 1$ in Markowitz mean-variance model for portfolio selection. In the objective, Θ forces closeness to data, Φ pushes the solution x to exhibit some priori expected features and $\lambda > 0$ is a parameter that controls the trade-off between the data-fitting term and the regularization term. A class of regularization terms is of the form

$$\Phi(x) = \sum_{i=1}^{r} \varphi(d_i^T x), \tag{6}$$

where $d_i \in R^n$, $i = 1, \ldots, r$ and $\varphi : R \to R_+$ is a continuous function. The regularization term Φ is also called a potential function [85] in image sciences and a penalty function [42] in statistics. The following fitting functions and regularization functions are widely used in practice.

- least squares: $\Theta(x) = \|Hx - c\|_2^2$,
- censored least absolute deviations [83]: $\Theta(x) = \|(Hx)_+ - c\|_1$,
- bridge penalty [61,67]: $\varphi(t) = |t|^p$,
- smoothly clipped absolute deviation (SCAD) [42]:
$$\varphi(t) = \int_0^{|t|} \min(1, (\alpha - s/\lambda)_+/(\alpha - 1))ds,$$
- minimax concave penalty (MCP) [106]: $\varphi(t) = \int_0^{|t|} (1 - s/(\alpha\lambda))_+ ds$,

– fraction penalty [52]: $\varphi(t) = \alpha|t|/(1+\alpha|t|)$,
– logistic penalty [84]: $\varphi(t) = \log(1+\alpha|t|)$,
– hard thresholding penalty function [41]: $\varphi(t) = \lambda - (\lambda - |t|)_+^2/\lambda$,

where $H \in R^{m_1 \times n}$ and $c \in R^{m_1}$ are given data, and α and p are positive constants. For the bridge penalty function, smoothness and convexity are dependent on the value of p. In particular, $|t|^p$ $(p > 1)$ is smooth, convex, $|t|^p$ $(p = 1)$ is nonsmooth, convex, and $|t|^p$ $(0 < p < 1)$ is non-Lipschitz, nonconvex. The other five regularization functions are nonsmooth, nonconvex.

There is considerable evidence that using nonsmooth, nonconvex regularization terms can provide better reconstruction results than using smooth convex or nonsmooth convex regularization functions, for piecewise constant image, signal reconstruction problems, variable selection and high dimensional estimations. Moreover, using non-Lipschitz regularization functions can produce desirable sparse approximations. See [13,30,42,61,62,84]. However, finding a global minimizer of (5) with a non-Lipschitz regularization function can be strongly NP-hard [23].

It is worth noting that the regularization term in (6) can be generalized to

$$\Phi(x) = \sum_{i=1}^{r} \varphi_i(d_i^T x) \tag{7}$$

and

$$\Phi(x) = \sum_{i=1}^{r} \varphi_i(D_i x_i), \tag{8}$$

where $D_i \in R^{\nu_i \times \gamma_i}$, $x_i \in R^{\gamma_i}$ and $\varphi_i(D_i x_i) = \log(1+\alpha\|D_i x_i\|)$ or $\varphi_i(D_i x_i) = (\|D_i x_i\|_1)^p$. In (7), various functions φ_i are used instead of using a single function φ. In (8) variables are divided into various groups which can be used for both group variable selections and individual variable selections [62].

2.2 Eigenvalue optimization

$$\min_x \; g(\Lambda(C + Y(x)^T Y(x))), \tag{9}$$

where $g : R^m \to R$ is a continuously differentiable function, C is an $m \times m$ symmetric positive semidefinite matrix and $Y(x)$ is an $r \times m$ matrix for any given $x \in R^n$. Suppose that each element Y_{ij} of Y is a continuously differentiable function from R^n to R. Let

$$B(x) = C + Y(x)^T Y(x).$$

Then $B(x)$ is an $m \times m$ symmetric positive semidefinite matrix and each element is a continuously differentiable function of x. We use

$$\Lambda(B(x)) = (\lambda_1(B(x)), \ldots, \lambda_m(B(x)))$$

to denote the vector of eigenvalues of $B(x)$. We assume that the eigenvalues are ordered in a decreasing order:

$$\lambda_1(B(x)) \geq \lambda_2(B(x)) \geq \ldots \geq \lambda_m(B(x)).$$

Consider the following eigenvalue optimization problems.

(i) Minimizing the condition number

$$g(\Lambda(B(x))) = \frac{\lambda_1(B(x))}{\lambda_m(B(x))}.$$

(ii) Minimizing the product of the k largest eigenvalues

$$g(\Lambda(B(x))) = \prod_{i=1}^{k} \lambda_i(B(x)).$$

(iii) Maximizing the minimal eigenvalue

$$g(\Lambda(B(x))) = -\lambda_m(B(x)).$$

Such nonsmooth, nonconvex problems arise from optimal control, design of experiments and distributions of points on the sphere [10,29,71,76].

2.3 Stochastic complementarity problems

Let $F : R^n \to R^n$ be a continuously differentiable function. The following nonsmooth and nonconvex minimization problem

$$\min_x \ \theta(x) := \| \min(x, F(x))\|_2^2 = \sum_{i=1}^{n} (\min(x_i, F_i(x)))^2 \tag{10}$$

is equivalent to the nonlinear complementarity problem, denoted by NCP(F),

$$0 \leq x \perp F(x) \geq 0, \tag{11}$$

in the sense that the solution sets of (10) and (11) coincide with each other if the NCP(F) has a solution.

The function $\theta : R^n \to R_+$ is a residual function of the NCP(F). A function $\gamma : R^n \to R$ is called a residual function (also called a merit function) of the NCP(F), if $\gamma(x) \geq 0$ for all $x \in R^n$ and $\gamma(x^*) = 0$ if and only if x^* is a solution of the NCP(F). There are many residual functions of the NCP(F), see [36, 40]. The function θ in (10) is called a natural residual function. We use θ here only for simplicity of explanation.

When F involves stochastic parameters $\xi \in \Xi \subseteq R^\ell$, in general, we can not find an x such that

$$0 \leq x \perp F(\xi, x) \geq 0, \qquad \forall\, \xi \in \Xi, \tag{12}$$

equivalently, we can not find an x such that

$$\theta(\xi, x) := \| \min(x, F(\xi, x)) \|_2^2 = 0 \qquad \forall\, \xi \in \Xi.$$

We may consider a variety of reformulations. For example, find $x \in R^n_+$ such that

$$\text{Prob}\big\{\theta(\xi, x) = 0\big\} \geq \alpha \qquad \text{or} \qquad 0 \leq x \perp E[F(\xi, x)] \geq 0,$$

where $\alpha \in (0, 1)$ and $E[\cdot]$ denotes the expected value over Ξ, a set representing future states of knowledge. The first formulation takes on the form of a "chance constraint". The second one is called Expected Value (EV) formulation, which is a deterministic NCP($\bar{F}$) with the expectation function $\bar{F}(x) = E[F(\xi, x)]$. See [53,65,95,100]. Using the residual function θ, the EV formulation can be equivalently written as a minimization problem

$$\min_x \| \min(x, E[F(\xi, x)]) \|_2^2. \tag{13}$$

There are other two ways to reformulate the stochastic NCP by using a residual function $\theta(\xi, x)$ as the recourse cost, which depends on both the random event ξ and the first-period decision x. By the definition of a residual function, the value of $\theta(\xi, x)$ reflexes the cost due to failure in satisfying the equilibrium conditions at x and ξ. Minimizing the expected values of cost in all possible scenarios gives the expected residual minimization (ERM) for the stochastic NCP

$$\min_{x \geq 0} E[\theta(\xi, x)]. \tag{14}$$

The ERM formulation (14) was introduced in [21] and applied to pricing American options, local control of discontinuous dynamics and transportation assignment [54,98,105]. Mathematical analysis and practice examples show that the ERM formulation is robust in the sense that its solutions have a minimal sensitivity with respect to random parameter variations in the model [32,43, 74,105].

On the other hand, taking on the form of robust optimization (best worst case) [3] gives

$$\min_x \max_{\xi \in \Xi} \ \theta(\xi, x). \tag{15}$$

In the reformulations above, we need prior commitments of probability distributions. However, in practice, we can only guess the distribution of ξ with limited information. To find a robust solution, we can adapt the ideas of "distributionally robust optimization" [38], which take into account knowledge about distribution's support and a confidence region for its mean and its covariance matrix. In particular, we define a set $\mathcal{P}$ of possible probability distributions that is assumed to include the true ρ_ξ, and the objective function is reformulated with respect to the worst case expect loss over the choice of a distribution in $\mathcal{P}$. Distributionally robust complementarity problems can be reformulated as

$$\min_{x \geq 0} \max_{\rho_\xi \in \mathcal{P}} E[\theta(\xi, x)]. \tag{16}$$

As the complementarity problem is a special subclass of the variational inequality problem, the stochastic complementarity problem is also a special subclass of the stochastic variational inequality problem. The EV reformulation and the ERM reformulation for finding a "here and now" solution x, for the stochastic variational inequality problem:

$$(y-x)^T F(\xi,x) \geq 0, \quad \forall y \in X_\xi, \quad \forall \xi \in \Xi,$$

can be found in [28,95], where X_ξ is a stochastic convex set.

3 Smooth approximations

The systematic study of nonsmooth functions has been a rich area of mathematical research for three decades. Clarke [34] introduced the notion of generalized gradient $\partial f(x)$ for Lipschitz continuous functions. Comprehensive studies of more recent developments can be found in [80,94]. The Clarke gradient and stationarity have been widely used in the construction and analysis of numerical methods for nonsmooth optimization problems. In addition to the study of general nonsmooth optimization, there is a large literature on more specialized problems, including semismooth and semiconvex functions [79], nonconvex polyhedral functions [87], composite nonsmooth functions [9,99,102] and piecewise smooth functions [93]. Furthermore, the study of smooth approximations for various specialized nonsmooth optimization and exact penalty functions has a long history [14,17,24,27,31,33,40,59,70,81,82,90,94].

