# Neural Network for Nonsmooth, Nonconvex Constrained Minimization via Smooth Approximation

Wei Bian and Xiaojun Chen

***Abstract*—A neural network based on smoothing approximation is presented for a class of nonsmooth, nonconvex constrained optimization problems, where the objective function is nonsmooth and nonconvex, the equality constraint functions are linear and the inequality constraint functions are nonsmooth, convex. This approach can find a Clarke stationary point of the optimization problem by following a continuous path defined by a solution of an ordinary differential equation. The global convergence is guaranteed if either the feasible set is bounded or the objective function is level-bounded. Specially, the proposed network does not require (i) the initial point to be feasible; (ii) a prior penalty parameter to be chosen exactly; (iii) a differential inclusion to be solved. Numerical experiments and comparisons with some existing algorithms are presented to illustrate the theoretical results and show the efficiency of the proposed network.**

***Index Terms*—Nonsmooth nonconvex optimization, neural network, smoothing approximation, Clarke stationary point, variable selection, condition number.**

## I. Introduction

The approach based on the use of analog neural networks for solving nonlinear programming problems and their engineering applications have received a great deal of attention in the last two decades. See [1]–[12], etc., and references therein. The neural network method is effective and particularly attractive in the applications where it is of crucial importance to obtain the optimal solutions in real time, as in some robotic control, signal processing and compressed sensing. Artificial neural networks can be used to model the dynamics of a system [13] and implemented physically by designed hardware such as specific integrated circuits where the computational procedure is distributed and parallel. Some dynamical properties of differential equation or differential inclusion networks make remarkable contributions to their applications in optimization [14]–[17].

In this paper, we consider the following constrained nonsmooth nonconvex minimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & Ax = b, \qquad g(x) \le 0, \end{array} \tag{1}$$

where $x \in \mathbb{R}^n$, $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, but not necessarily differentiable or convex, $A \in \mathbb{R}^{r\times n}$ is of full row rank, $b \in \mathbb{R}^r$, $g : \mathbb{R}^n \to \mathbb{R}^m$ and $g_i$ is convex but not necessarily differentiable, $i = 1, 2, \cdots, m$.

Nonsmooth and nonconvex optimization problem arises in a variety of scientific and engineering applications. For example, the constrained nonsmooth nonconvex optimization model

$$\min_{x\in\mathcal{C}} \|Ax-b\|_2^2 + \lambda \sum_{i=1}^{r} \varphi(|d_i^T x|), \tag{2}$$

where $\lambda > 0$, $r$ is a positive integer, $d_i \in \mathbb{R}^n$, $\mathcal{C}$ is a closed convex subset of $\mathbb{R}^n$ and $\varphi$ is a given penalized function. Problem (2) attracts great attention in variable selection and sparse reconstruction [18]–[22]. Moreover, the problem of minimizing condition number is also an important class of nonsmooth nonconvex optimization problems, which has been widely used in the sensitivity analysis of interpolation and approximations [23].

Recently, some discrete iterative algorithm, statistical algorithms and dynamic subgradient algorithms are proposed for constrained nonsmooth nonconvex optimization problems. Among them, the smoothing projected gradient method [24] is a discrete iterative method, which uses smoothing approximations and has global convergence. The sequence quadratic programming algorithm based on gradient sampling (SQP-GS) [25] is a statistical method, which uses a process of gradient sampling around each iterate $x^k$, and have global convergence to find a Clarke stationary point "with probability one". The network in [7] uses exact penalty functions to find a Clarke stationary point via a differential inclusion. To avoid estimating an upper bound of the Lipschitz constant of the inequality constrained functions over a compact set needed in [7], Liu and Wang [9] propose another network to solve nonconvex optimization problem (1). A neural network via smoothing techniques is proposed in [12] for solving a class of non-Lipschitz optimization, where the objective function is non-Lipschitz with specific structure and the constraint is so simple such that its projection has a closed form. Moreover, the network in [12] is to find a scaled stationary point of the considered problem, which may not be a Clarke stationary point of Lipschitz optimization. Although these methods can efficiently solve some nonsmooth, nonconvex optimization problems, some difficulties still remain. For instance, the statistical gradient sampling methods relay on the number of the individuals largely and require that the functions are

Wei Bian is with the Department of Mathematics, Harbin Institute of Technology, Harbin, China (bianweilvse520@163.com). The author's work was supported by the Hong Kong Polytechnic University Postdoctoral Fellowship Scheme and the NSF foundation (11101107,11126218,11271099) of China.

Xiaojun Chen is with the Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (maxjchen@polyu.edu.hk). The author's work was supported in part by Hong Kong Research Grant Council grant (PolyU5003/10P).

differentiable at all iterates for global convergence analysis; the algorithms based on projection methods have difficulties in handling complex constraints; the dynamic subgradient methods need exact penalty parameters and solutions of differential inclusions.

The main contributions of this paper are as follows. First, the proposed network can solve the nonconvex optimization problem with general convex constraints without needing to give the exact penalty parameter in advance. To find an exact penalty parameter, most existing results need the Lipschitz constants of the objective and constraint functions and the boundedness of the feasible region [4], [7], [9]. However, estimating these values is very difficult in most cases. Moreover, too large penalty parameter may bring numerical overflow in calculation and let the network ill-conditioned. To overcome these difficulties, smoothing method is introduced into the network, which leads the differentiability of the approximated objective and penalty functions. Then the penalty parameter can be updated on line following some values, such as the gradient information of the approximated functions and the smoothing parameter. Second, by the smoothing methods, the proposed network is modeled by a differential equation not differential inclusion and can be implemented directly by circuits and mathematical softwares. For the networks modeled by a differential inclusion, one needs to know the element in the right hand set-valued map which equals to $\dot{u}(t)$ almost everywhere. This is crucial for the implementation of networks and relays on the geometry property of the set-valued map. Third, the smoothing parameter in the proposed network is updated continuously, which is different from the updating rules in the previous iterative algorithms. Fourth, the proposed network does not need large sampling for approximation, which is used in the statistical optimization methods.

This paper is organized as follows. In Section II, we define a class of smoothing functions and give some properties of smoothing functions for the composition of two functions. In Section III, the proposed neural network via smoothing techniques is present. In Section IV, we study the existence and limit behavior of solutions of the proposed network. In Section V, some numerical results and comparisons show that the proposed network is promising and performs well.

Let $\|\cdot\|$ denote the 2-norm of a vector and a matrix. For a subset $U \subseteq \mathbb{R}^n$, let int$(U)$, bd$(U)$ and $U^C$ denote the interior, boundary and complementary sets of $U$, respectively.

## II. Smoothing Approximation

Many smoothing approximations for nonsmooth optimization problems have been developed in the past decades [26]–[30]. The main feature of smoothing methods is to approximate the nonsmooth functions by parameterized smooth functions.

*Definition 2.1:* Let $h : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz. We call $\tilde{h} : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}$ a smoothing function of $h$, if $\tilde{h}$ satisfies the following conditions.

(i) For any fixed $\mu \in (0,\infty)$, $\tilde{h}(\cdot,\mu)$ is continuously differentiable in $\mathbb{R}^n$, and for any fixed $x \in R^n$, $\tilde{h}(x,\cdot)$ is differentiable in $(0,\infty)$.

(ii) For any fixed $x \in \mathbb{R}^n$, $\lim_{\mu\downarrow 0} \tilde{h}(x,\mu) = h(x)$.

(iii) $\{\lim_{z\to x,\mu\downarrow 0} \nabla_z \tilde{h}(z,\mu)\} \subseteq \partial h(x)$.

(iv) There is a positive constant $\kappa_{\tilde{h}} > 0$ such that $|\nabla_\mu \tilde{h}(x,\mu)| \le \kappa_{\tilde{h}}$, $\forall \mu \in (0,\infty)$, $x \in \mathbb{R}^n$.

From (iv) of Definition 2.1, for any $\mu \ge \bar{\mu} > 0$, we have

$$|\tilde{h}(x,\mu) - \tilde{h}(x,\bar{\mu})| \le \kappa_{\tilde{h}}(\mu - \bar{\mu}), \quad \forall x \in \mathbb{R}^n,$$

letting $\bar{\mu} \downarrow 0$ in the above inequality and from (ii) of Definition 2.1, it gives

$$|\tilde{h}(x,\mu) - h(x)| \le \kappa_{\tilde{h}}\mu, \quad \forall \mu \in (0,\infty), \quad x \in \mathbb{R}^n. \quad (3)$$

For any fixed $z, x \in \mathbb{R}^n$, from (3), we obtain

$$\begin{aligned} |\tilde{h}(z,\mu) - h(x)| \le& |\tilde{h}(z,\mu) - h(z)| + |h(z) - h(x)| \\ \le& \kappa_{\tilde{h}}\mu + |h(z) - h(x)|, \end{aligned}$$

which implies

$$\lim_{z\to x,\mu\downarrow 0} \tilde{h}(z,\mu) = h(x). \quad (4)$$

The following proposition gives four important properties for the compositions of smoothing functions. The proof of Proposition 2.1 can be found in Appendix.

*Proposition 2.1:* (a) Let $\tilde{f}_1, \ldots, \tilde{f}_m$ be smoothing functions of $f_1, \ldots, f_m$, then $\sum_{i=1}^m \alpha_i \tilde{f}_i$ is a smoothing function of $\sum_{i=1}^m \alpha_i f_i$ with $\kappa_{\sum_{i=1}^m \alpha_i \tilde{f}_i} = \sum_{i=1}^m \alpha_i \kappa_{\tilde{f}_i}$ when $\alpha_i \ge 0$ and $f_i$ is regular [31] for any $i = 1, 2, \ldots, m$.

(b) Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and $\psi : \mathbb{R} \to \mathbb{R}$ be continuously differentiable and globally Lipschitz with a Lipschitz constant $l_\psi$. If $\tilde{\varphi}$ is a smoothing function of $\varphi$, then $\psi(\tilde{\varphi})$ is a smoothing function of $\psi(\varphi)$ with $\kappa_{\psi(\tilde{\varphi})} = l_\psi \kappa_{\tilde{\varphi}}$.

(c) Let $\varphi : \mathbb{R}^m \to \mathbb{R}$ be regular and $\psi : \mathbb{R}^n \to \mathbb{R}^m$ be continuously differentiable. If $\tilde{\varphi}$ is a smoothing function of $\varphi$, then $\tilde{\varphi}(\psi)$ is a smoothing function of $\varphi(\psi)$ with $\kappa_{\tilde{\varphi}(\psi)} = \kappa_{\tilde{\varphi}}$.

(d) Let $\varphi : \mathbb{R}^n \to \mathbb{R}$ be locally Lipschitz and $\psi : \mathbb{R} \to \mathbb{R}$ be globally Lipschitz with a Lipschitz constant $l_\psi$. If $\tilde{\varphi}$ and $\tilde{\psi}$ are smoothing functions of $\varphi$ and $\psi$, $\tilde{\psi}(\cdot,\mu)$ and $\tilde{\varphi}(\cdot,\mu)$ are convex, and $\tilde{\psi}(\cdot,\mu)$ is non-decreasing, then $\tilde{\psi}(\tilde{\varphi})$ is a smoothing function of $\psi(\varphi)$ with $\kappa_{\tilde{\psi}(\tilde{\varphi})} = \kappa_{\tilde{\psi}} + l_\psi \kappa_{\tilde{\varphi}}$.