In this section, we consider a class of nonsmooth functions which can be expressed by composition of the plus function $(t)_+$ with smooth functions. All motivational problems in section 2 belong to this class.

Chen and Mangasarian construct a class of smooth approximations of the function $(t)_+$ by convolution [17,90,94] as follows. Let $\rho : R \to R_+$ be a piecewise continuous density function satisfying

$$\rho(s) = \rho(-s) \quad \text{and} \quad \kappa := \int_{-\infty}^{\infty} |s|\rho(s)ds < \infty.$$

Then

$$\phi(t,\mu) := \int_{-\infty}^{\infty} (t-\mu s)_+ \rho(s)ds \tag{17}$$

from $R \times R_+$ to R_+ is well defined. Moreover, for any fixed $\mu > 0$, $\phi(\cdot,\mu)$ is continuously differentiable, convex, strictly increasing, and satisfies [14]

$$0 \leq \phi(t,\mu) - (t)_+ \leq \kappa\mu. \tag{18}$$

Inequalities in (18) imply that for any $t \in R$,

$$\lim_{t_k \to t, \mu \downarrow 0} \phi(t_k,\mu) = (t)_+. \tag{19}$$

Hence ϕ is a smoothing function of $(t)_+$ by Definition 1. Moreover, from $0 \leq \nabla_t \phi(t,\mu) = \int_{-\infty}^{\frac{t}{\mu}} \rho(s)ds \leq 1$, we have

$$\lim_{t_k \to t, \mu_k \downarrow 0} \nabla_t \phi(t_k, \mu_k) \in \partial(t)_+ = \begin{cases} 1 & \text{if } t > 0 \\ 0 & \text{if } t < 0 \\ [0,1] & \text{if } t = 0, \end{cases} \tag{20}$$

where $\partial(t)_+$ is the Clarke subdifferential of $(t)_+$. The subdifferential associated with the smoothing function ϕ at a point t is

$$G_\phi(t) = \text{con}\{\tau \mid \nabla_t \phi(t_k, \mu_k) \to \tau,\ t_k \to t,\ \mu_k \downarrow 0\}.$$

At $t = 0$, taking two sequences $t_k \downarrow 0$ and $t_k \uparrow 0$, and setting $\mu_k = t_k^2$, from $0 \leq \nabla_t \phi(t_k, \mu_k) = \int_{-\infty}^{\frac{t_k}{\mu_k}} \rho(s)ds \leq 1$, we find $G_\phi(0) = [0,1]$. Hence the Clarke subdifferential coincides with the subdifferential associated with the smoothing function ϕ, namely, we have

$$G_\phi(t) = \partial(t)_+. \tag{21}$$

Many nonsmooth optimization problems including all motivational problems in section 2 can be reformulated by using the plus function $(t)_+$. We list some problems as follows

$$\begin{aligned} |x| &= (x)_+ + (-x)_+ \\ \max(x,y) &= x + (y-x)_+ \\ \min(x,y) &= x - (x-y)_+ \\ \text{mid}(x,\ell,u) &= \min(\max(\ell,x),u), \quad \text{for given} \quad \ell, u \end{aligned}$$

$$x_{r_1}x_{r_2}x_{r_3} = \max\{x_i x_j x_k : \quad i < j < k, \quad i,j,k \in \{1,\ldots,n\}\},$$

$$\text{where } x_{r_1} \geq \ldots \geq x_{r_n} \geq 0.$$

We can choose a smooth approximation of $(t)_+$ to define smooth approximations for these nonsmooth functions and their compositions.

Example 1 Consider minimizing the following function

$$f(x) = \sum_{i=1}^{r} |(h_i^T x)_+ - c_i|^p,$$

in censored least absolute deviations [83], where $p > 0$ and $h_i \in R^n$, $c_i \in R$, $i = 1,\ldots,r$.

Choose $\rho(s) = \begin{cases} 0 & \text{if } |s| \geq \frac{1}{2} \\ 1 & \text{if } |s| < \frac{1}{2}. \end{cases}$ Then

$$\phi(t,\mu) = \int_{-\infty}^{\infty} (t-\mu s)_+ \rho(s) ds = \begin{cases} (t)_+ & \text{if } |t| \geq \frac{\mu}{2} \\ \frac{t^2}{2\mu} + \frac{t}{2} + \frac{\mu}{8} & \text{if } |t| < \frac{\mu}{2} \end{cases}$$

is a smoothing function of $(t)_+$. This function is called the uniform smoothing function.

From $|t| = (t)_+ + (-t)_+$, we know

$$\psi(t,\mu) := \phi(t,\mu) + \phi(-t,\mu) = \begin{cases} |t| & \text{if } |t| \geq \frac{\mu}{2} \\ \frac{t^2}{\mu} + \frac{\mu}{4} & \text{if } |t| < \frac{\mu}{2}. \end{cases}$$

is a smoothing function of $|t|$ and

$$\tilde{f}(x,\mu) = \sum_{i=1}^{r} \psi(\phi(h_i^T x, \mu) - c_i, \mu)^p$$

is a smoothing function of $f(x) = \sum_{i=1}^{r} |(h_i^T x)_+ - c_i|^p$.

Example 2 [29,76] Consider minimizing the condition number of a symmetric positive definite matrix $B(x)$ in section 2.2,

$$f(x) = \frac{\lambda_1(B(x))}{\lambda_m(B(x))}.$$

Let $h(x) = \max(x_1, \ldots, x_n)$. Choose $\rho(s) = e^{-s}/(1+e^{-s})^2$. Then

$$\phi(t,\mu) = \int_{-\infty}^{\infty} (t-\mu s)_+ \rho(s) ds = \begin{cases} \mu \ln(1+e^{t/\mu}) & \text{if } \mu > 0 \\ (t)_+ & \text{if } \mu = 0 \end{cases}$$

is a smoothing function of $(t)_+$. This function is called the Neural Networks smoothing function.

For $n = 2$, from $\max(x_1, x_2) = x_1 + (x_2 - x_1)_+$, we know

$$x_1 + \phi(x_2 - x_1, \mu) = x_1 + \mu \ln(e^{(x_2-x_1)/\mu} + 1) = \mu \ln(e^{x_1/\mu} + e^{x_2/\mu})$$

for $\mu > 0$. Hence $x_1 + \phi(x_2 - x_1, \mu)$ is a smoothing function of $\max(x_1, x_2)$.

Suppose $\mu \ln \sum_{i=1}^{n-1} e^{x_i/\mu}$ for $\mu > 0$ is a smoothing function of $\max(x_1, \ldots, x_{n-1})$. Then from $\max(x_1, \ldots, x_{n-1}, x_n) = \max(x_n, \max(x_1, \ldots, x_{n-1}))$, we find

$$\tilde{h}(x,\mu) = \mu \ln \sum_{i=1}^{n} e^{x_i/\mu} \approx \mu \ln(e^{x_n/\mu} + e^{\max(x_1,\ldots,x_{n-1})/\mu})$$

for $\mu > 0$ is a smoothing function of $h(x) = \max(x_1, \ldots, x_n)$.

Moreover, from $\min(x_1,\ldots,x_n) = -\max(-x_1,\ldots,-x_n)$, we have $-\tilde{h}(-x,\mu) = -\mu\ln\sum_{i=1}^n e^{-x_i/\mu}$ for $\mu > 0$ is a smoothing function of $\min(x_1,\ldots,x_n)$. Hence, we get

$$\tilde{f}(x,\mu) = \begin{cases} -\frac{\ln\sum_{i=1}^n e^{\lambda_i(B(x))/\mu}}{\ln\sum_{i=1}^n e^{-\lambda_i(B(x))/\mu}} & \text{if} \quad \mu > 0 \\ f(x) & \text{if} \quad \mu = 0 \end{cases}$$

is a smoothing function of $f(x) = \frac{\lambda_1(B(x))}{\lambda_m(B(x))}$. In computation, we use a stable form

$$\tilde{f}(x,\mu) = \frac{\lambda_1(B(x)) + \mu\ln\sum_{i=1}^n e^{(\lambda_i(B(x))-\lambda_1(B(x)))/\mu}}{\lambda_m(B(x)) - \mu\ln\sum_{i=1}^n e^{(\lambda_m(B(x))-\lambda_i(B(x)))/\mu}}$$

for $\mu > 0$.

Example 3 [21,32,104] Consider the stochastic nonlinear complementarity problem and let

$$f(x) = E[\|\min(x, F(\xi,x))\|_2^2].$$

Choose $\rho(s) = \frac{2}{(s^2+4)^{\frac{3}{2}}}$. Then

$$\phi(t,\mu) = \int_{-\infty}^{\infty}(t-\mu s)_+\rho(s)ds = \frac{1}{2}(t+\sqrt{t^2+4\mu^2})$$

is a smoothing function of $(t)_+$. This smoothing function is called the CHKS(Chen-Harker-Kanzow-Smale) smoothing function [15,66,96].

From $\min(x, F(\xi,x)) = x - (x - F(\xi,x))_+$, we have

$$\tilde{f}(x,\mu) = \frac{1}{2}E[\sum_{i=1}^n \left(x_i + F_i(\xi,x) - \sqrt{(x_i - F_i(\xi,x))^2 + 4\mu^2}\right)^2]$$

is a smoothing function of $f(x) = E[\|\min(x, F(\xi,x))\|_2^2]$.

Remark 1 The plus function $(t)_+$ has been widely used for penalty functions, barrier functions, complementarity problems, semi-infinite programming, optimal control, mathematical programs with equilibrium constraints etc. Examples of the class of smooth approximations ϕ of $(t)_+$ defined by (17) can be found in a number of articles, we refer to the book [40] and [2,5,14,16–18,22, 24,27–31,33,39,64,92,104].

Remark 2 Some smoothing functions can not be expressed by the Chen-Mangasarian smoothing function. For example, the smoothing function $\frac{1}{2}(t+s-\sqrt{t^2+s^2+2\mu^2})$ of the Fischer-Bumeister function [45] $\frac{1}{2}(t+s-\sqrt{t^2+s^2})$ proposed by Kanzow [66] for complementarity problems, and the smoothing function

$$\phi(t,\mu) = \begin{cases} \frac{1}{p}|t|^p - (\frac{1}{p} - \frac{1}{2})\mu^p & \text{if} \quad |t| > \mu \\ \frac{1}{2}\mu^{p-2}t^2 & \text{if} \quad 0 \le |t| \le \mu \end{cases}$$

of $|t|^p$ proposed by Hintermüller and Wu [58] for the ℓ_p-norm regularized minimization problems where $0 < p < 1$.

4 Analysis of smoothing functions

The plus function $(t)_+$ is convex and globally Lipschitz continuous. Any smoothing function $\phi(t,\mu)$ of $(t)_+$ defined by (17) is also convex and globally Lipschitz and has nice properties (18)-(21). In addition, for any fixed t, ϕ is continuously differentiable, monotonically increasing and convex with respect to $\mu > 0$ and satisfies

$$0 \le \phi(t,\mu_2) - \phi(t,\mu_1) \le \kappa(\mu_2 - \mu_1), \quad \text{for} \quad \mu_2 \ge \mu_1. \tag{22}$$

See [14]. These properties are important for developing efficient smoothing methods for nonsmooth optimization. In this section, we study properties of the smoothing functions defined by composition of the smoothing function $\phi(t,\mu)$. Specially, we investigate the subdifferential associated with a smoothing function in such class of smoothing functions.

4.1 Locally Lipschitz continuous functions

In this subsection, we assume that f is locally Lipschitz continuous. According to Rademacher's theorem, f is differentiable almost everywhere. The Clarke subdifferential of f at a point x is defined by

$$\partial f(x) = \mathrm{con}\partial_B f(x),$$

where

$$\partial_B f(x) = \{v \,|\, \nabla f(z) \to v,\ f \text{ is differentiable at } z,\ z \to x\}.$$

For a locally Lipschitz function f, the gradient consistency

$$\partial f(x) = \mathrm{con}\{\lim_{x^k \to x, \mu_k \downarrow 0} \nabla_x f(x^k, \mu_k)\} = G_{\tilde{f}}(x), \quad \forall x \in R^n \tag{23}$$

between the Clarke subdifferential and the subdifferential associated with a smoothing function of f is important for the convergence of smoothing methods. Rockafellar and Wets [94] show that for any locally Lipschitz function f, we can construct a smoothing function by using convolution

$$\tilde{f}(x,\mu) = \int_{R^n} f(x-y)\psi_\mu(y)dy = \int_{R^n} f(y)\psi_\mu(x-y)dy, \tag{24}$$

where $\psi : R^n \to R$ is a smooth kernel function (a mollifier), which has the gradient consistency (23).

Now we show the gradient consistency of smoothing composite functions using ϕ in (17) for the plus function. For a vector function $h(x) = (h_1(x), \ldots, h_m(x))^T$ with components $h_i : R^n \to R$, we denote $(h(x))_+ = ((h_1(x))_+, \ldots, (h_m(x))_+)^T$ and

$$\phi(h(x),\mu) = (\phi(h_1(x),\mu), \ldots, \phi(h_m(x),\mu))^T.$$

Theorem 1 *Let $f(x) = g((h(x))_+)$, where $h : R^n \to R^m$ and $g : R^m \to R$ are continuously differentiable functions. Then $\tilde{f}(x,\mu) = g(\phi(h(x),\mu))$ is a smoothing function of f with the following properties.*

(i) For any x, $\{\lim_{x^k \to x, \mu_k \downarrow 0} \nabla_x \tilde{f}(x^k, \mu_k)\}$ is nonempty and bounded, and $\partial f(x) = G_{\tilde{f}}(x)$.

(ii) If g, h_i are convex and g is monotonically nondecreasing, then for any fixed $\mu > 0$, $\tilde{f}(\cdot, \mu)$ is convex.

Proof By the chain rule for continuously differentiable functions, $\tilde{f}$ is a smoothing function of f with

$$\nabla_x \tilde{f}(x,\mu) = \nabla h(x) \mathrm{diag}(\nabla_y \phi(y,\mu)|_{y=h_i(x)}) \nabla g(z)|_{z=\phi(h(x),\mu)}.$$

(i) Taking two sequences $x^k \to x$ and $\mu_k \downarrow 0$, we have

$$\begin{aligned}
&\{\lim_{x^k \to x, \mu_k \downarrow 0} \nabla_x \tilde{f}(x^k, \mu_k)\} \\
&= \{\lim_{x^k \to x, \mu_k \downarrow 0} \nabla h(x^k) \mathrm{diag}(\nabla_y \phi(y^k, \mu_k)|_{y^k = h_i(x^k)}) \nabla g(z^k)|_{z^k = \phi(h(x^k), \mu_k)}\} \\
&\subseteq \nabla h(x) \mathrm{diag}(\partial (h_i(x))_+) \nabla g(z)|_{z=(h(x))_+}.
\end{aligned}$$

Since $(t)_+$ is monotonically nondecreasing and h is continuously differentiable, by Proposition 2.3.6 in [34], $(h(x))_+$ is Clarke regular. From Theorem 2.3.9 in [34], we have

$$\nabla h(x) \mathrm{diag}(\partial (h_i(x))_+) \nabla g(z)|_{z=(h(x))_+} = \partial f(x).$$

Hence, we obtain

$$G_{\tilde{f}}(x) = \mathrm{con}\{\lim_{x^k \to x, \mu_k \downarrow 0} \nabla_x \tilde{f}(x^k, \mu_k)\} \subseteq \partial f(x).$$

By the continuous differentiability of g, h, the function f is locally Lipschitz, and $\partial f(x)$ is bounded. On the other hand, according to Theorem 9.61 and (b) of Corollary 8.47 in [94], $\partial f(x) \subseteq G_{\tilde{f}}(x)$. Hence we obtain $\partial f(x) = G_{\tilde{f}}(x)$.

(ii) For any fixed μ, the smoothing function $\phi(t,\mu)$ of the plus function $(t)_+$ is convex and monotonically nondecreasing. Hence for any $\lambda \in (0,1)$, we have

$$\begin{aligned}
\tilde{f}(\lambda x + (1-\lambda)y, \mu) &= g(\phi(h(\lambda x + (1-\lambda)y), \mu)) \\
&\le g(\phi(\lambda h(x) + (1-\lambda)h(y), \mu)) \\
&\le g(\lambda \phi(h(x),\mu) + (1-\lambda)\phi(h(y),\mu)) \\
&\le \lambda g(\phi(h(x),\mu)) + (1-\lambda) g(\phi(h(y),\mu)) \\
&= \lambda \tilde{f}(x,\mu) + (1-\lambda)\tilde{f}(y,\mu).
\end{aligned}$$

Corollary 1 *Let $f(x) = (g((h(x))_+))_+$, where $h : R^n \to R^m$ and $g : R^m \to R$ are continuously differentiable functions. Then $\tilde{f}(x,\mu) = \phi(g(\phi(h(x),\mu)),\mu)$ is a smoothing function of f with the following properties.*

(i) For any x, $\{\lim_{x^k\to x,\mu_k\downarrow 0}\nabla_x\tilde{f}(x^k,\mu_k)\}$ is nonempty and bounded, and $\partial f(x)=G_{\tilde{f}}(x)$.
(ii) If g,h_i are convex and g is monotonically nondecreasing, then for any fixed $\mu>0$, $\tilde{f}(\cdot,\mu)$ is convex.

Proof By Proposition 2.3.6 in [34], $(h(x))_+$, $g((h(x))_+)$ and $f(x)$ are Clarke regular. From Theorem 2.3.9 in [34] and Proposition 1, we derive this corollary.

Many useful properties of the Clarke subdifferential can be applied to smoothing functions. For example, the following proposition is derived from Theorem 2.3.9 in [34] for the Chain rules.

Proposition 1 *[5] Let $\tilde{g}$ and $\tilde{h}$ be smoothing functions of locally Lipschitz functions $g:R^m\to R$ and $h:R^n\to R^m$ with $\partial g(x)=G_{\tilde{g}}(x)$ and $\partial h(x)=G_{\tilde{h}}(x)$, respectively. Set $\tilde{g}=g$ or $\tilde{h}=h$ when g or h is differentiable. Then $\tilde{f}=\tilde{g}(\tilde{h})$ is a smoothing function of $f=g(h)$ with $\partial f(x)=G_{\tilde{f}}(x)$ under any one of the following conditions for any $x\in R^n$.*

1. *g is regular at $h(x)$, each h_i is regular at x and $\nabla_z\tilde{g}(z,\mu)|_{z=h(x)}\geq 0$.*
2. *g is continuously differentiable at $h(x)$ and $m=1$.*
3. *g is regular at $h(x)$ and h is continuously differentiable.*

Using the chain rules for smoothing functions, we can quickly find that the smoothing functions $\tilde{f}$ in Example 1 with $p\geq 1$ and Examples 2-3, satisfies $\partial f(x)=G_{\tilde{f}}(x)$.

4.2 Non-Lipschitz continuous functions

Lipschitz continuity is important in nonsmooth optimization. Generalized gradient and Jacobian of Lipschitz continuous functions have been well studied [24,34,80,94]. However, we have lack of theory and algorithms for non-Lipschitz optimization problems. Recently, non-Lipschitz optimization problems have attracted considerable attention from variable selection, sparse approximation, signal reconstruction, image restoration, exact penalty functions, eigenvalue optimization, etc [42,62,70,75,78]. Reformulation or approximation of non-Lipschitz continuous functions by Lipschitz continuous functions have been proposed in [1,19,70,107]. In this subsection, we show that smoothing functions defined by composition of smoothing functions (17) of the plus function for a class of non-Lipschitz continuous functions have both differentiability and Lipschitz continuity.