*Example 2.1:* Four popular smoothing functions of $\phi(s) = \max\{0, s\}$ are

$$\tilde{\phi}_1(s,\mu) = s + \mu \ln(1 + e^{-\frac{s}{\mu}}), \qquad \tilde{\phi}_2(s,\mu) = \frac{1}{2}(s + \sqrt{s^2 + 4\mu^2}),$$

$$\tilde{\phi}_3(s,\mu) = \begin{cases} \max\{0,s\} & \text{if } |s| > \mu \\ \dfrac{(s+\mu)^2}{4\mu} & \text{if } |s| \le \mu, \end{cases}$$

$$\tilde{\phi}_4(s,\mu) = \begin{cases} s + \dfrac{\mu}{2} e^{-\frac{s}{\mu}} & \text{if } s > 0 \\ \dfrac{\mu}{2} e^{-\frac{s}{\mu}} & \text{if } s \le 0. \end{cases}$$

It is easy to find that the four functions satisfy the four conditions in Definition 2.1 with $\kappa_{\tilde{\phi}_1} = \ln 2$, $\kappa_{\tilde{\phi}_2} = 1$, $\kappa_{\tilde{\phi}_3} = 1/4$ and $\kappa_{\tilde{\phi}_4} = 1$. For $i = 1, 2, 3, 4$, $\tilde{\phi}_i(s,\mu)$ is convex and non-decreasing for any fixed $\mu > 0$, and non-decreasing for any fixed $s \in \mathbb{R}$. Moreover, we note that the four smoothing functions have a common property that

$$\nabla_s \tilde{\phi}_i(s,\mu) \geq \frac{1}{2}, \qquad \forall s \in [0,\infty), \quad \mu \in (0,+\infty). \tag{5}$$

Since $|s| = \max\{0,s\} + \max\{0,-s\}$, then we can also obtain some smoothing functions of $|s|$ by the above smoothing functions of $\max\{0,s\}$, where one frequently used is

$$\tilde{\theta}(s,\mu) = \begin{cases} |s| & \text{if } |s| > \dfrac{\mu}{2} \\ \dfrac{s^2}{\mu} + \dfrac{\mu}{4} & \text{if } |s| \leq \dfrac{\mu}{2}. \end{cases} \tag{6}$$

Note that $\tilde{\theta}(\cdot,\mu)$ is convex for any fixed $\mu > 0$, $\tilde{\theta}(s,\cdot)$ is non-decreasing for any fixed $s \in \mathbb{R}$ and $\kappa_{\tilde{\theta}} = 1/4$.

Among many existing smoothing methods, simple structure is one of most important factors for the neural network design. For example, $\tilde{\phi}_3(s,\mu)$ is a better choice for $\phi(s)$. High order smoothness and maintaining the features of the original nonsmooth function as much as possible are also crucial for the produced smoothing function. See [26]–[30] for other smoothing functions and relative analysis. Moreover, the scheme on updating the smoothing parameter will affect the convergence rate. How to choose a better performance smoothing function and scheme of updating smoothing parameter gives us a topic for further research.

## III. Proposed Neural Network

Denote the feasible set of (1) by $\mathbb{X} = \mathbb{X}_1 \cap \mathbb{X}_2$, where $\mathbb{X}_1 = \{\, x \mid Ax = b\}$ and $\mathbb{X}_2 = \{\, x \mid g(x) \leq 0\}$. We always assume the following conditions hold in this paper.
(A1) There is $\hat{x} \in \mathbb{X}_1 \cap \mathrm{int}(\mathbb{X}_2)$.
(A2) The feasible region $\mathbb{X}$ is bounded.

Let $c = A^T(AA^T)^{-1}b$, $P = I_n - A^T(AA^T)^{-1}A$, $q(x) = \sum_{i=1}^m \max\{0, g_i(x)\}$. In what follows, we use a smoothing function $\tilde{\phi}$ of $\max\{0,s\}$ given in Example 2.1. Since $\max_{1\leq i\leq 4}\{\kappa_{\tilde{\phi}_i}\} \leq 1$, we let $\kappa_{\tilde{\phi}} = 1$ in our following theoretical analysis.

Let $\tilde{f} : \mathbb{R}^n \times [0,\infty) \to \mathbb{R}$ be a smoothing function of $f$ and the smoothing function of $q$ be given as

$$\tilde{q}(x,\mu) = \sum_{i=1}^m \tilde{\phi}(\tilde{g}_i(x,\mu),\mu), \tag{7}$$

where $\tilde{g}_i : \mathbb{R}^n \times (0,\infty) \to \mathbb{R}$ is a smoothing function of $g_i$, $i = 1,2,\cdots,m$. Since $g_i$ is convex, $\tilde{g}_i(x,\mu_2) \geq \tilde{g}_i(x,\mu_1)$ for $\mu_2 \geq \mu_1 > 0$ in most smoothing functions, which implies

$$\tilde{g}_i(x,\mu) \geq g_i(x), \quad \forall x \in \mathbb{R}^n,\ \mu \in (0,\infty),\ i = 1,\dots,m. \tag{8}$$

Thus, we suppose $\tilde{g}_i(\cdot,\mu)$ is convex and $\tilde{g}_i(x,\cdot)$ is non-decreasing and denote

$$\kappa = \max_{1\leq i\leq m}\{\kappa_{\tilde{g}_i}\}.$$

From (c) of Proposition of 2.1, we get that $\tilde{q}$ is a smoothing function of $q$ with

$$\kappa_{\tilde{q}} = m\kappa_{\tilde{\phi}} + \sum_{i=1}^m \kappa_{\tilde{g}_i} = m + \sum_{i=1}^m \kappa_{\tilde{g}_i} \leq m(1+\kappa). \tag{9}$$

From condition (A1), we denote

$$\beta = -\frac{\max_{1\leq i\leq m} g_i(\hat{x})}{4}, \quad \mu_0 = \frac{-\max_{1\leq i\leq m} g_i(\hat{x})}{2\kappa + 4(m-1)}.$$

*Remark 3.1:* If $g_1,\dots,g_m$ are smooth, we can define $\mu_0 = 1$ for $m = 1$ and $\mu_0 = \frac{-\max_{1\leq i\leq m} g_i(\hat{x})}{4(m-1)}$, for $m > 1$.

The affine equality constraints are very difficult to handle in optimization, especially in large dimension problems. One of the most important methods is the projection method. However, when the matrix dimension $m$ is large and the structure of $A$ is not simple, it is difficult and expensive to calculate $P$. We should state that the proposed network in this paper is applicative for the problems where the matrix $P$ can be calculated effectively. Then, we consider the following unconstrained optimization problem

$$\min \quad f(Px+c) + \sigma q(Px+c), \tag{10}$$

where $\sigma > 0$ is a positive penalty parameter. It is known that if $f$ and $g$ are smooth, there is a $\hat{\sigma} > 0$ such that for all $\sigma \geq \hat{\sigma}$, if $x^* \in \mathbb{X}$ is a stationary point of (10), then $Px^* + c = x^*$ is a stationary point of (1) [32, Theorem 17.4]. However, choosing such $\hat{\sigma}$ is very difficult. To overcome these difficulties, we adopt a parametric penalty function defined as

$$\sigma(x,\mu) = \frac{(|\langle P\nabla_x \tilde{f}(x,\mu), P\nabla_x \tilde{q}(x,\mu)\rangle| + \lambda\beta\mu)\|x-\hat{x}\|^2}{\max\{\beta^2, \|P\nabla_x \tilde{q}(x,\mu)\|^2\|x-\hat{x}\|^2\}}, \tag{11}$$

where $\lambda$ is a positive parameter defined as

$$\lambda = \frac{2q(u_0) + 4m(1+\kappa)\mu_0}{\beta\mu_0}.$$

The main goal of this paper is to present a stable and continuous path $u \in C^1[0,\infty)$, which leads to the set $\mathbb{X}^*$ of the Clarke stationary points[1] of (1) from any starting point $x_0 \in \mathbb{R}^n$.

We consider the following network modeled by a class of ordinary differential equations (ODEs)

$$\begin{aligned} &\dot{u}(t) = -P(\nabla_u \tilde{f}(u(t),\nu(t)) + \sigma(u(t),\nu(t))\nabla_u \tilde{q}(u(t),\nu(t))), \\ &u(0) = Px_0 + c, \end{aligned} \tag{12}$$

where $x_0 \in \mathbb{R}^n$ and $\nu(t) = \mu_0 e^{-t}$.

In order to implement (12) by circuits, we can use the reformulated form of (12) as following

$$\begin{aligned} &\dot{u}(t) = -P(\nabla_u \tilde{f}(u(t),\nu(t)) + \sigma(u(t),\nu(t))\nabla_u \tilde{q}(u(t),\nu(t))), \\ &\dot{\nu}(t) = -\nu(t) \\ &u(0) = Px_0 + c,\ \nu(0) = \mu_0. \end{aligned} \tag{13}$$

(13) can be seen as a network with two input and two output variables. A simple block structure of the network (13) implemented by circuits is presented in Fig. 1. The blocks $P\nabla_u \tilde{f}$ and $P\nabla_u \tilde{q}$ can be realized by matrix $P$, $\nabla_u \tilde{f}(u,\nu)$ and $\nabla_u \tilde{q}(u,\nu)$ based on the adder and multiplier components.

[1] $x^*$ is called a Clarke stationary point of (1) if $x^* \in \mathbb{X}$ and there is a $\xi^* \in \partial f(x^*)$ such that

$$\langle x - x^*, \xi^*\rangle \geq 0, \qquad \forall x \in \mathbb{X}.$$

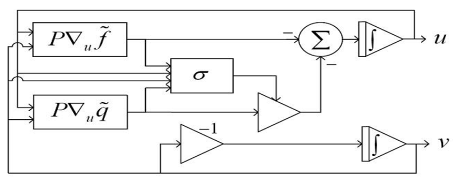


Fig. 1: Schematic block structure of a neural network described by (12)

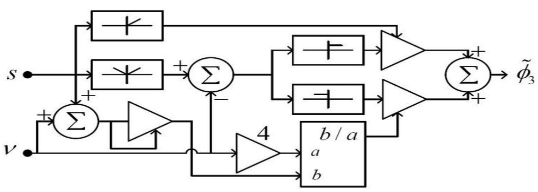


Fig. 2: Circuit implementation of term $\tilde{\phi}_3(s,\nu)$ by circuits

Fig. 2 and Fig. 3 show the implementation methods on $\phi_3(s,\nu)$ and $\nabla_s\phi_3(s,\nu)$, which give some hints on how to implement $\nabla_u\tilde{f}(u,\nu)$ and $\nabla_u\tilde{q}(u,\nu)$. $\sigma$ is a block with scalar output based on the information of $u$, $\nu$, $\nabla_u\tilde{f}(u,\nu)$ and $\nabla_u\tilde{q}(u,\nu)$. A detailed architecture flow structure of the block $\sigma$ is given in Fig. 4, where $F_i$ and $Q_i$ denote the $i$th output of the blocks $P\nabla_u\tilde{f}$ and $P\nabla_u\tilde{q}$, respectively. From Figs. 1-4, we can see that network (12) can be implemented by the adder, multiplier, divider and comparator components in circuits. Through the expression of $\sigma(u,\nu)$ looks complex, it can be realized based on the existing blocks $P\nabla_u\tilde{f}$ and $P\nabla_u\tilde{q}$, which shows that it will not bring expensive components in circuit implementation. The readers can refer to [33] for the detailed techniques on this topic.