We consider the following non-Lipschitz function

$$f(x)=\|(g(x))_+\|_p^p+\|h(x)\|_p^p=\sum_{i=1}^r((g_i(x))_+)^p+\sum_{i=1}^m|h_i(x)|^p, \qquad (25)$$

where $g:R^n\to R^r$ and $h:R^n\to R^m$ are continuously differentiable functions and $0<p<1$. This function includes ℓ_p norm regularization and ℓ_p exact

penalty functions for $0 < p < 1$ as special cases [13,30,42,62,70,75,78]. It is easy to see that f is a continuous function from R^n to R_+. Moreover, f is continuously differentiable at x if all components of g and h at x are nonzero. However, f is possibly non-Lipschitz continuous at x, if one of components of g or h at x is zero.

We use a smoothing function $\phi(t,\mu)$ in the class of Chen-Mangasarian smoothing functions (17) for $(t)_+$ and $\psi(t,\mu) = \phi(t,\mu) + \phi(-t,\mu)$ for $|t| = (t)_+ + (-t)_+$ to define a smoothing function of f as the following

$$\tilde{f}(x,\mu) = \sum_{i=1}^{r} (\phi(g_i(x),\mu))^p + \sum_{i=1}^{m} (\psi(h_i(x),\mu))^p. \tag{26}$$

Lemma 1 *$(\phi(t,\mu))^p$ and $(\psi(t,\mu))^p$ are smoothing functions of $((t)_+)^p$ and $|t|^p$ in R_+ and R, respectively. Moreover, the following statements hold.*

(i) Let $\kappa_0 = (\frac{\kappa}{2})^p$ where $\kappa := \int_{-\infty}^{\infty} |s|\rho(s)ds$. Then for any $t \in R$,

$$0 \le (\phi(t,\mu))^p - ((t)_+)^p \le \kappa_0\mu^p \quad \text{and} \quad 0 \le (\psi(t,\mu))^p - |t|^p \le 2\kappa_0\mu^p.$$

(ii) For any fixed $\mu > 0$, $(\phi(t,\mu))^p$ and $(\psi(t,\mu))^p$ are Lipschitz continuous in R_+ and R, respectively. In particular, their gradients are bounded by $p(\phi(0,\mu))^{p-1}$, that is,

$$0 \le p(\phi(t,\mu))^{p-1}\nabla_t\phi(t,\mu) \le p(\phi(0,\mu))^{p-1}, \quad \text{for} \quad t \in R_+$$

and

$$|p(\psi(t,\mu))^{p-1}\nabla_t\psi(t,\mu)| \le 2p(\phi(0,\mu))^{p-1}, \quad \text{for} \quad t \in R.$$

(iii) Assume there is $\rho_0 > 0$ such that $|\rho(s)| \le \rho_0$. Let

$$\nu_\mu = p(1-p)\phi(0,\mu)^{p-2} + \frac{1}{\mu}p\phi(0,\mu)^{p-1}\rho_0$$

then for any fixed $\mu > 0$, the gradients of $(\phi(t,\mu))^p$ and $(\psi(t,\mu))^p$ are Lipschitz continuous in R_+ and R, with Lipschitz constants ν_μ and $2\nu_\mu$, respectively. In particular, if ρ is continuous at t then $|\nabla_t^2(\phi(t,\mu))^p| \le \nu_\mu$ for $t \in R_+$ and $|\nabla_t^2(\psi(t,\mu))^p| \le 2\nu_\mu$ for $t \in R$.

In addition, if $\rho(s) > 0$ for $s \in (-\infty,\infty)$, then $(\phi(t,\mu))^p$ is a smoothing function of $((t)_+)^p$ in R, and for any fixed $\mu > 0$, $(\phi(t,\mu))^p$ and its gradient are Lipschitz continuous in R.

Proof We first prove this lemma for $(\phi(t,\mu))^p$. Since $\phi(t,\mu)$ is a smoothing function of $(t)_+$ and $\nabla_t\phi(t,\mu) > 0$ in R_+, $(\phi(t,\mu))^p$ is a smoothing function of $((t)_+)^p$.

(i) Since t^p is a monotonically increasing function in R_+, from (18), we have $(\phi(t,\mu))^p - ((t)_+)^p \ge 0$. Moreover, from

$$\nabla_\mu\phi(t,\mu) = -\int_{-\infty}^{t/\mu} s\rho(s)ds, \quad \text{for}\, t \ge 0,$$

the difference between $(\phi(t,\mu))^p$ and $((t)_+)^p$ has the maximal value at $t=0$, that is,

$$0 \le (\phi(t,\mu))^p - ((t)_+)^p \le (\phi(0,\mu))^p = (\frac{\kappa}{2})^p \mu^p.$$

(ii)-(iii) Note that $0 \le \nabla_t \phi(t,\mu) = \int_{-\infty}^{t/\mu} \rho(s)ds \le 1$. Straightforward calculation shows that for any $t \in R_+$,

$$p(\phi(t,\mu))^{p-1}\nabla_t\phi(t,\mu) \le p(\phi(0,\mu))^{p-1}\int_{-\infty}^{t/\mu}\rho(s)ds \le p(\phi(0,\mu))^{p-1}$$

and if ρ is continuous at t, then

$$|\nabla_t^2(\phi(t,\mu))^p| = |p(p-1)(\phi(t,\mu))^{p-2}(\nabla_t\phi(t,\mu))^2 + p(\phi(t,\mu))^{p-1}\nabla_t^2\phi(t,\mu)|$$

$$\le p(1-p)(\phi(0,\mu))^{p-2} + p\phi(0,\mu)^{p-1}\rho(\frac{t}{\mu})\frac{1}{\mu} \le \nu_\mu.$$

Since ρ is piecewise continuous, the gradient of $(\phi(t,\mu))^p$ is locally Lipschitz and the second derivative $\nabla_t^2(\phi(t,\mu))^p$ almost everywhere exists in R_+. By the mean value theorem in [34], we find for any $t_1, t_2 \in R_+$,

$$|\nabla_t(\phi(t_1,\mu))^p - \nabla_t(\phi(t_2,\mu))^p| \le \nu_\mu|t_1 - t_2|.$$

In addition, if $\rho(s) > 0$ for $s \in (-\infty,\infty)$, then $0 < \nabla_t\phi(t,\mu) = \int_{-\infty}^{t/\mu}\rho(s)ds < 1$. Hence, we complete the proof for $(\phi(t,\mu))^p$.

By the symmetrical characteristic of $|t| = (t)_+ + (-t)_+$ and $\psi(t,\mu) \ge 2\phi(0,\mu) = \kappa\mu > 0$ for $t \in R$, $\mu > 0$, we can prove this lemma for ψ by the same argument above.

Remark 3 If the density function $\rho(s) = 0$ for some $s \in (-\infty,\infty)$, $(\phi(t,\mu))^p$ is not necessarily a smoothing function of $((t)_+)^p$ in R. For example, the function ρ in Example 1 is a piecewise continuous function with $\rho(s) = 0$ for $|s| \ge \frac{1}{2}$. Consider the gradient of $(\phi(t,\mu))^p$ at $t = -\frac{\mu}{2}$, we have

$$\lim_{\tau\uparrow 0}\frac{(\phi(t+\tau,\mu))^p - (\phi(t,\mu))^p}{\tau} = \lim_{\tau\uparrow 0}\frac{0-0}{\tau} = 0$$

and

$$\lim_{\tau\downarrow 0}\frac{(\phi(t+\tau,\mu))^p - (\phi(t,\mu))^p}{\tau} = \lim_{\tau\downarrow 0}\frac{(\phi(t+\tau,\mu))^p}{\tau} = \lim_{\tau\downarrow 0}(\frac{1}{2\mu})^p\tau^{2p-1}.$$

Hence $(\phi(t,\mu))^p$ is a smoothing function of $((t)_+)^p$ in R if and only if $p > \frac{1}{2}$.

On the other hand, $(\psi(t,\mu))^p$ is smoothing functions of $|t|^p$ in R. Again consider the gradient of $(\psi(t,\mu))^p$ at $t = -\frac{\mu}{2}$,

$$\lim_{\tau\uparrow 0}\frac{(\phi(t+\tau,\mu))^p - (\phi(t,\mu))^p}{\tau} = p|t|^{p-1}\mathrm{sign}(t) = -p(\frac{\mu}{2})^{p-1}$$

$$= \lim_{\tau\downarrow 0}\frac{(\phi(t+\tau,\mu))^p - (\phi(t,\mu))^p}{\tau} = p(\frac{t^2}{\mu} + \frac{\mu}{4})^{p-1}\frac{2t}{\mu} = -p(\frac{\mu}{2})^{p-1}.$$

Proposition 2 *Let f and $\tilde{f}$ be defined as in (25) and (26), and let $\Omega = \{x|g(x) \geq 0\}$. Then $\tilde{f}$ is a smoothing function of f in Ω. Moreover, the following statements hold.*

(i) There is a positive constant α_0 such that

$$0 \leq \tilde{f}(x,\mu) - f(x) \leq \alpha_0 \mu^p.$$

(ii) For any fixed $\mu > 0$, $\tilde{f}(\cdot,\mu)$ and $\nabla_x \tilde{f}(\cdot,\mu)$ are locally Lipschitz continuous in Ω.

(iii) If g and h are globally Lipschitz continuous, then $\tilde{f}(\cdot,\mu)$ is globally Lipschitz continuous in Ω; If ∇g and ∇h are globally Lipschitz continuous, then $\nabla_x \tilde{f}(\cdot,\mu)$ is globally Lipschitz continuous in Ω.

In addition, if $\rho(s) > 0$ for $s \in (-\infty,\infty)$, then $\tilde{f}$ is a smoothing function of f in R^n, and (i) and (ii) hold in R^n.

Proposition 2 can be easily proved by the Chain rules and Lemma 1. Hence, we omit the proof.

Now we compare the smoothing function $\tilde{f}$ with the robust regularization

$$\bar{f}_\mu(x) = \sup\{\, f(y) \,: y \in X,\, \|x - y\| \leq \mu \,\}$$

for approximating a non-Lipschitz function $f : X \subset R^n \to R$. The robust regularization $\bar{f}_\mu$ is proposed by Lewis and Pang [70]. They use the following example

$$f(t) = \begin{cases} -t & \text{if } t < 0 \\ \sqrt{t} & \text{if } t \geq 0 \end{cases} \tag{27}$$

to illustrate the robust regularization. The function f is non-Lipschitz, since $\lim_{\epsilon \downarrow 0} \frac{f(\epsilon)-f(0)}{\epsilon - 0} = \lim_{\epsilon \downarrow 0} \frac{1}{\sqrt{\epsilon}} = \infty$. The minimizer of f is 0. The robust regularization of f is

$$\bar{f}_\mu(t) = \begin{cases} \mu - t & \text{if } t < \alpha(\mu) \\ \sqrt{\mu + t} & \text{if } t \geq \alpha(\mu), \end{cases}$$

where $\mu > \alpha(\mu) = \frac{1+2\mu-\sqrt{1+8\mu}}{2} > -\mu$. The robust regularization $\bar{f}_\mu(t)$ is Lipschitz continuous with the Lipschitz modulus $\frac{1}{2\sqrt{\mu+\alpha(\mu)}}$ for any fixed $\mu > 0$, but not differentiable. The minimizer of $\bar{f}_\mu$ is $\alpha(\mu)$ which is different from the minimizer of f.

Now we use the smoothing function $\psi(t,\mu)$ of $|t|$ in Example 1 to define a smoothing function of f in (27). To simplify the notation, we replace μ by 4μ in $\psi(t,\mu)$. Then we have

$$\tilde{f}(t,\mu) = \begin{cases} \psi(t,\mu) & \text{if } t < 0 \\ \sqrt{\psi(t,\mu)} + \mu - \sqrt{\mu} & \text{if } t \geq 0 \end{cases}$$

$$= \begin{cases} -t & \text{if} \quad t < -2\mu \\ \frac{t^2}{4\mu} + \mu & \text{if} \quad -2\mu \leq t < 0 \\ \sqrt{\frac{t^2}{4\mu} + \mu} + \mu - \sqrt{\mu} & \text{if} \quad 0 \leq t < 2\mu \\ \sqrt{t} + \mu - \sqrt{\mu} & \text{if} \quad t \geq 2\mu. \end{cases}$$

Compared with the robust regularization, the smoothing function $\tilde{f}(t,\mu)$ is not only Lipschitz continuous, but also continuously differentiable with $|\nabla_t \tilde{f}(t,\mu)| \le \frac{1}{2\sqrt{2\mu}} \le \frac{1}{2\sqrt{\mu+\alpha(\mu)}}$ for any fixed $\mu > 0$. Moreover, the minimizer of $\tilde{f}(\cdot,\mu)$ is 0, which is the same as the minimizer of f.

For a non-Lipschitz function f, we can not define the gradient of f. The scaled gradient, and scaled first order and second order necessary conditions for non-Lipschitz optimization have been studied [4,25,30]. Now we consider the scaled gradient consistency for the following non-Lipschitz function

$$f(x) := \theta(x) + \lambda \sum_{i=1}^{m} \varphi(d_i^T x), \tag{28}$$

where $\theta : R^n \to R_+$ is locally Lipschitz continuous, $\lambda \in R_+$, $d_i \in R^n, i = 1, \cdots, m$, and $\varphi : R \to R_+$ is continuously differentiable in $(-\infty, 0) \cup (0, +\infty)$. Many penalty functions satisfy the conditions, for example, the six penalty functions in subsection 2.1. For a given vector $x \in R^n$, we set

$$I_x = \{i \mid d_i^T x = 0, i = 1, \cdots, m\} \quad \text{and} \quad J_x = \{i \mid d_i^T x \neq 0, i = 1, \cdots, m\}.$$

Let Z_x be an $n \times \ell$ matrix whose columns are an orthonormal basis for the null space of $\{d_i \mid i \in I_x\}$ [25].

Let $\tilde{f}(x,\mu) = \tilde{\theta}(x,\mu) + \lambda \sum_{i=1}^m \phi(d_i^T x, \mu)$, where $\tilde{\theta}$ is a smoothing function of θ which satisfies the gradient consistency $\partial\theta(x) = G_{\tilde{\theta}}(x)$, and ϕ is a smoothing function of φ which satisfies the gradient consistency $\varphi'(t) = G_\phi(t)$ at $t \neq 0$.

Theorem 2 *For any $x \in R^n$, we have*

$$Z_x^T G_{\tilde{f}}(x) = Z_x^T \partial(\theta(x) + \lambda \sum_{i \in J_x} \varphi(d_i^T x)). \tag{29}$$

Proof If $x = 0$, then $J_x = \emptyset$, and (29) holds. If $x \neq 0$, then $\mathrm{rank}(Z_x) = \ell > 0$. Moreover, we have

$$\begin{aligned} Z_x^T \nabla_x \tilde{f}(x^k,\mu) &= Z_x^T (\nabla_x \tilde{\theta}(x^k,\mu) + \lambda \sum_{i=1}^m d_i \nabla_t \phi'(t^k,\mu)|_{t^k = d_i^T x^k}) \\ &= Z_x^T \nabla_x \tilde{\theta}(x^k,\mu) + \lambda \sum_{i \in J_x} Z_x^T d_i \nabla_t \phi'(t^k,\mu)|_{t^k = d_i^T x^k}, \end{aligned}$$

where the last equality uses $Z_x^T d_i = 0, \forall\, i \in I_x$. Since φ is continuously differentiable at $d_i^T x$ for $i \in J_x$, we obtain (29) by the gradient consistency of $\tilde{\theta}$ and ϕ.

5 Smoothing Algorithms

Numerical methods for solving nonsmooth, nonconvex optimization problems have been studied extensively [7,10,12,35,37,51,59,68,77,102]. To explain numerical implementation and convergence analysis of smoothing methods in a clear and simple way, we use a smoothing gradient method to show how to update the smoothing parameter to guarantee the global convergence.

Burke, Lewis and Overton [10] proposed a robust gradient sampling algorithm for nonsmooth, nonconvex, Lipschitz continuous unconstrained minimization problems. Basically, it is a stabilized steepest descent algorithm. At each iteration of the algorithm, a descent direction is obtained by evaluating the gradient at the current point and at additional nearby points if the function is differentiable at these points, and then computing the vector in the convex hull of these gradients with least ℓ_2 norm. A standard line search is then used to obtain a new point. If at one of these points, the function is not differentiable, the algorithm will terminate. Kiwiel [69] slightly revised this algorithm and showed that any accumulation point generated by the revised algorithm is a Clarke stationary point *with probability one*. Under the same conditions of [10,69], we show that any accumulation point generated by the smoothing gradient method is a Clarke stationary point. Moreover, we show that any accumulation point generated by the smoothing gradient method is a scaled Clarke stationary point for non-Lipschitz optimization problems.

We consider (1) with $X = R^n$. Assume that $f : R^n \to R$ is locally Lipschitz, level-bounded, and continuously differentiable on an open dense subset of R^n [10,69]. Let $\tilde{f}$ be a smoothing function with the gradient consistency (4).

Smoothing gradient method

Step 1. Choose constants $\sigma, \rho \in (0,1)$, $\gamma > 0$ and an initial point x^0. Set $k = 0$.

Step 2. Compute the gradient

$$g^k = \nabla_x \tilde{f}(x^k, \mu_k)$$

and the step size ν_k by the Armijo line search, where $\nu_k = \max\{\rho^0, \rho^1, \cdots\}$ and ρ^i satisfies

$$\tilde{f}(x^k - \rho^i g^k, \mu_k) \le \tilde{f}(x^k, \mu_k) - \sigma \rho^i (g^k)^T g^k.$$

Set $x^{k+1} = x^k - \nu_k g^k$.

Step 3. If $\|\nabla_x \tilde{f}(x^{k+1}, \mu_k)\| \ge \gamma \mu_k$, then set $\mu_{k+1} = \mu_k$; otherwise, choose $\mu_{k+1} = \sigma \mu_k$.

Theorem 3 *Any accumulation point generated by the smoothing gradient method is a Clarke stationary point.*

This theorem can be proved in a similar way in [33, Theorem 2.6]. For completeness, we give a simple proof.

Proof Denote $K = \{k \mid \mu_{k+1} = \sigma\mu_k\}$. If K is finite, then there exists an integer $\bar{k}$ such that for all $k > \bar{k}$,

$$\|\nabla_x \tilde{f}(x^{k+1}, \mu_k)\| \geq \gamma\mu_k, \tag{30}$$

and $\mu_k =: \bar{\mu}$ for all $k \geq \bar{k}$ in Step 2 of the smoothing gradient method. Since $\tilde{f}(\cdot, \bar{\mu})$ is a smooth function, the gradient method for solving

$$\min \tilde{f}(x, \bar{\mu})$$

satisfies

$$\liminf_{k\to\infty} \|\nabla_x \tilde{f}(x^k, \bar{\mu})\| = 0, \tag{31}$$

which contradicts with (30). This shows that K must be infinite and $\lim_{k\to\infty} \mu_k = 0$.

Since K is infinite, we can assume that $K = \{k_0, k_1, \ldots\}$ with $k_0 < k_1 < \ldots$. Then we have

$$\lim_{i\to\infty} \|\nabla_x \tilde{f}(x^{k_i+1}, \mu_{k_i})\| \leq \gamma \lim_{i\to\infty} \mu_{k_i} = 0. \tag{32}$$

Let $\bar{x}$ be an accumulation point of $\{x^{k_i+1}\}$. Then by the gradient consistency, we have $0 \in \partial f(\bar{x})$. Hence $\bar{x}$ is a Clarke stationary point.