## IV. Existence and limit behavior

In this section, we study the existence and limit behavior of the solutions of (12). For readability, we put the proof of all theoretical results in Appendix.

*Theorem 4.1:* For any $x_0 \in \mathbb{R}^n$, (12) has a solution $u \in C^1[0,\infty)$. Moreover, there is a $\rho > 0$ such that for any solution $u$ of (12) in $C^1[0,\infty)$, we have $\sup_{t\in[0,\infty)}\|u(t)\| \le \rho$.

*Remark 4.1:* We know that a finite penalty parameter is very important for implementation. From Theorem 4.1, $\sigma(u(t),\nu(t))$ is uniformly bounded on $[0,\infty)$.

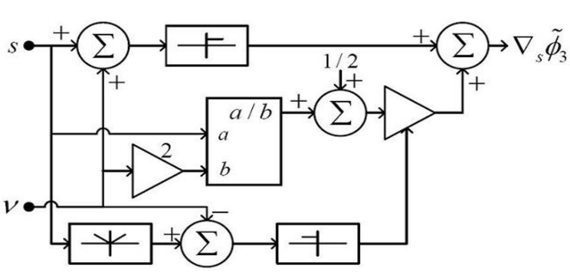


Fig. 3: Circuit implementation of term $\nabla_s\tilde{\phi}_3(s,\nu)$ by circuits

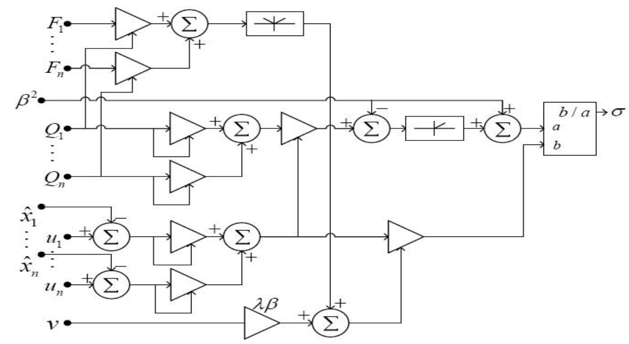


Fig. 4: Circuit implementation of term $\sigma(u,\nu)$ by circuits

Furthermore, locally Lipschitz property of the proposed smoothing functions can guarantee the uniqueness of the solution of (12).

*Proposition 4.1:* When $\nabla_x\tilde{f}(\cdot,\mu)$ and $\nabla_x\tilde{q}(\cdot,\mu)$ are locally Lipschitz for any fixed $\mu \in (0,\mu_0]$, then (12) has a unique solution.

The following theorem shows the feasibility and limit behavior of $u(t)$ as $t \to \infty$.

*Theorem 4.2:* Any solution $u(t)$ of (12) in $C^1[0,\infty)$ satisfies $\{\lim_{t\to\infty} u(t)\} \subseteq \mathbb{X}$.

Note that $q$ is convex on $\mathbb{R}^n$ and $\partial q(x)$ exists for all $x \in \mathbb{R}^n$. From [31, Corollary 1 of Proposition 2.3.3 and Theorem 2.3.9], we have the expression of $\partial q(x)$, and from [31, Corollary 1 and Cororllary 2 of Theorem 2.4.7], the normal cones to the three sets can be expressed as follows:

$$\begin{aligned} N_{\mathbb{X}_1}(x) &= \{A^T\xi \mid \xi \in \mathbb{R}^m\}, && \forall x \in \mathbb{X}_1,\\ N_{\mathbb{X}_2}(x) &= \cup_{\tau\ge 0}\tau\partial q(x), && \forall x \in \mathbb{X}_2,\\ N_{\mathbb{X}}(x) &= N_{\mathbb{X}_1}(x) + N_{\mathbb{X}_2}(x), && \forall x \in \mathbb{X}. \end{aligned}$$

*Theorem 4.3:* Any solution $u(t)$ of (12) in $C^1[0,\infty)$ satisfies

(i) $\dot{u}(t) \in L^2[0,\infty)$;
(ii) $\lim_{t\to\infty} f(u(t))$ exists and $\lim_{t\to\infty}\|\dot{u}(t)\| = 0$;
(iii) $\{\lim_{t\to\infty} u(t)\} \subseteq \mathbb{X}^*$, where $\mathbb{X}^*$ is the set of Clarke stationary points of (1).

*Remark 4.2:* If the objective function $f$ is level-bounded[1], there is $R > 0$ such that $\|x - \hat{x}\|^2 \le R$ holds for all $x \in \{x : f(x) \le f(\hat{x})\}$. By adding constraint $\|x-\hat{x}\|^2 \le R$ to the original optimization problem, the extension problem satisfies assumption (A2) and has the same optimal solutions as the original problem.

*Remark 4.3:* If $f$ is pseudoconvex on $\mathbb{X}$[2], which may be nonsmooth and nonconvex, from Theorem 4.2 and Theorem 4.3, any solution of (12) converges to the optimal solution set of (1). Some pseudoconvex functions in engineering and economic problems are given in [11], [34].

[1]We call $f$ is level-bounded, if the level set $\{x \in \mathbb{R}^n \mid f(x) \le \eta\}$ is bounded for any $\eta > 0$.

[2]We call $f$ is pseudoconvex on $\mathbb{X}$ if for any $x', x'' \in \mathbb{X}$, we have

$$\exists\zeta(x') \in \partial\phi(x') : \langle\zeta(x'), x''-x'\rangle \ge 0 \Rightarrow f(x'') \ge f(x').$$

Network (12) reduces to

$$\begin{cases} \dot{u}(t) = -\nabla_u \tilde{f}(u(t),\nu(t)) - \sigma(u(t),\nu(t))\nabla_u \tilde{q}(u(t),\nu(t)), \\ u_0 = x_0 \end{cases} \tag{14}$$

for a special cases of (1), that is

$$\min \quad f(x) \qquad \text{s.t.} \quad g(x) \le 0. \tag{15}$$

Similarly, we can obtain that the conclusions of Theorem 4.3 hold for (14) to solve (15).

Moreover, when we consider problem

$$\min \quad f(x) \qquad \text{s.t.} \quad Ax = b, \tag{16}$$

which is also a special form of (1), the feasible region $\mathbb{X}$ is unbounded. When $f$ is level-bounded, we can use the analysis in Remark 4.2 to solve it and we can obtain the results in Theorems 4.1-4.3 with the simpler network

$$\begin{cases} \dot{u}(t) = -P\nabla_u \tilde{f}(u(t),\nu(t)), \\ u_0 = Px_0 + c. \end{cases} \tag{17}$$

*Corollary 4.1:* For any $x_0 \in \mathbb{R}^n$, if $f$ is level-bounded and $\mu_0 \le 1$, the conclusions of Theorem 4.3 hold for (17) to solve (16).

*Remark 4.4:* If we can find an exact parameter $\hat{\sigma}$ such that the solutions of (10) are just the solutions of (1), then we can define

$$\sigma(x,\mu) = \hat{\sigma},$$

which brings (12) a simpler structure. All the results in this paper can be obtained by similar proofs.

*Remark 4.5:* From the proof of the above results, it is not too rigorous for the choose of $\mu_0$ and $\lambda$. Actually, all the results hold with

$$\mu_0 \le \frac{-\max_{1\le i\le m} g_i(\hat{x})}{2\kappa + 4(m-1)}, \ \lambda \ge \frac{2q(u_0)}{\mu_0} + 4m(1+\kappa).$$

## V. Numerical Examples

In this section, we use five numerical examples to illustrate the performance of network (12) and compare it with the network in [9], Lasso, Best Subset and IRL1 methods in [36], and the SQP-GS algorithm in [25]. The numerical testing was carried out on a Lenovo PC (3.00GHz, 2.00GB of RAM) with the use of Matlab 7.4. In our report for numerical results, we use the following notations.

- SNN: Use codes for ODE in Matlab to implement (12). We use `ode15s` for Examples 5.1-5.3, and `ode23` for Examples 5.4-5.5.
- $u_{t_k}$ : numerical solution of SNN at the $k$th iteration.
- $\bar{x}$: numerical solution obtained by the corresponding algorithms.
- time: CPU time in second.
- fea-err($x$): value of the infeasibility measure at $x$, which is evaluated by fea-err($x$)=$\|Ax-b\| + \sum_{i=1}^n \max\{0, g_i(x)\}$.
- $\tilde{\theta}(s,\mu)$: a smoothing function of $\theta(s) = |s|$ given in (6).

We choose $\nu(t) = \mu_0 e^{-t}$ in Examples 5.1-5.3 and $\nu(t) = \mu_0 e^{-\alpha t}$ in Examples 5.4-5.5. It is trivial to get all results in Sections IV for $\nu(t) = \mu_0 e^{-\alpha t}$ by resetting $t = \alpha t$.

*Example 5.1:* [35] Find the minimizer of a nonsmooth Rosenbrock function with constraints:

$$\begin{array}{ll} \min & 8|x_1^2 - x_2| + (x_1-1)^2 \\ \text{s.t.} & x_1 - \sqrt{2}x_2 = 0, \ x_1^2 + |x_2| - 4 \le 0. \end{array} \tag{18}$$

$x^* = (\frac{\sqrt{2}}{2}, \frac{1}{2})^T$ is the unique optimal solution of (18) and the objective function is nonsmooth at $x^*$ with the optimal value $f(x^*) = \frac{3-2\sqrt{2}}{2}$.

It is easy to see $\hat{x} = (0,0)^T \in \mathbb{X}_1 \cap \text{int}(\mathbb{X}_2)$. Let the smoothing functions of $f$ and $q$ be

$$\tilde{f}(x,\mu) = 8\tilde{\theta}(x_1^2 - x_2, \mu) + (1-x_1)^2,$$
$$\tilde{q}(x,\mu) = \tilde{\phi}_2(x_1^2 + \tilde{\theta}(x_2,\mu) - 4, \mu),$$

where $\tilde{\phi}_2$ is defined in Example 2.1. In [35], Gurbuzbalaban and Overton state that it is an interesting topic that whether the solution obtained by their proposed algorithm is the global minimizer, but not the other Clarke stationary points. Besides $x^*$, (18) has another Clarke stationary point $(0,0)^T$. We test the SNN with the 491 different initial points in $[-5,5]\times[-5,5]$, where 441 initial points are $x_0 = (-5 + 0.5i, -5+0.5j)^T$, $i,j = 0,1,\ldots,20$ and the other 50 initial points are also in $[-5,5]\times[-5,5]$ and uniformly distributed on the vertical centerline of $x^*$ and $(0,0)^T$. Through this numerical testing, we suggest for this example that

$$\text{if} \quad \|u_0 - x^*\| \le \|x_0\|, \text{ then} \lim_{t\to\infty} u(t) = x^*,$$
$$\text{otherwise} \quad \lim_{t\to\infty} u(t) = (0,0)^T.$$

However, we can not obtain this result by a theoretical proof.