Corollary 2 *Any accumulation point $\bar{x}$ generated by the smoothing gradient method for solving (28) is an affine-scaled Clarke stationary point, that is,*

$$0 \in Z_{\bar{x}}^T \partial(\theta(\bar{x}) + \lambda \sum_{i\in J_{\bar{x}}} \varphi(d_i^T \bar{x})). \tag{33}$$

Proof Following the proof of Theorem 3, we can show that there is a subsequence $K = \{k_0, k_1, \ldots\}$ with $k_0 < k_1 < \ldots$ such that (32) holds. Let $\bar{x}$ be an accumulation point of $\{x^{k_i+1}\}$. From Theorem 2, we have

$$0 = \lim_{i\to\infty} Z_{\bar{x}}^T \nabla_x \tilde{f}(x^{k_i+1}, \mu_{k_i}) \in Z_{\bar{x}}^T \partial(\theta(\bar{x}) + \lambda \sum_{i\in J_{\bar{x}}} \varphi(d_i^T \bar{x})).$$

Remark 4 Developing a smoothing method for solving a nonsmooth, nonconvex optimization problem (1) involves three main parts. (i) Define a smoothing function $\tilde{f}$ by using the methods and propositions in sections 3-4. (ii) Choose an algorithm for solving the smooth optimization problem (2) with the objective function $\tilde{f}$. For instance, the smoothing gradient method uses the gradient method in its Step 2. (iii) Update the smoothing parameter μ_k. The updating scheme will depend on the convergence of the algorithm used to solve the smooth optimization problem. As shown above, since the gradient method has the convergence property (31), we set the updating condition $\|\nabla_x \tilde{f}(x^{k+1}, \mu_k)\| < \gamma\mu_k$ in Step 3 in the smoothing gradient method. The efficiency of smoothing methods will depend on the smoothing function $\tilde{f}$, the method for solving the smooth optimization problem (2) and the scheme for updating the smoothing parameter μ_k.

There are rich theory and abundant efficient algorithms for solving smooth optimization problems [86]. With adaptive smoothing functions and updating schemes, these theory and algorithms can be powerful for solving nonsmooth optimization problems.

6 Numerical implementation

Many smoothing methods have been developed to solve nonsmooth, nonconvex optimization [4,5,18,22,24,26–31,33,39,40,46–48,54,64,66,72,73,81,82,89,90,104]. Recently, Garmanjani and Vicente [50] propose a smoothing direct search (DS) algorithm basing on smooth approximations and derivative free methods for nonsmooth, nonconvex, Lipschitz continuous unconstrained minimization problems. The smoothing DS algorithm takes at most $O(-\varepsilon^{-3}\log\varepsilon)$ function evaluations to find an x such that $\|\nabla_x\tilde{f}(x,\mu)\|_\infty \le \varepsilon$ and $\mu \le \varepsilon$. In [6], Bian and Chen propose a smoothing sequential quadratic programming (SSQP) algorithm for solving regularized minimization problem (5) where Θ is continuously differentiable but the penalty function may be non-Lipschitz. The worst-case complexity of the SSQP algorithm for finding an ε affine-scaled stationary point is $O(\varepsilon^{-2})$. Moreover, if Φ is locally Lipschitz, the SSQP algorithm with a slightly modified updating scheme can obtain an ε Clarke stationary point at most $O(\varepsilon^{-3})$ steps.

In the following, we illustrate numerical implementation of smoothing methods using the SSQP algorithm for solving the ℓ_2-ℓ_p problem

$$\min_{x\in R^n} f(x) := \|Ax-b\|_2^2 + \lambda\|x\|_p^p, \tag{34}$$

where $A \in R^{m\times n}, b \in R^m, \|x\|_p^p = \sum_{i=1}^n |x_i|^p, p \in (0,1)$. It is shown that problem (34) is strongly NP hard in [23]. Let $X = \mathrm{diag}(x)$. The affine-scaled first order necessary condition for local minimizers of (34) is

$$G(x) := 2XA^T(Ax-b) + \lambda p|x|^p = 0 \tag{35}$$

and the affine-scaled second order necessary condition is that

$$2XA^TAX + \lambda p(p-1)|X|^p \tag{36}$$

is positive semi-definite, where $|X|^p = \mathrm{diag}(|x|^p)$ [30]. Obviously, (34) is a special case of (28) and (35) is a special case of (33) with the ith column $(Z_x)_i = (X)_i/|x_i|$ for $i \in J_x$. Moreover, (35) is equivalent to

$$2(A^T(Ax-b))_i + \lambda p|x_i|^{p-1}\mathrm{sign}(x_i) = 0, \quad i \in J_x = \{i \mid x_i \neq 0\}.$$

Using (35) and (36), several good properties of the ℓ_2-ℓ_p problem (34) have been derived [23,30], including the lower bounds for nonzero entries of its local minimizers and the sparsity of its local minimizers. Moreover, a smoothing trust region Newton method for finding a point x satisfying (35) and (36) is proposed in [25].

Let us use the smoothing function

$$s(t,\mu) = \begin{cases} |t| & \text{if } |t| > \mu \\ \frac{t^2}{2\mu} + \frac{\mu}{2} & \text{if } |t| \le \mu \end{cases}$$

of $|t|$ to define the smoothing function

$$\tilde{f}(x,\mu) = \|Ax-b\|_2^2 + \lambda \sum_{i=1}^{n} s(x_i,\mu)^p$$

of f. Then we can easily find that

$$\begin{aligned}
&\lim_{x^k \to x, \mu \downarrow 0} X^k \nabla_x \tilde{f}(x^k,\mu) \\
&= \lim_{x^k \to x, \mu \downarrow 0} 2X^k A^T(Ax^k - b) + \lambda p \sum_{i=1}^{n} s(x_i^k,\mu)^{p-1} s'(x_i^k,\mu) x_i^k e_i \\
&= 2XA^T(Ax-b) + \lambda p |x|^p \\
&= \lim_{x^k \to x, \mu \downarrow 0} X \nabla_x \tilde{f}(x^k,\mu).
\end{aligned}$$

We apply the SSQP method [6] to solve (34).

SSQP Algorithm

Choose $x^0 \in R^n$, $\mu_0 > 0$ and $\sigma \in (0,1)$. Set $k=0$ and $z^0 = x^0$.

For $k \geq 0$, set

$$\begin{aligned}
x^{k+1} &= \arg\min_{x \in R^n} \tilde{f}(x^k,\mu_k) + (x-x^k)^T g^k + \tfrac{1}{2}(x-x^k)^T D_k (x-x^k), \\
\mu_{k+1} &= \begin{cases} \mu_k & \text{if } \tilde{f}(x^{k+1},\mu_k) - \tilde{f}(x^k,\mu_k) \leq -4\alpha p \mu_k^p \\ \sigma\mu_k & \text{othewise,} \end{cases} \\
z^{k+1} &= \begin{cases} x^k & \text{if } \mu_{k+1} = \sigma\mu_k \\ z^k & \text{othewise.} \end{cases}
\end{aligned}$$

Here D_k is a diagonal matrix whose diagonal elements $d_i^k, i = 1,\ldots,n$ are defined by

$$d_i^k = \begin{cases} \max\{2\|A^TA\| + 8\lambda p |\frac{x_i^k}{2}|^{p-2}, \frac{|g_i^k|}{2^{\frac{p}{2}-1}\mu^{\frac{p}{2}}|x_i^k|^{1-\frac{p}{2}}}\} & \text{if } |x_i^k| > 2\mu \\ \max\{2\|A^TA\| + 8\lambda p \mu^{p-2}, \frac{|g_i^k|}{\mu}\} & \text{if } |x_i^k| \leq 2\mu, \end{cases}$$

where $g^k = \nabla_x \tilde{f}(x^k,\mu_k)$. The solution of the QP problem can be given as $x_i^{k+1} = x_i^k - g_i^k/d_i^k, i = 1,\ldots,n$.

It is shown in [6] that for any $\epsilon > 0$, the SSQP algorithm with any starting point z^0 takes at most $O(\epsilon^{-2})$ steps to find a z^k such that $\|G(z^k)\|_\infty \leq \epsilon$.

Example 4 We use Example 3.2.1: Prostate cancer in [55] to show numerical performance of the SSQP algorithm. The data for this example consists of the medical records of 97 men who were about to receive a radical prostatectomy, which is divided into a training set with 67 observations and a test set with 30 observations. The variables are eight clinical measures: lcavol, lweight, age, lbph, svi, lcp, pleason and pgg45. The aim is to find few main factors with small prediction error.

We use the training set to build model (34) with matrix $A \in R^{67\times 8}$, $b \in R^{67}$ and $p = 0.5$. The test set is used to compute the prediction error and judge the performance of the selected methods. Table 1 reports numerical

Table 1 Example 4: Variable selection by SSQP, Lasso, Best subset methods

	SSQP			Lasso	Best subset
λ	7.734	7.78	22.1		
x_1^*(lcavol)	0.6302	0.6437	0.7004	0.533	0.740
x_2^*(lweight)	0.2418	0.2765	0.1565	0.169	0.316
x_3^*(age)	0	0	0	0	0
x_4^*(lbph)	0.0755	0	0	0.002	0
x_5^*(svi)	0.1674	0.1327	0	0.094	0
x_6^*(lcp)	0	0	0	0	0
x_7^*(pleason)	0	0	0	0	0
x_8^*(pgg45)	0	0	0	0	0
Number of nonzero	4	3	2	4	2
Error	0.4294	0.4262	0.488	0.479	0.492

results of the SSQP algorithm with algorithm parameters $\sigma = 0.99$, $\mu_0 = 10$, $x^0 = (0, \ldots, 0)^T \in R^8$ and stop criterion $\|G(z^k)\|_\infty \leq 10^{-4}$ and $\mu_k \leq 10^{-4}$. Moreover, results of the best two methods (Lasso, Best subset) from Table 3.3 in [55] are also listed in Table 1. We use Figure 1 and Figure 2 to show the convergence of $\{z^k\}$,$\tilde{f}(z^k, \mu_k)$, $f(z^k)$, μ_k and $\|G(z^k)\|_\infty$ generated by the SSQP algorithm for (34) with $\lambda = 7.734$ and $\lambda = 7.78$.

Theoretical and numerical results show that smoothing methods are promising for nonsmooth, nonconvex, non-Lipschitz optimization problems.

Acknowledgements. We would like to thank Prof. Masao Fukushima for reading this paper carefully and giving many helpful comments. Thanks to Dr. Wei Bian for helpful discussions concerning analysis of smoothing functions and numerical experience.

References

1. G. Alefeld and X. Chen, A regularized projection method for complementarity problems with non-Lipschitzian functions, Math. Comput., 77(2008) 379-395.
2. A. Auslender, How to deal with the unbounded in optimization: theory and algorithm, Math. Program., 79 (1997) 3-18.
3. A. Ben-Tal, L. EI Ghaoui and A. Nemirovski, Robust Optimization, Princeton University Press, Princeton, NJ, 2009.
4. W. Bian and X. Chen, Smoothing neural network for constrained non-Lipschitz optimization with applications, IEEE Trans. Neural Netw., 23(2012) 399-411.
5. W. Bian and X. Chen, Neural network for nonsmooth, nonconvex constrained minimization via smooth approximation, Preprint, 2011.
6. W. Bian and X. Chen, Smoothing SQP algorithm for non-Lipschitz optimization with complexity analysis, Preprint, 2012.
7. W. Bian and X.P. Xue, Subgradient-based neural network for nonsmooth nonconvex optimization problem, IEEE Trans. Neural Netw., 20(2009) 1024-1038.
8. A. M. Bruckstein, D. L. Donoho and M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Review, 51(2009) 34-81.
9. J.V. Burke, Descent methods for composite nondifferentiable optimization problems, Math. Program., 33(1985) 260-279.

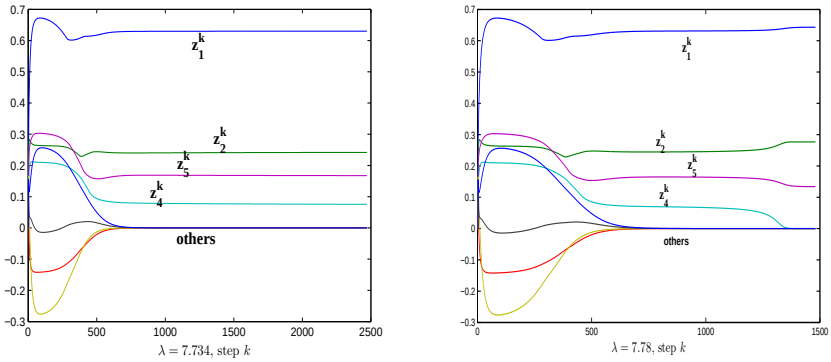

Fig. 1 Convergence of $\{z^k\}$ generated by the SSQP for (34).

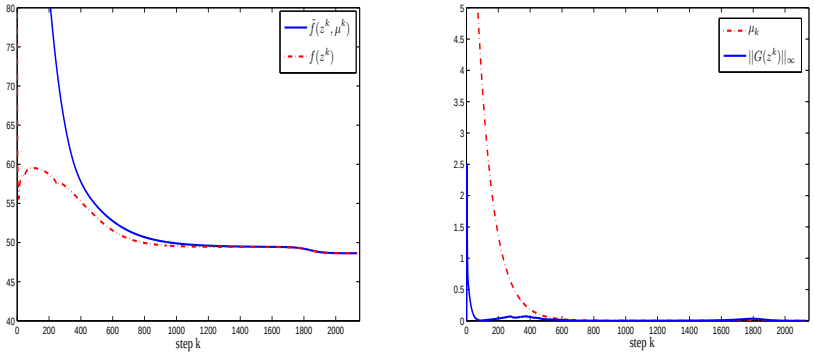

Fig. 2 Convergence of $\tilde{f}(z^k, \mu_k)$, $f(z^k)$, μ_k and $\|G(z^k)\|_\infty$ with $\lambda = 7.78$

10. J.V. Burke, A.S. Lewis and M.L. Overton, A robust gradient sampling algorithm for nonsmooth, nonconvex optimization, SIAM J. Optim., 15(2005) 751-779.
11. J.V. Burke, D. Henrion, A.S. Lewis and M.L. Overton, Stabilization via nonsmooth, nonconvex optimization, IEEE Trans. Automat. Control, 51(2006) 1760-769.
12. C. Catis, N. I. M. Gould and P. Toint, On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming, SIAM J. Optim., 21(2011) 1721-1739.
13. R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, IEEE Signal Proc. Lett., 14(2007) 707-710.
14. B. Chen and X. Chen, A global and local superlinear continuation-smoothing method for P_0 and R_0 NCP or monotone NCP, SIAM J. Optim., 9(1999) 624-645.
15. B. Chen and P.T. Harker, A non-interior-point continuation method for linear complementarity problems, SIAM J. Matrix Anal. Appl., 14(1993) 1168-1190.
16. B. Chen and N. Xiu, A global linear and local quadratic noninterior continuation method for nonlinear complementarity problems based on Chen-Mangasarian smoothing functions, SIAM J. Optim., 9(1999) 605-623.
17. C. Chen and O.L. Mangasarian, A class of smoothing functions for nonlinear and mixed complementarity problems, Math. Program., 71(1995) 51-70.
18. X. Chen, Smoothing methods for complementarity problems and their applications: a survey, J. Oper. Res. Society Japan, 43 (2000) 32-46.
19. X. Chen, A superlinearly and globally convergent method for reaction and diffusion problems with a non-Lipschitzian operator, Computing[Suppl], 15(2001) 79-90.

20. X. Chen, First order conditions for nonsmooth discretized constrained optimal control problems, SIAM J. Control Optim., 42(2004) 2004-2015.
21. X. Chen and M. Fukushima, Expected residual minimization method for stochastic linear complementarity problems, Math. Oper. Res., 30(2005) 1022-1038.
22. X. Chen and M. Fukushima, A smoothing method for a mathematical program with P-matrix linear complementarity constraints, Comp. Optim. Appl., 27(2004) 223-246.
23. X. Chen, D. Ge, Z. Wang and Y.Ye, Complexity of unconstrained L_2-L_p minimization, Preprint, 2011.
24. X. Chen, Z. Nashed and L. Qi, Smoothing methods and semismooth methods for nondifferentiable operator equations, SIAM J. Numer. Anal., 38(2000) 1200-1216.
25. X. Chen, L. Niu and Y. Yuan, Optimality conditions and smoothing trust region Newton method for non-Lipschitz optimization, Preprint, 2012.
26. X. Chen and L. Qi, A parameterized Newton method and a quasi-Newton method for nonsmooth equations. Comput. Optim. Appl., 3(1994) 157-179.
27. X. Chen, L. Qi and D. Sun, Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities, Math. Comput., 67(1998) 519-540.
28. X. Chen, R. J-B Wets and Y. Zhang, Stochastic variational inequalities: residual minimization smoothing/sample average approximations, to appear in SIAM J. Optim.
29. X. Chen, R. Womersley and J. Ye, Minimizing the condition number of a Gram matrix, SIAM J. Optim., 21(2011) 127-148.
30. X. Chen, F. Xu and Y. Ye, Lower bound theory of nonzero entries in solutions of l_2-l_p minimization, SIAM J. Sci. Comput., 32(2010) 2832-2852.
31. X. Chen and Y. Ye, On homotopy-smoothing methods for variational inequalities, SIAM J. Control Optim., 37(1999) 587-616.
32. X. Chen, C. Zhang and M. Fukushima, Robust solution of monotone stochastic linear complementarity problems, Math. Program., 117(2009) 51-80.
33. X. Chen and W. Zhou, Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization, SIAM J. Imag. Sci., 3(2010) 765-790.
34. F.H. Clarke, Optimization and Nonsmooth Analysis, John Wiley, New York, 1983.
35. A.R. Conn, K. Scheinberg and L. N. Vicente, Introduction to Derivative-Free Optimization, MPS-SIAM Book Series on Optimization, SIAM, Philadelphia, 2009.
36. R.W. Cottle, J.S. Pang and R.E. Stone, The Linear Complementarity Problem, Academic Press, Inc., Boston, 1992.
37. A. Daniilidis, C. Sagastizbal and M. Solodov, Identifying structure of nonsmooth convex functions by the bundle technique, SIAM J. Optim., 20(2009) 820-840.
38. E. Delage and Y. Ye, Distributionally robust optimization under monment uncertainty with application to data-driven problems, Oper. Res., 58(2010) 595-612.
39. F. Facchinei, H. Jiang and L. Qi, A smoothing method for mathematical programs with equilibrium constraints, Math. Program., 85(1999) 107-134.
40. F. Facchinei and J.S. Pang, Finite-Dimensional Variational Inequalities and Complementarity Problems, Springer-Verlag, New York, 2003.
41. J. Fan, Comments on Wavelets in stastics: a review by A. Antoniadis, Stat. Method. Appl., 6(1997) 131V138.
42. J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Amer. Statist. Assoc., 96(2001) 1348-1360.
43. H. Fang, X. Chen and M. Fukushima, Stochastic R_0 matrix linear complementarity problems, SIAM J. Optim., 18(2007) 482-506.
44. M. Ferris, Extended mathematical programming: competition and stochasticity, SIAM News, 45(2012) 1-2.
45. A. Fischer: A special Newton-type optimization method, Optim., 24(1992) 269- 284.
46. M. Fukushima, Z.-Q. Luo and J.-S. Pang, A globally convergent sequential quadratic programming algorithm for mathematical programs with linear complementarity constraints, Comp. Optim. Appl., 10(1998) 5-34.
47. M. Fukushima, Z.-Q. Luo and P. Tseng, Smoothing functions for second-order-cone complementarity problems, SIAM J. Optim., 12(2002) 436-460.