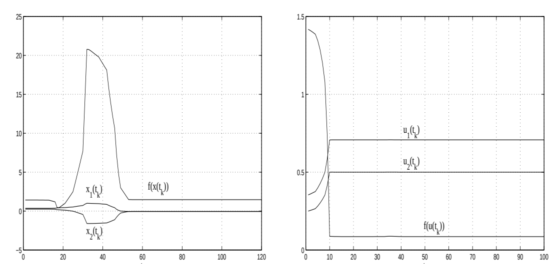


Fig. 5: *Convergence of the network in [9](left); Convergence of the SNN(right)*

Recently, Liu and Wang [9] proposed a one layer recurrent neural network to solve nonsmooth nonconvex optimization problems, which improves the network in [7]. We test the network in [9] to solve (18), where we choose $\sigma = 73$ and $\epsilon = 10^{-1}$. With initial point $(\sqrt{2}/4, 1/4)^T$, the left figure of Fig. 5 gives the convergence of the solution of network in [9], while the right figure of Fig. 5 gives the convergence of the solutions of the SNN. From these two figures, we can find that the SNN is more robust than the NN in [9] for solving (18). However, we should state that the network in [9] can also find the minimizer of (18) with some initial points.

*Example 5.2:* We consider a nonsmooth variant of Nesterov's problem [35]

$$\begin{array}{ll} \min & 4|x_2 - 2|x_1| + 1| + |1 - x_1| \\ \text{s.t.} & 2x_1 - x_2 = 1,\ x_1^2 + |x_2| - 4 \le 0. \end{array} \tag{19}$$

$x^* = (1,1)^T$ is the unique optimal solution of (19) and $f(x^*) = 0$. The objective function and the inequality constrained function are nonsmooth.

From Example 2 of [35], $x^*$ and $(0,-1)^T$ are two Clarke stationary points of (19) without constraints. By simple calculation, the two points are also Clarke stationary points of (19). We choose the smoothing function

$$\tilde{f}(x,\mu) = \tilde{\theta}(x_2 - 2\tilde{\theta}(x_1,\mu) + 1, \mu) + \tilde{\theta}(1 - x_1, \mu)$$

for $f$ and $\tilde{q}(x,\mu)$ for $q(x)$ given in Example 5.1.

We choose $\hat{x} = (0,-1)^T \in \mathbb{X}_1 \cap \mathrm{int}(\mathbb{X}_2)$. The left figure of Fig. 6 shows the convergence of $\|u_{t_k} - x^*\|$ with 40 different initial points, which are $(10\cos(\frac{i\pi}{20}), 10\sin(\frac{i\pi}{20}))^T$, $\quad i = 0, 1, \dots, 39$. The SNN performs well for solving (19) from any of the 40 initial points, which are on the boundary of the circle with center $(0,0)^T$ and radius 10. The right figure of Fig. 6 shows the solution of the SNN with $x_0 = (-10, 0)^T$, which converges to $x^*$.

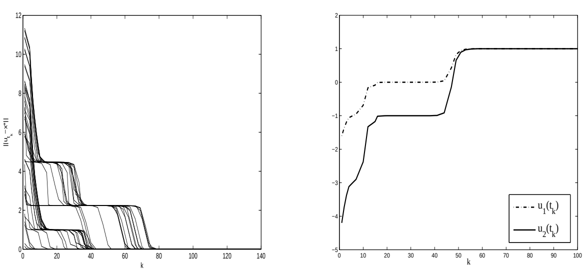

Fig. 6: $\|u(t_k) - x^*\|$ *with 40 given initial points(left); the solution of SNN with the initial point* $x_0 = (-10,0)^T$*(right)*

*Example 5.3:* In this example, we consider

$$\begin{array}{ll} \min & \kappa(Q(x)) \\ \text{s.t.} & 0 \le x \le \mathbf{1},\ \mathbf{1}^T x = 1, \end{array} \tag{20}$$

where $\mathbf{1} = (1, \dots, 1)^T \in \mathbb{R}^n$, $Q(x) = \sum_{i=1}^n x_i Q_i$, $Q_i$ are given matrices in $S_m^{++}$, the cone of symmetric positive definite $m \times m$ matrices.

This example comes from an application of minimizing condition number. It is difficult to evaluate the Lipschitz constant of $\kappa(Q(x))$ over the feasible region. From the constraints in (20), when $x \in \mathbb{X}$, $Q(x) \in S_m^{++}$. Then the condition number of $Q(x)$ is defined by $\kappa(Q(x)) = \frac{\lambda_1(Q(x))}{\lambda_m(Q(x))}$, where $\lambda_1(Q(x)), \dots, \lambda_m(Q(x))$ are the non-increasing ordered eigenvalues of $Q(x)$. In this example, we want to find a matrix in $co\{Q_1, \dots, Q_n\}$ such that it has the smallest condition number, where "$co$" denotes the convex hull.

For given $l, u \in \mathbb{R}_{++}$ with $l \le u$, the following Matlab code is used to generate $Q_1, \dots, Q_n \in S_m^{++}$.

```
R=randn(m,n);[U,D,V]=svd(R(:,1+m*(i-1):m*i));
for j=1:m
D(j,j)=median([l, u, D(j,j)]);
```

| $[l,u]$ | $\lambda_1(Q(\bar{x}))$ | $\lambda_{20}(Q(\bar{x}))$ | $\kappa(Q(u_0))$ | $\kappa(Q(\bar{x}))$ |
|---|---|---|---|---|
| $[0.5, 64]$ | 31.6774 | 10.6844 | 26.5687 | 2.9648 |
| $[5, 50]$ | 29.5896 | 11.9007 | 8.2545 | 2.4864 |
| $[20, 30]$ | 27.5574 | 20.556 | 1.4803 | 1.3406 |
| $[24, 26]$ | 25.2444 | 24.1842 | 1.0820 | 1.0438 |

TABLE I: Numerical results of the SNN for Example 5.3

```
end
Q=U'*D*U;
```

We choose $\hat{x} \in \mathbb{X}_1 \cap \mathrm{int}(\mathbb{X}_2)$ with $\hat{x}_i = \frac{1}{n}$. We use the smoothing function of the objective function given in [23], specially,

$$\tilde{f}(x,\mu) = -\frac{\ln(\sum_{i=1}^m e^{\lambda_i(Q(x))/\mu})}{\ln(\sum_{i=1}^m e^{-\lambda_i(Q(x))/\mu})}.$$

We define $\tilde{q}(x,\mu) = \sum_{i=1}^{2n} \tilde{\phi}_3(g_i(x), \mu)$, where $\tilde{\phi}_3$ is defined in Example 2.1, $g_i(x) = -x_i$ and $g_{n+i}(x) = x_{n+i} - 1$, $i = 1, 2, \dots, n$.

Table I presents the numerical results using the SNN to solve (20) with $n = 10$, $m = 20$ and initial point $x_0 = (0.5, 0.5, 0, \dots, 0)^T$. When $l = 0.5$ and $u = 64$, the left figure of Fig. 7 shows the convergence of $\lambda_1(Q(u_t))$, $\lambda_{20}(Q(u_t))$ and $\kappa(Q(u_t))$ of the SNN with this initial point.

It is known that the condition number function $\kappa$ is nonsmooth nonconvex. $\kappa$ is not differentiable at $x$ when $Q(x)$ has multiple eigenvalues. In order to show the effectiveness of the SNN, we consider a special case, in which we generate $Q_1$ by $l = u = 30$, $Q_i$ by $l = 5$ and $u = 50$ for $i = 2, \dots, 10$. Then, the optimal solution of (20) is $x^* = (1, 0, \dots, 0)^T$ and $\kappa(Q(x^*)) = 1$. The right figure of Fig. 7 shows the eigenvalues of $Q(x)$ at initial point $x_0 = u_0 = (0.1, \dots, 0.1)^T$ and $\bar{x}$ obtained by the SNN.

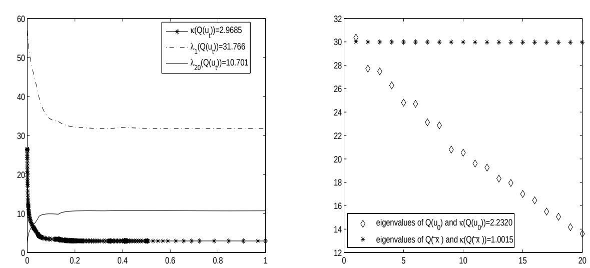

Fig. 7: *Convergence of* $\lambda_1(Q(u_t))$, $\lambda_{20}(Q(u(t)))$ *and* $\kappa(Q(u(t)))$*(left);* $\lambda_i(Q(u_0))$ *and* $\lambda_i(Q(\bar{x}))$, $i = 1, \dots, 20$ *(right).*

*Example 5.4:* In this example, we test our proposed network into the Prostate cancer problem in [36]. The date is consisted of the records of 97 men, which is divided into a training set with 67 observations and a test set with 30 observations. The predictors are eight clinical measures: lcavol, lweight, age, lbph, svi, lcp, pleason and pgg45. In this example, we want to find fewer main factors with smaller prediction error, where the prediction error is the mean square error of the 30 observations in the test set. Then the considered

| | SNN | | | LASSO | BS | IRL1 |
|---|---|---|---|---|---|---|
| $\lambda$ | 6.5 | 6.875 | 18.95 | | | |
| $\bar{x}_1$ | 0.6524 | 0.6817 | 0.7641 | 0.533 | 0.740 | 0.619 |
| $\bar{x}_2$ | 0.2390 | 0.2797 | 0.1267 | 0.169 | 0.316 | 0.236 |
| $\bar{x}_3$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\bar{x}_4$ | 0.0878 | 0 | 0 | 0.002 | 0 | 0.100 |
| $\bar{x}_5$ | 0.1599 | 0.0957 | 0 | 0.094 | 0 | 0.186 |
| $\bar{x}_6$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\bar{x}_7$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $\bar{x}_8$ | 0 | 0 | 0 | 0 | 0 | 0 |
| $N^*$ | 4 | 3 | 2 | 4 | 2 | 4 |
| Error | 0.4321 | 0.4389 | 0.4772 | 0.479 | 0.492 | 0.468 |

TABLE II: Variable selection by the SNN, Lasso, Best subset and Iterative Reweighted $l_1$ norm methods

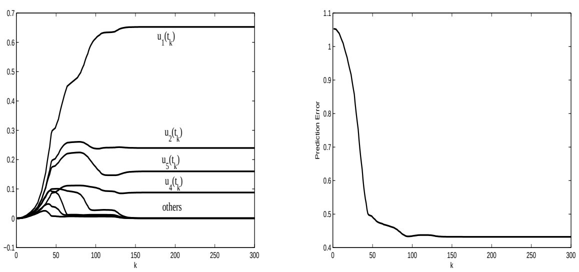


Fig. 8: Convergence of $u(t_k)$ and prediction error

optimization is modeled as

$$\begin{aligned} \min \quad & \|Ax-b\|^2 + \lambda \sum_{i=1}^{8} \frac{3|x_i|}{1+3|x_i|}, \\ \text{s.t.} \quad & 0 \le x_i \le 1, \quad i = 1, 2, \ldots, 8, \end{aligned} \tag{21}$$

where $A \in \mathbb{R}^{67\times 8}$, $b \in \mathbb{R}^{67}$. The objective function is nonsmooth and nonconvex.