48. M. Fukushima and J.-S. Pang, Convergence of a smoothing continuation method for mathematical programs with complementarity constraints, Lecture Notes in Economics and Mathematical Systems, Vol. 477, M. Thera and R. Tichatschke (eds.), Springer-Verlag, Berlin/Heidelberg, (1999) 99-110.
49. S.A. Gabriel and J.J. More, Smoothing of mixed complementarity problems. In: M.C. Ferris and J.S. Pang (eds.): Complementarity and Variational Problems: State of the Art, SIAM Philadelphia, Pennsylvania, (1997) 105-116.
50. R. Garmanjani and L.N. Vicente, Smoothing and worst case complexity for direct-search methods in non-smooth optimization, Preprint, 2011.
51. D. Ge, X. Jiang and Y. Ye, A note on the complexity of L_p minimization, Math. Program., 21(2011) 1721-1739.
52. D. Geman and G. Reynolds, Constrained restoration and the recovery of discontinuities, IEEE Trans. Pattern Anal. Mach. Intell., 14(1992) 357-383.
53. G. Gürkan, A.Y. Özge and S.M. Robinson, Sample-path solution of stochastic variational inequalities, Math. Program., 84(1999) 313-333.
54. K. Hamatani and M. Fukushima, Pricing American options with uncertain volatility through stochastic linear complementarity models, Comp. Optim. Appl., 50(2011) 263-286.
55. T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning Data Mining, Inference, and Prediction, Springer-Verlag, New York, 2009.
56. S. Hayashi, N. Yamashita and M. Fukushima, A combined smoothing and regularization method for monotone second-order cone complementarity problems, SIAM J. Optim., 15(2005) 593-615.
57. M. Heinkenschloss, C.T. Kelley and H.T. Tran, Fast algorithms for nonsmooth compact fixed-point problems, SIAM J. Numer. Anal., 29(1992) 1769-1792.
58. M. Hintermueller and T. Wu, Nonconvex TV^q-models in image restoration: analysis and a trust-region regularization based superlinearly convergent solver, Preprint, 2011.
59. J.B. Hiriart-Urruty and C. Lemareéchal, Convex Analysis and Minimization Algorithms II: Advanced Theory and Boundle Methods, Springer-Verlag, Berlin, 1993.
60. M. Hu and M. Fukushima, Smoothing approach to Nash equilibrium formulations for a class of equilibrium problems with shared complementarity constraints, to appear in Comp. Optim. Appl.,
61. J. Huang, J. L. Horowitz and S. Ma, Asymptotic properties of bridge estimators in sparse high-dimensional regression models, Ann. Stat., 36(2008) 587-613.
62. J. Huang, S. Ma, H. Xie and C.-H. Zhang, A group bridge approach for variable selection, Biometrika, 96(2009) 339-355.
63. Z. Huang, L. Qi and D. Sun, Sub-quadratic convergence of a smoothing Newton algorithm for the P_0- and monotone LCP, Math. Program., 99(2004) 423-441.
64. H. Jiang and D. Ralph, Smooth SQP methods for mathematical programs with nonlinear complementarity constraints, Comp. Optim. Appl., 25(2002) 123-150.
65. H. Jiang and H. Xu, Stochastic approximation approaches to the stochastic variational inequality problem, IEEE. Trans. Autom. Control, 53(2008) 1462-1475.
66. C. Kanzow, Some noninterior continuation methods for linear cornplementarity problems, SIAM J. Matrix Anal. Appl., 17(1996) 851-868.
67. K. Knight and W.J. Fu, Asymptotics for lasso-type estimators, Ann. Stat., 28(2000) 1356-1378.
68. K.C. Kiwiel, Methods of Descent for Nondifferentiable Optimization, Lecture Notes in Math. 1133, Springer-Verlag, Berlin, New York, 1985.
69. K.C. Kiwiel, Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization, SIAM J. Optim., 18(2007) 379-388.
70. A.S. Lewis and C.H.J. Pang, Lipschitz behavior of the robust regularization, SIAM, J. Control Optim., 48(2010) 3080-3104.
71. A.S. Lewis and M.L. Overton, Eigenvalue optimization, Acta Numerica, 5(1996) 149-190.
72. D. Li and M. Fukushima, Smoothing Newton and quasi-Newton methods for mixed complementarity problems, Comp. Optim. Appl., 17(2000) 203-230.
73. G.H. Lin, X. Chen and M. Fukushima, Solving stochastic mathematical programs with equilibrium constraints via approximation and smoothing implicit programming with penalization, Math. Program., 116(2009) 343-368.

74. G.H. Lin and M. Fukushima, Stochastic equilibrium problems and stochastic mathematical programs with equilibrium constraints: A survey, Pacific J. Optim., 6(2010) 455-482.
75. Z-Q. Luo, J-S.Pang, D. Ralph and S. Wu, Exact penalization and stationarity conditions of mathematical programs with equilibrium constraints, Math. Program., 75(1996) 19-76.
76. P. Maréchal and J.J. Ye, Optimizing condition numbers, SIAM J. Optim., 20(2009) 935-947.
77. J.M. Martínez and A.C. Moretti, A trust region method for minimization of nonsmooth functions with linear constraints, Math. Program., 76(1997) 431-449.
78. K. Meng and X. Yang, Optimality conditions via exact penalty functions, SIAM J. Optim., 20(2010) 3205-3231.
79. R. Mifflin, Semismooth and semiconvex functions in constrained optimization, SIAM J. Control Optim., 15(1977) 957-972.
80. B.S. Mordukhovich, Variational Analysis and Generalized Differentiation I and II, Springer, Berlin, 2006.
81. Y. Nesterov, Smooth minimization of nonsmooth functions, Math. Program., 103(2005) 127-152.
82. Y. Nesterov, Smoothing technique and its applications in semidefinite optimization, Math. Program., 110(2007) 245-259.
83. W.K. Newey and D. McFadden, Large sample estimation and hypothesis testing, Handbook of econometrics, Vol.IV, North-Holland, Amsterdam, (1994) 2111-2245.
84. M. Nikolova, Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares, SIAM J. Multiscale Model. Simul., 4(2005) 960-991.
85. M. Nikolova, M. K. Ng, S. Zhang and W. Ching, Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization, SIAM J. Imag. Sci., 1(2008) 2-25.
86. J. Nocedal and S.J. Wright, Numerical Optimization, 2nd Edition, Springer, New York, 2006.
87. M.R. Osborne, Finite Algorithms in Optimizations in Optimization and Data Analysis, John Wiley, Chichester, UK, New York, 1985.
88. R.A. Polyak, A Nonlinear rescaling vs. smoothing technique in convex optimization, Math. Program., 92(2002) 197-235.
89. H.D. Qi and L.Z. Liao, A smoothing Newton method for extended vertical linear complementarity problems, SIAM J. Matrix Anal. Appl., 21(1999) 45-66.
90. L. Qi and X. Chen, A globally convergent successive approximation method for severely nonsmooth equations, SIAM J. Control Optim., 33(1995) 402-418.
91. L. Qi, C. Ling, X. Tong and G. Zhou, A smoothing projected Newton-type algorithm for semi-infinite programming, Comput. Optim. Appl., 42(2009) 1-30.
92. L. Qi, D. Sun and G. Zhou, A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities, Math. Program., 87(2000) 1-35.
93. R.T. Rockafellar, A property of piecewise smooth functions, Comput. Optim. Appl., 25(2003) 247-250.
94. R.T. Rockafellar and R.J-B Wets, Variational Analysis, Springer-Verlag, New York, 1998.
95. A. Ruszczynski and A. Shapiro, Stochastic Programming, Handbooks in Operations Research and Management Science, Elsevier, 2003.
96. S. Smale, Algorithms for solving equations. Proceedings of the International Congress of Mathematicians, Berkeley, California, (1986) 172-195.
97. J. Sun, D. Sun and L. Qi, A squared smoothing Newton method for nonsmooth matrix equations and its applications in semidefinite optimization problems, SIAM J. Optim., 14(2004) 783-806.
98. Y. Tassa and E. Todorov, Stochastic complementarity for local control of discontinuous dynamics, in Proceedings of Robotics: Science and Systems (RSS), 2010.
99. S. J. Wright, Convergence of an inexact algorithm for composite nonsmooth optimization, IMA J. Numer. Anal., 9(1990) 299-321.

100. H. Xu, Sample average approximation methods for a class of stochastic variational inequality problems, Asia Pac. J. Oper. Res., 27(2010) 103-119.
101. Z. Xu, H. Zhang, Y. Wang and X. Chang, $L_{1/2}$ regularizer, Science in China Series F-Inf Sci., 53(2010) 1159-1169.
102. Y. Yuan, Conditions for convergence of a trust-region method for nonsmooth optimization, Math. Program., 31(1985) 220-228.
103. I. Zang: A smoothing-out technique for min-max optimization, Math. Program., 19(1980) 61-71.
104. C. Zhang and X. Chen, Smoothing projected gradient method and its application to stochastic linear complementarity problems, SIAM J. Optim., 20(2009) 627-649.
105. C. Zhang, X. Chen and A. Sumalee, Wardrop's user equilibrium assignment under stochastic environment, Transport. Res. B, 45(2011) 534-552.
106. C.-H. Zhang, Nearly unbiased variable selection under minimax concave penalty, Ann. Statist. 38(2010) 894-942.
107. G. L. Zhou, L. Caccetta and K.L Teo, A superlinearly convergent method for a class of complementarity problems with non-Lipschitzian functions, SIAM J. Optim., 20(2010) 1811-1827.