Choose $\hat{x} = (0.5, \ldots, 0.5)^T$ and $\nu(t) = \mu_0 e^{-0.1t}$. With initial point $x_0 = (0, \ldots, 0)^T$, the numerical results of using the SNN for solving (21) with different values of $\lambda$ are listed in Table II, where $N^*$ is the number of nonzero elements in $\bar{x}$. The results with three famous methods LASSO , BS (Best Subsets) and IRL1 ( Iterative Reweighted $l_1$ norm) [36] are also given in Table II. We can find that the SNN can find the main factors with smaller prediction error. Additionally, the solution $u(t_k)$ of the SNN and the prediction error along this solution with $\lambda = 6.5$ are illustrated in Fig. 8.

*Example 5.5:* We consider the following nonsmooth nonconvex optimization

$$\begin{aligned} \min \quad & \|Hx-p\|^2 + 0.002 \sum_{i=1}^{n} \psi(x_i) \\ \text{s.t.} \quad & \mathbf{1}^T x = \gamma,\ 0 \le x \le \mathbf{1}, \end{aligned} \tag{22}$$

where $H = (H_{ij})_{n\times n}$, $p = (p_i)_{n\times 1}$ are defined as $H_{ij} = e^{-2(i/3)^2-2(j/3)^2}$, $p_i = 1/i$, $i, j = 1, 2, \ldots, n$, $\gamma = \mathbf{1}^T p$ and $\psi : \mathbb{R} \to \mathbb{R}$ is defined by $\psi(s) = \frac{0.5|s|}{1+0.5|s|}$.

Optimization problem (22) arises frequently in a number of engineering and economic applications, including image restoration, signal processing, system identification, filter design, regression analysis and robot control [18]–[23].

| $x_0^1$ | SQP-GS | | | SNN | | |
|---|---|---|---|---|---|---|
| $n$ | time | $f(\bar{x})$ | fea-err($\bar{x}$) | time | $f(\bar{x})$ | fea-err($\bar{x}$) |
| 16 | 136.3 | 0.2337 | 0 | 0.6045 | 0.2337 | 8.88E-16 |
| 32 | 624.9 | 0.2642 | 8.88E-16 | 1.3795 | 0.2642 | 4.44E-15 |
| 64 | 665.0 | 0.2802 | 8.88E-16 | 5.9176 | 0.2803 | 1.77E-15 |
| $x_0^2$ | SQP-GS | | | SNN | | |
| $n$ | time | $f(\bar{x})$ | fea-err($\bar{x}$) | time | $f(\bar{x})$ | fea-err($\bar{x}$) |
| 16 | 136.1 | 0.2337 | 0 | 0.6120 | 0.2337 | 0 |
| 32 | 623.8 | 0.2642 | 0 | 1.2501 | 0.2642 | 0 |
| 64 | 679.5 | 0.2801 | 8.88E-16 | 2.3482 | 0.2801 | 8.88E-16 |

TABLE III: The SQP-GS and the SNN for Example 5.5

| $x_0^3$ | | | | $x_0^2$ | | | |
|---|---|---|---|---|---|---|---|
| $n$ | time | $f(\bar{x})$ | fea-err($\bar{x}$) | n | time | $f(\bar{x})$ | fea-err($\bar{x}$) |
| 16 | 0.9674 | 0.2337 | 0 | 256 | 7.545 | 0.2930 | 0 |
| 32 | 2.2694 | 0.2642 | 0 | 1024 | 82.17 | 0.2973 | 2.88E-16 |
| 64 | 8.7204 | 0.2802 | 2.66E-15 | 4096 | 387.1 | 0.3392 | 3.38E-14 |

TABLE IV: The SNN for Example 5.5 with $x_0^2$ and $x_0^3$ in (23)

Choose $\hat{x} = \frac{\gamma}{n} e$ and $\nu(t) = \mu_0 e^{-0.1t}$. We define the smoothing functions $\tilde{\psi}(x,\mu) = \psi(\tilde{\theta}(x,\mu))$ of $\psi$ and $\tilde{q}(x,\mu)$ of $q$ with the format given in Example 5.1. Let

$$\begin{aligned} & x_0^1 = (r, 1, \cdots, 1)^T \notin \mathbb{X}_1 \cup \mathbb{X}_2, \quad x_0^2 = p \in \mathbb{X}, \\ & x_0^3 = (r, 0, \cdots, 0)^T \in \mathbb{X}_1 \cap \mathbb{X}_2^C. \end{aligned} \tag{23}$$

Table III shows numerical results of the SQP-GS [25] and the SNN for solving (22) with initial points $x_0^1$ and $x_0^2$. From this table, we can see that the SNN performs better than the SQP-GS in the sense that the SNN can obtain almost same values of $f(\bar{x})$ and fea-err($\bar{x}$) with much less CPU time. In [25], the SQP-GS needs the objective function to be differentiable at the initial point. Table IV shows that the SNN is effective with initial point $x_0^3$, at which the objective function is not differentiable. Table IV also illustrates that the SNN performs well for solving (22) with high dimensions.

Since there is an affine equality constraint in (22), the proposed network is very sensitive and the computation time is long when the dimension $n$ is large. To the best of our knowledge, it is an open and interesting problem on how to solve the large dimension nonsmooth nonconvex optimization problem with affine equality constraints effectively and fast.

## VI. Conclusions

In this paper, we propose a neural network described by an ordinary differential equation, to solve a class of nonsmooth nonconvex optimization problems, which have wide applications in engineering, sciences and economics. Based on the closed form expression of the project operator on the constraints defined by a class of affine equalities, we choose the neural network with projection. Additionally, the penalty function method is also introduced into our system to handle the convex inequality constraints. To avoid solving the differential inclusion and overcome the difficulty in choosing the exact penalty parameter, we make use of the smoothing techniques to approximate the nonsmooth functions and construct a continuous function to replace the fixed parameter. Only with the initial point belonging to the equality constraints, which can be calculated easily by the project operator, we can prove theoretically that any solution of the proposed network converges to the critical point set of the optimization problem. Finally, in order to show the efficiency and superiority of the proposed network, some numerical examples and comparisons are presented, including the Rosenbrock function, the Nesterov's problem, the minimization of condition number, and a familiar optimization model in image restoration, signal processing and identification. By the numerical experiments, it is as expected that the proposed network in this paper performs better than the neural network method in [9], the two famous iterative algorithms Lasso and IRL1, and the well-known statistical optimization algorithms Best Subset and SQP-GS [25]. There are two possible reasons why the proposed network can provide better numerical results than these existing methods. The first is that the smoothing parameter is updating continuously in the proposed network and the global convergence can also be guaranteed. The second reason is that the continuous penalty parameter $\sigma(u(t),\nu(t))$ controls the proposed network and let it solve the constrained optimization effectively. However, we can not prove these two reasons in theory, which inspires us to explore the reasons in further work.

## VII. APPENDIX

*Proof of Proposition 2.1* It is easy to see that these compositions satisfy (i) and (ii) of Definition 2.1. We only need to consider (iii) and (iv) of Definition 2.1. By the chain rules of the subgradient in [31], (a) holds naturally.

(b) Condition (iii) of Definition 2.1 is derived as the following

$$\begin{aligned}&\{\lim_{z\to x,\mu\downarrow 0}\nabla_z(\psi(\tilde{\varphi}(z,\mu))\}\\ =&\nabla_s\psi(s)_{s=\varphi(x)}\{\lim_{z\to x,\mu\downarrow 0}\nabla_z\tilde{\varphi}(z,\mu)\}\\ \subseteq&\nabla_s\psi(s)_{s=\varphi(x)}\partial\varphi(x)=\partial(\psi\circ\varphi)(x),\end{aligned}$$

where we use $\{\lim_{z\to x,\mu\downarrow 0}\nabla_z\tilde{\varphi}(z,\mu)\}\subseteq\partial\varphi(x)$ and [31, Theorem 2.3.9 (ii)].

Condition (iv) of Definition 2.1 follows from

$$|\nabla_\mu\psi(\tilde{\varphi}(x,\mu))|\le|\nabla_s\psi(s)_{s=\tilde{\varphi}(x,\mu)}||\nabla_\mu\tilde{\varphi}(x,\mu)|\le l_\psi\kappa_{\tilde{\varphi}}.$$

Similar to the analysis in (a), we omit the proof of (c).

(d) Denote $\tilde{\psi}\circ\tilde{\varphi}(x,\mu)=\tilde{\psi}(\tilde{\varphi}(x,\mu),\mu)$. For any fixed $\mu>0$, since $\tilde{\psi}(\cdot,\mu)$ is convex and non-decreasing, and $\tilde{\varphi}(\cdot,\mu)$ is convex , we get that $\tilde{\psi}\circ\tilde{\varphi}(\cdot,\mu)$ is convex. Hence, for any fixed $\mu>0$, $z,v\in\mathbb{R}^n$ and $\tau>0$, we have

$$\frac{\tilde{\psi}\circ\tilde{\varphi}(z+\tau v,\mu)-\tilde{\psi}\circ\tilde{\varphi}(z,\mu)}{\tau}\ge\langle\nabla_z(\tilde{\psi}\circ\tilde{\varphi})(z,\mu),v\rangle.$$

Let $z\to x$ and $\mu\downarrow 0$, and then passing $\tau$ to 0 in the above inequality, we have

$$\psi(\varphi)'(x;v)\ge\langle\lim_{z\to x,\mu\downarrow 0}\nabla_z(\tilde{\psi}\circ\tilde{\varphi})(z,\mu),v\rangle,\quad\forall v\in\mathbb{R}^n.$$

By the definition of the subgradient, we obtain

$$\{\lim_{z\to x,\mu\downarrow 0}\nabla_z(\tilde{\psi}\circ\tilde{\varphi})(z,\mu)\}\subseteq\partial\psi(\varphi(x)),$$

which proves that $\tilde{\psi}\circ\tilde{\varphi}$ satisfies (iii) of Definition 2.1.

Condition (iv) of Definition 2.1 follows from

$$|\nabla_\mu\tilde{\psi}(\tilde{\varphi}(x,\mu),\mu)|\le\kappa_{\tilde{\psi}}+l_\psi\kappa_{\tilde{\varphi}}.$$

■

In order to give the proof of Theorem 4.1, we need the following preliminary analysis.

For a given $x\in\mathbb{R}^n$, denote

$$I^+(x)=\{\,i\,|\,g_i(x)>0,\},\ I^0(x)=\{\,i\,|\,g_i(x)=0\}.$$

We need the following lemmas to obtain our main results.

*Lemma 7.1:* The following inequality holds

$$\langle x-\hat{x},\nabla_x\tilde{q}(x,\mu)\rangle\ge\beta,\quad\forall x\notin\mathbb{X}_2,\ \mu\in(0,\,\mu_0].$$

*Proof:* For any $x\notin\mathbb{X}_2$, $I^+(x)\ne\emptyset$. (8) implies

$$\tilde{g}_i(x,\mu)\ge g_i(x)>0,\quad\forall i\in I^+(x).\tag{24}$$

From (iv) of Definition 2.1, we have

$$\tilde{g}_i(\hat{x},\mu)\le g_i(\hat{x})+\kappa_{\tilde{g}_i}\mu\le-4\beta+\kappa\mu,\quad i=1,2,\ldots,m.\tag{25}$$

From the convexity of $\tilde{g}_i(\cdot,\mu)$, (24) and (25), for any $i\in I^+(x)$, we have

$$\langle x-\hat{x},\nabla_x\tilde{g}_i(x,\mu)\rangle\ge\tilde{g}_i(x,\mu)-\tilde{g}_i(\hat{x},\mu)\ge 4\beta-\kappa\mu.\tag{26}$$

For $\mu\le\mu_0$, (5) and (26) imply that for any $x\notin\mathbb{X}_2$, $i\in I^+(x)$,

$$\langle x-\hat{x},\nabla_x\tilde{\phi}(\tilde{g}_i(x,\mu),\mu)\rangle\ge\frac{1}{2}(4\beta-\kappa\mu).\tag{27}$$

When $\mu\le\mu_0$, $\tilde{g}_i(\hat{x},\mu)\le 0$, $i=1,2,\ldots,m$. Since $\tilde{\phi}(\cdot,\mu)$ is convex and $\tilde{\phi}(s,\cdot)$ is non-decreasing, for all $i=1,2,\ldots,m$, we obtain

$$\begin{aligned}&\langle x-\hat{x},\nabla_x\tilde{\phi}(\tilde{g}_i(x,\mu),\mu)\rangle\\ \ge&\tilde{\phi}(\tilde{g}_i(x,\mu),\mu)-\tilde{\phi}(\tilde{g}_i(\hat{x},\mu),\mu)\ge-\mu.\end{aligned}\tag{28}$$

Combining (27) and (28), when $x\notin\mathbb{X}_2$ and $\mu\le\mu_0$, we obtain

$$\langle x-\hat{x},\nabla_x\tilde{q}(x,\mu)\rangle\ge\frac{1}{2}(4\beta-\kappa\mu)-(m-1)\mu\ge\beta.$$

■

*Lemma 7.2:* For any $x_0\in\mathbb{R}^n$, there is a $T>0$ such that (12) has a solution $u\in C^1[0,T)$. Moreover, any solution of (12) in $C^1[0,T)$ satisfies $u(t)\in\mathbb{X}_1$ for all $t\in[0,T)$.

*Proof:* Since the right hand function in the system (12) is continuous, there are a $T>0$ and $u\in C^1[0,T)$ such that $u(t)$ satisfies (12) for all $t\in[0,T)$, see [37]. Differentiating $\frac{1}{2}\|Au(t)-b\|^2$ along this solution, from $AP=0$, we obtain

$$\frac{d}{dt}\frac{1}{2}\|Au(t)-b\|^2=\langle A^T(Au(t)-b),\dot{u}(t)\rangle=0,$$

which derives that $\|Au(t)-b\|^2=\|Au_0-b\|^2,\quad\forall t\in[0,T)$. Since $u_0=Px_0+c\in\mathbb{X}_1$, we get $\|Au_0-b\|^2=0$. Hence, $\|Au(t)-b\|^2=0$ and $u(t)\in\mathbb{X}_1$, $\forall t\in[0,T)$. ■

*Lemma 7.3:* The level set $\{x\in\mathbb{X}_1\,|\,\tilde{q}(x,\mu_0)\le\eta\}$ is bounded for any $\eta>0$.

*Proof:* First, we prove that for any $\eta>0$, the level set $\Gamma=\{x\in\mathbb{X}_1\mid\max_{1\le i\le m}\tilde{g}_i(x,\mu_0)\le\eta\}$ is bounded. Since

$\mathbb{X}$ is bounded and $\Gamma$ is a subset of $\mathbb{X}_1$, $\Gamma \bigcap \mathbb{X}_2$ is bounded. In order to prove the boundedness of $\Gamma$, we need to consider the set $\Gamma \bigcap \mathbb{X}_2^C$. Assume on contradiction that there exist $\bar{\eta} > 0$ and a sequence $\{x_k\} \subseteq \mathbb{X}_1 \bigcap \mathbb{X}_2^C$ such that

$$\max_{1\le i\le m} \tilde{g}_i(x_k, \mu_0) \le \bar{\eta} \quad \text{and} \quad \lim_{k\to\infty} \|x_k\| = \infty. \tag{29}$$

Denote $\psi_k(\tau) = \max_{1\le i\le m} \tilde{g}_i((1-\tau)\hat{x} + \tau x_k, \mu_0)$, $k = 1, 2, \ldots$. Since $\tilde{g}_i(\cdot, \mu_0)$ is convex, $i = 1, 2, \ldots, m$, $\psi_k$ is convex on $[0, +\infty)$, $k = 1, 2, \ldots$. From (3) and $\mu_0 \le \frac{2\beta}{\kappa}$, for $k = 1, 2, \ldots$,

$$\psi_k(0) = \max_{1\le i\le m} \tilde{g}_i(\hat{x}, \mu_0) \le \max_{1\le i\le m} g_i(\hat{x}) + \kappa_{\tilde{g}_i}\mu_0 \le -2\beta,$$
$$\psi_k(1) = \max_{1\le i\le m} \tilde{g}_i(x_k, \mu_0) \ge \max_{1\le i\le m} g_i(x_k) > 0.$$

Then, for each $k = 1, 2, \ldots$, there exists $\tau_k \in (0, 1)$ such that

$$\psi_k(\tau_k) = \max_{1\le i\le m} \tilde{g}_i((1-\tau_k)\hat{x} + \tau_k x_k, \mu_0) = 0. \tag{30}$$

Since $\psi_k$ is convex, $\nabla\psi_k$ is non-decreasing, $k = 1, 2, \ldots$. From the mean value theorem, for each $k = 1, 2, \ldots$, there exists $\hat{\tau}_k \in [0, \tau_k]$ such that

$$\nabla\psi_k(\tau_k) \ge \nabla\psi_k(\hat{\tau}_k) = \frac{\psi_k(\tau_k) - \psi_k(0)}{\tau_k} \ge \frac{2\beta}{\tau_k}. \tag{31}$$

Using the non-decreasing of $\tilde{g}_i(x, \cdot)$ and (30), for all $i = 1, 2, \ldots, m$, we have

$$g_i((1-\tau_k)\hat{x} + \tau_k x_k) \le \tilde{g}_i((1-\tau_k)\hat{x} + \tau_k x_k, \mu_0) \le 0,$$

which implies that $(1-\tau_k)\hat{x} + \tau_k x_k \in \mathbb{X}_2$, $k = 1, 2, \ldots$. Combining this with $(1-\tau_k)\hat{x} + \tau_k x_k \in \mathbb{X}_1$, $k = 1, 2, \ldots$, we have

$$(1-\tau_k)\hat{x} + \tau_k x_k \in \mathbb{X}, \quad k = 1, 2, \ldots. \tag{32}$$

Since $\mathbb{X}$ is bounded, there exists $R > 0$ such that $\|x - \hat{x}\| \le R$, $\forall x \in \mathbb{X}$. Hence, (32) implies

$$\|(1-\tau_k)\hat{x} + \tau_k x_k - \hat{x}\| = \tau_k\|\hat{x} - x_k\| \le R, \quad k = 1, 2, \ldots.$$

Since $\lim_{k\to\infty} \|x_k\| = \infty$, from the above inequality, we obtain $\lim_{k\to\infty} \tau_k = 0$. Owning to (31), $\lim_{k\to\infty} \nabla\psi_k(\tau_k) = \infty$.

From the convex inequality of $\psi_k$, for all $k = 1, 2, \ldots$,

$$\psi_k(1) \ge \psi_k(\tau_k) + (1-\tau_k)\nabla\psi_k(\tau_k) = (1-\tau_k)\nabla\psi_k(\tau_k),$$

which follows that $\lim_{k\to\infty} \max_{1\le i\le m} \tilde{g}_i(x_k, \mu_0) = \lim_{k\to\infty} \psi_k(1) = \infty$. This is a contradiction to (29). Hence, the level set $\{x \in \mathbb{X}_1 \mid \max_{1\le i\le m} \tilde{g}_i(x, \mu_0) \le \eta\}$ is bounded for any $\eta > 0$.

From the definition of $\tilde{q}$ and non-decreasing of $\tilde{\phi}(s, \cdot)$, we obtain

$$\tilde{q}(x, \mu_0) \ge \sum_{i=1}^{m} \max\{0, \tilde{g}_i(x, \mu_0)\} \ge \max_{1\le i\le m} \tilde{g}_i(x, \mu_0). \tag{33}$$

Thus, for any $\eta > 0$, $\{x \in \mathbb{X}_1 \mid \tilde{q}(x, \mu_0) \le \eta\} \subseteq \{x \in \mathbb{X}_1 \mid \max_{1\le i\le m} \tilde{g}_i(x, \mu_0) \le \eta\}$. Since $\{x \in \mathbb{X}_1 \mid \max_{1\le i\le m} \tilde{g}_i(x, \mu_0) \le \eta\}$ is bounded, $\{x \in \mathbb{X}_1 \mid \tilde{q}(x, \mu_0) \le \eta\}$ is bounded. ∎

*Proof of Theorem 4.1* From Lemma 7.2, there is a $T > 0$ such that (12) has a solution $u(t) \in C^1[0, T)$, where $[0, T)$ is the maximal existence interval of $t$. We suppose $T < \infty$.

Differentiating $\tilde{q}(u(t), \nu(t)) + \kappa_{\tilde{q}}\nu(t)$ along this solution of (12), we have

$$\frac{d}{dt}(\tilde{q}(u(t), \nu(t)) + \kappa_{\tilde{q}}\nu(t)) = \langle \nabla_u \tilde{q}(t), \dot{u} \rangle + (\nabla_\nu \tilde{q}(t) + \kappa_{\tilde{q}})\dot{\nu}.$$

From $\nabla_\nu \tilde{q}(t) + \kappa_{\tilde{q}} \ge 0$, $\dot{\nu}(t) \le 0$ and $P^2 = P$, we have

$$\begin{aligned} &\frac{d}{dt}(\tilde{q}(u(t), \nu(t)) + \kappa_{\tilde{q}}\nu(t)) \\ \le&\langle \nabla_u \tilde{q}(t), -P(\nabla_u \tilde{f}(t) + \sigma(t)\nabla_u \tilde{q}(t)) \rangle \\ \le&|\langle P\nabla_u \tilde{q}(t), P\nabla_u \tilde{f}(t) \rangle| - \sigma(t)\|P\nabla_u \tilde{q}(t)\|^2. \end{aligned} \tag{34}$$

Since $u(t) \in \mathbb{X}_1$, we have $P(u(t) - \hat{x}) = u(t) - \hat{x}$, $\forall t \in [0, T)$. Meantime, if $u(t) \in \mathbb{X}_1 \cap \mathbb{X}_2^C$ for some $t \in [0, T)$, from Lemma 7.1, we have

$$\begin{aligned} &\langle u(t) - \hat{x}, P\nabla_u \tilde{q}(u(t), \nu(t)) \rangle \\ =&\langle u(t) - \hat{x}, \nabla_u \tilde{q}(u(t), \nu(t)) \rangle \ge \beta, \end{aligned}$$

which implies that $\|P\nabla_u \tilde{q}(u(t), \nu(t))\|\|u(t) - \hat{x}\| \ge \beta$.

Thus, for any $t \in [0, T)$ such that $u(t) \in \mathbb{X}_1 \cap \mathbb{X}_2^C$, we have

$$\begin{aligned} &\max\{\beta^2, \|P\nabla_u \tilde{q}(u(t), \nu(t))\|^2\|u(t) - \hat{x}\|^2\} \\ =&\|P\nabla_u \tilde{q}(u(t), \nu(t))\|^2\|u(t) - \hat{x}\|^2. \end{aligned}$$

Since $u(t) \in \mathbb{X}_1$, $\forall t \in [0, T)$, using the above result, the definition of $\sigma(u, \nu)$ and (34), when $u(t) \notin \mathbb{X}_2$, we find

$$\frac{d}{dt}(\tilde{q}(u(t), \nu(t)) + \kappa_{\tilde{q}}\nu(t)) \le -\lambda\beta\nu(t). \tag{35}$$

This implies that $\tilde{q}(u(t), \nu(t)) + \kappa_{\tilde{q}}\nu(t)$ is a non-increasing function of $t$ when $u(t) \in \mathbb{X}_2^C$. On the other hand, when $u(t) \in \mathbb{X}_2$, $\tilde{q}(u(t), \nu(t)) \le \kappa_{\tilde{q}}\nu(t) \le \kappa_{\tilde{q}}\mu_0$. Thus,

$$\begin{aligned} \tilde{q}(u(t), \nu(t)) \le& \max\{\tilde{q}(u_0, \mu_0) + \kappa_{\tilde{q}}\mu_0, \kappa_{\tilde{q}}\mu_0\} \\ \le& q(u_0) + 2\kappa_{\tilde{q}}\mu_0, \quad \forall t \in [0, T). \end{aligned}$$

From Definition 2.1, for all $t \in [0, T)$,

$$\tilde{q}(u(t), \mu_0) \le \tilde{q}(u(t), \nu(t)) + \kappa_{\tilde{q}}|\nu(t) - \mu_0|.$$

Thus, for all $t \in [0, T)$,

$$\tilde{q}(u(t), \mu_0) \le q(u_0) + 3\kappa_{\tilde{q}}\mu_0.$$

Form Lemma 7.3, $u : [0, T) \to \mathbb{R}^n$ is bounded. Then this solution of (12) can be extended. Therefore, this solution of (12) exists globally.

Similarly, we can obtain $\tilde{q}(u(t), \mu_0) \le q(u_0) + 3\kappa_{\tilde{q}}\mu_0$, $\forall t \in [0, \infty)$. Thus, $u : [0, \infty) \to \mathbb{R}^n$ is uniformly bounded, which means that there is a $\rho > 0$ such that $\|u(t)\| \le \rho$, $\forall t \ge 0$.

*Proof of Proposition 4.1*

Denote $u, v \in C^1[0, \infty)$ two solutions of (12) with an initial point $u_0 = Px_0 + c$ and suppose there exists $\hat{t}$ such that $\hat{t} = \inf_{t\ge 0, u(t)\ne v(t)} t$. From Theorem 4.1, there is a $\rho > 0$ such that $\|u(t)\| \le \rho$, $\|v(t)\| \le \rho$, $\forall t \ge 0$.

Denote $r(x, \mu) = -P(\nabla_x \tilde{f}(x, \mu) + \sigma(x, \mu)\nabla_x \tilde{q}(x, \mu))$. When $\nabla_x \tilde{f}(\cdot, \mu)$ and $\nabla_x \tilde{q}(\cdot, \mu)$ are locally Lipschitz for any $\mu \in (0, \infty)$, $\sigma(\cdot, \mu)$ is locally Lipschitz for any $\mu \in (0, \infty)$. Then $r(\cdot, \mu)$ is locally Lipschitz for any $\mu \in (0, \infty)$.

Since $u(t)$, $v(t)$ and $\nu(t)$ are continuous and bounded on

$[0,\infty)$, there is an $L$ such that for any $t \in [\hat{t}, \hat{t}+1]$,

$$\|r(u(t),\nu(t)) - r(v(t),\nu(t))\| \leq L\|u(t)-v(t)\|.$$

Differentiating $\frac{1}{2}\|u(t)-v(t)\|^2$ along the two solutions of (12), we have

$$\frac{d}{dt}\frac{1}{2}\|u(t)-v(t)\|^2 \leq L\|u(t)-v(t)\|^2, \quad \forall t \in [\hat{t}, \hat{t}+1].$$

Applying Gronwall's inequality into the integration of the above inequality, it gives $u(t) = v(t)$, $\forall t \in [\hat{t}, \hat{t}+1]$, which leads a contradiction.■

*Proof of Theorem 4.2* Let $u \in C^1[0,\infty)$ be a solution of (12) with initial point $x_0$. When $u(t) \notin \mathbb{X}_2$, from (35), we have

$$\frac{d}{dt}(\tilde{q}(t) + \kappa_{\tilde{q}}\nu(t)) \leq -\lambda\beta\nu(t) = -\lambda\beta\mu_0 e^{-t}, \ \forall t \geq 0. \quad (36)$$

Integrating the above inequality from 0 to $t$, we get

$$\begin{aligned}&\tilde{q}(u(t),\nu(t)) + \kappa_{\tilde{q}}\nu(t) - \tilde{q}(u_0,\mu_0) - \kappa_{\tilde{q}}\mu_0\\ \leq& -\lambda\beta\mu_0\int_0^t e^{-s}ds = -\lambda\beta\mu_0(1-e^{-t}).\end{aligned}$$

Owning to $\tilde{q}(t) + \kappa_{\tilde{q}}\nu(t) \geq q(u(t)) \geq 0$, $\forall t \geq 0$, we obtain

$$0 \leq \tilde{q}(u_0,\mu_0) + \kappa_{\tilde{q}}\mu_0 - \lambda\beta\mu_0(1-e^{-t}). \quad (37)$$

From (3) and (9), we have

$$q(u_0) + 2m(1+\kappa)\mu_0 \geq \tilde{q}(u_0,\mu_0) + \kappa_{\tilde{q}}\mu_0,$$

then

$$\lambda = \frac{2q(u_0) + 4m(1+\kappa)\mu_0}{\beta\mu_0} \geq \frac{2(\tilde{q}(u_0,\mu_0) + \kappa_{\tilde{q}}\mu_0)}{\beta\mu_0}. \quad (38)$$

(37) and (38) lead to $t \leq \ln 2$.

Therefore, $u(t)$ hits the feasible region $\mathbb{X}_2$ in finite time.

For $t > \ln 2$ and $u(t) \notin \mathbb{X}_2$. Denote $\hat{t} = \sup_{0\leq s<t, u(s)\in\mathbb{X}_2} s$. Then, $u(s) \notin \mathbb{X}_2$ when $s \in (\hat{t}, t]$. Integrating (36) from $\hat{t}$ to $t$, we get

$$\begin{aligned}&\tilde{q}(u(t),\nu(t)) + \kappa_{\tilde{q}}\nu(t)\\ \leq&\tilde{q}(u(\hat{t}),\nu(\hat{t})) + \kappa_{\tilde{q}}\nu(\hat{t}) - \lambda\beta\mu_0\int_{\hat{t}}^t e^{-s}ds \\ \leq&2\kappa_{\tilde{q}}\nu(\hat{t}) + \lambda\beta(\nu(t) - \nu(\hat{t})).\end{aligned} \quad (39)$$

Applying $\lambda \geq 2\kappa_{\tilde{q}}/\beta$ to (39), we get

$$\begin{aligned}q(u(t)) \leq&\tilde{q}(u(t),\nu(t)) + \kappa_{\tilde{q}}\nu(t)\\ \leq&2\kappa_{\tilde{q}}\nu(\hat{t}) + 2\kappa_{\tilde{q}}(\nu(t)-\nu(\hat{t})) = 2\kappa_{\tilde{q}}\nu(t).\end{aligned} \quad (40)$$

Moreover, combining (40) with $q(u(t)) \leq 0$ when $u(t) \in \mathbb{X}_2$, we have that

$$q(u(t)) \leq 2\kappa_{\tilde{q}}\nu(t), \quad \forall t \geq \ln 2.$$

Passing to the suplimit of the above inequality, we obtain

$$0 \leq \limsup_{t\to\infty} q(u(t)) \leq \lim_{t\to\infty} 2\kappa_{\tilde{q}}\nu(t) = 0.$$

Therefore, we deduce that $\lim_{t\to\infty} q(u(t)) = 0$, which means $\{\lim_{t\to\infty} u(t)\} \subseteq \mathbb{X}_2$. Combining this with $u(t) \in \mathbb{X}_1$, $\forall t \in [0,\infty)$, we have $\{\lim_{t\to\infty} u(t)\} \subseteq \mathbb{X}$. ■

To prove the global convergence of (12) to the Clarke stationary point set of (1), we need a lemma on the relationship between the normal cones and the subgradients.

*Lemma 7.4:* If $\lim_{k\to\infty}\mu_k = 0$ and $\lim_{k\to\infty} x_k = x^* \in \mathbb{X}$, then

$$\begin{aligned}&\{\lim_{k\to\infty} P(\nabla_x\tilde{f}(x_k,\mu_k) + \sigma(x_k,\mu_k)\nabla_x\tilde{q}(x_k,\mu_k)\}\\ \subseteq&\partial f(x^*) + N_{\mathbb{X}}(x^*).\end{aligned}$$

*Proof:* From (iii) of Definition 2.1, we have

$$\{\lim_{k\to\infty}\nabla_x\tilde{q}(x_k,\mu_k)\} \subseteq \partial q(x^*).$$

If $x^* \in \mathrm{bd}(\mathbb{X}_2)$, $\partial q(x^*) = \sum_{i\in I^0(x^*)}[0,1]\partial g_i(x^*)$. Since $g_i$ is convex, for any $\tau > 0$,

$$\tau\partial q(x^*) = \sum_{i\in I^0(x^*)}[0,\tau]\partial g_i(x^*).$$

Since $\lim_{k\to\infty} x_k = x^*$, we have $0 < \sigma(x_k,\mu_k) < \infty$, $k = 1,2,\ldots$. Thus,

$$\begin{aligned}&\{\lim_{k\to\infty} P(\nabla_x\tilde{f}(x_k,\mu_k) + \sigma(x_k,\mu_k)\nabla_x\tilde{q}(x_k,\mu_k))\}\\ \subseteq&P(\partial f(x^*) + [0,\infty)\partial q(x^*)).\end{aligned}$$

For any fixed $\tau > 0$, since $A^T(AA^T)^{-1}A(-\partial f(x^*) - \tau\partial q(x^*)) \subseteq N_{\mathbb{X}_2}(x^*)$, we have

$$\begin{aligned}&P(\partial f(x^*) + \tau\partial q(x^*))\\ =&\partial f(x^*) + \tau\partial q(x^*) - A^T(AA^T)^{-1}A(\partial f(x^*) + \tau\partial q(x^*))\\ \subseteq&\partial f(x^*) + N_{\mathbb{X}_1}(x^*) + N_{\mathbb{X}_2}(x^*) = \partial f(x^*) + N_{\mathbb{X}}(x^*).\end{aligned}$$

■

*Proof of Theorem 4.3* From Theorem 4.1, there is a $\rho > 0$ such that $\|u(t)\| \leq \rho$ for all $t \geq 0$ which implies that there is $R > 0$ such that $\|u(t) - \hat{x}\| \leq R$ for all $t \geq 0$. Since $f$ and $q$ are locally Lipschitz, there exist $l_f$ and $l_q$ such that $\|\xi\| \leq l_f$, $\|\eta\| \leq l_q$, $\forall \xi \in \partial f(x)$, $\eta \in \partial q(x)$, $\|x\| \leq \rho$. From (iii) of Definition 2.1, we confirm that $\limsup_{t\to\infty}\|\nabla_u\tilde{f}(u(t),\nu(t))\| \leq l_f$ and $\limsup_{t\to\infty}\|\nabla_u\tilde{q}(u(t),\nu(t))\| \leq l_q$, which means that there are $l_{\tilde{f}}$ and $l_{\tilde{q}}$ such that $\|\nabla_u\tilde{f}(u(t),\nu(t))\| \leq l_{\tilde{f}}$ and $\|\nabla_u\tilde{q}(u(t),\nu(t))\| \leq l_{\tilde{q}}$, $\forall t \geq 0$.

(i) From (12) and $P^2 = P$, we have

$$\begin{aligned}&\langle\nabla_u\tilde{f}(u(t),\nu(t)) + \sigma(u(t),\nu(t))\nabla_u\tilde{q}(u(t),\nu(t)), \dot{u}(t)\rangle\\ =&-\|P(\nabla_u\tilde{f}(u(t),\nu(t)) + \sigma(u(t),\nu(t))\nabla_u\tilde{q}(u(t),\nu(t)))\|^2\\ =&-\|\dot{u}(t)\|^2\end{aligned} \quad (41)$$

Calculating $\frac{d}{dt}\tilde{f}(u(t),\nu(t)) + \sigma(u(t),\nu(t))\frac{d}{dt}\tilde{q}(u(t),\nu(t))$ along this solution of (12), from (41), we obtain

$$\begin{aligned}&\frac{d}{dt}\tilde{f}(u(t),\nu(t)) + \sigma(u(t),\nu(t))\frac{d}{dt}\tilde{q}(u(t),\nu(t))\\ =&-\|\dot{u}(t)\|^2 + (\nabla_\nu\tilde{f}(t) + \sigma(t)\nabla_\nu\tilde{q}(t))\dot{\nu}(t).\end{aligned} \quad (42)$$

On the other hand, we have

$$\begin{aligned}&\frac{d}{dt}\tilde{f}(u(t),\nu(t)) - \sigma(u(t),\nu(t))\frac{d}{dt}\tilde{q}(u(t),\nu(t))\\ =&-\|P\nabla_u\tilde{f}(t)\|^2 + \sigma^2(t)\|P\nabla_u\tilde{q}(t))\|^2\\ &+(\nabla_\nu\tilde{f}(t) - \sigma(t)\nabla_\nu\tilde{q}(t))\dot{\nu}(t).\end{aligned} \quad (43)$$

Adding (42) and (43) gives

$$2\frac{d}{dt}\tilde{f}(u(t),\nu(t)) = -\|\dot{u}(t)\|^2 - \|P\nabla_u \tilde{f}(t)\|^2 + 2\nabla_\nu \tilde{f}(t)\dot{\nu}(t) + \sigma^2(t)\|P\nabla_u \tilde{q}(t))\|^2. \quad (44)$$

From the definition of $\sigma(x,\mu)$ in (12), for all $t \in [0,\infty)$, we get

$$\begin{aligned}&\sigma(u(t),\nu(t))\|P\nabla_u \tilde{q}(u(t),\nu(t))\|\\ \leq&\frac{|\langle P\nabla_x \tilde{f}(x,\mu), P\nabla_x \tilde{q}(x,\mu)\rangle|}{\|P\nabla_x \tilde{q}(x,\mu)\|}\\ &+\frac{\lambda\beta\nu(t)\|u(t)-\hat{x}\|^2\|P\nabla_u \tilde{q}(u(t),\nu(t))\|}{\beta^2},\\ \leq&\|P\nabla_u \tilde{f}(u(t),\nu(t))\| + \varrho\nu(t),\end{aligned} \quad (45)$$

where $\varrho = \frac{\lambda R^2 l_{\tilde{q}}}{\beta}$. Substituting (45) into (44) and using $\|P\| = 1$, we have

$$\begin{aligned}&2\frac{d}{dt}(\tilde{f}(u(t),\nu(t)) + \kappa_{\tilde{f}}\nu(t))\\ \leq& -\|\dot{u}(t)\|^2 + 2\varrho l_{\tilde{f}}\nu(t) + \varrho^2\nu^2(t).\end{aligned} \quad (46)$$

Let $\delta = 2\varrho l_{\tilde{f}}\mu_0 + \frac{\varrho^2\mu_0^2}{2}$. Integrating (46) from 0 to $t$, we have

$$\begin{aligned}&\int_0^t \|\dot{u}_s\|^2 ds\\ \leq&2\tilde{f}(u_0,\mu_0) - 2\tilde{f}(u(t),\nu(t)) + 2\kappa_{\tilde{f}}\mu_0 - 2\kappa_{\tilde{f}}\nu(t) + \delta\\ \leq&2f(u_0) - 2\min_{\|x\|\leq\rho} f(x) + 4\kappa_{\tilde{f}}\mu_0 + \delta.\end{aligned}$$

(ii) Let

$$w(t) = 2\tilde{f}(u(t),\nu(t)) + 2\kappa_{\tilde{f}}\nu(t) + 2\varrho l_{\tilde{f}}\nu(t) + \frac{1}{2}\varrho^2\nu^2(t).$$

From (46) and $\nu(t) = \mu_0 e^{-t}$, we have

$$\frac{d}{dt}w(t) \leq -\|\dot{u}(t)\|^2 \leq 0. \quad (47)$$

In addition, we have $w(t) \geq 2\min_{\|x\|\leq\rho} f(x)$. Since $w(t)$ is bounded from below and non-increasing on $[0,\infty)$, we have

$$\lim_{t\to\infty} w(t) \quad \text{exists and} \quad \lim_{t\to\infty}\frac{d}{dt}w(t) = 0. \quad (48)$$

From (3) and $\lim_{t\to\infty}\nu(t) = 0$, we have

$$\lim_{t\to\infty} f(u(t)) = \lim_{t\to\infty}\tilde{f}(u(t),\nu(t)) \quad \text{exists.}$$

Moreover, (47) and (48) imply that $\lim_{t\to\infty}\|\dot{u}(t)\| = 0$.

(iii) If $x^* \in \{\lim_{t\to\infty} u(t)\}$, there exists a sequence $\{t_k\}$ such that $\lim_{k\to\infty} u(t_k) = x^*$ and $\lim_{k\to\infty}\nu(t_k) = 0$ as $\lim_{k\to\infty} t_k = \infty$. From Theorem 4.2, we know that $x^* \in \mathbb{X}$. From Lemma 7.4 and result (ii) of this theorem, we get $0 \in \partial f(x^*) + N_{\mathbb{X}}(x^*)$, which implies that there exists $\xi \in \partial f(x^*)$ such that $\langle \xi, x - x^*\rangle \geq 0$, $\forall x \in \mathbb{X}$. Thus, $x^*$ is a Clarke stationary point of (1).■

*Proof of Corollary 4.1* Denote $u : [0,T) \to \mathbb{R}^n$ a solution of (17) with initial point $x_0$, where $[0,T)$ is the maximal existence interval of $t$. From Theorems 4.1, 4.2 and 4.3, we only need to prove the boundedness of $u(t)$ on $[0,T)$. Differentiating $\tilde{f}(u(t),\nu(t))$ along this solution of (17), from $P^2 = P$, we have

$$\begin{aligned}\frac{d}{dt}\tilde{f}(u(t),\nu(t)) =&\langle\nabla_u \tilde{f}(t), -P\nabla_u \tilde{f}(t)\rangle + \langle\nabla_\nu \tilde{f}(t), \dot{\nu}(t)\rangle\\ \leq& -\|P\nabla_u \tilde{f}(t)\|^2 - \kappa_{\tilde{f}}\dot{\nu}(t),\end{aligned}$$

which follows that $\frac{d}{dt}(\tilde{f}(u(t),\nu(t)) + \kappa_{\tilde{f}}\nu(t)) \leq 0$. Thus, $\tilde{f}(u(t),\nu(t)) + \kappa_{\tilde{f}}\nu(t) \leq \tilde{f}(u_0,\nu_0) + \kappa_{\tilde{f}}\nu_0$, $\forall t \in [0,T)$. Similar to the analysis in Theorem 4.1, when $f$ is level-bounded, we get that $u(t)$ is bounded on $[0,T)$.■

## ACKNOWLEDGEMENT

The authors would like to thank the Editor in Chief, Associate Editor and the reviewers for their insightful and constructive comments, which help us to enrich the content and improve the presentation of the results in this paper.

## REFERENCES

[1] A. Cichocki and R. Unbehauen, *Neural Networks for Optimization and Signal Processing*. New York: Wiley, 1993.

[2] E.K.P. Chong, S. Hui, S.H. Żak, "An analysis of a class of neural networls for solving linear programming problems", IEEE Trans. Autom. Control, vol. 44, no. 11, pp. 1995-2005, 1999.

[3] M. Forti, P. Nistri and M. Quincampoix, "Generalized neural network for nonsmooth nonlinear programming problems", IEEE Trans. Neural Netw., vol. 51, no. 9, pp. 1741-1754, 2004.

[4] M. Forti, P. Nistri and M. Quincampoix, "Convergence of neural networks for programming problems via a nonsmooth Łojasiewicz inequality", IEEE Trans. Neural Netw., vol. 17, no. 6, pp. 1471-1486, 2006.

[5] Y.S. Xia, G. Feng and J. Wang, "A novel recurrent neural network for solving nonlinear optimization problems with inequality constraints", IEEE Trans. Neural Netw., vol. 19, no. 8, pp. 1340-1353, 2008.

[6] J.L. Liang, Z.D. Wang, X.H. Liu, "State estimation for coupled uncertain stochastic networks with missing measurements and time-varying delays: the discrete-time case", IEEE Trans. Neural Netw., vol. 20, no. 5, pp. 781-793, 2009.

[7] W. Bian and X.P. Xue, "Subgradient-based neural network for nonsmooth nonconvex optimization problem", IEEE Trans. Neural Netw., vol. 20, no. 6, pp.1024-1038, 2009.

[8] Q.S. Liu and J. Wang, "Finite-time convergent recurrent neural network with a hard-limiting activation function for constrained optimization with piecewise-linear objective functions", IEEE Trans. Neural Netw., vol. 22, no. 4, pp. 601-613, 2011.

[9] Q.S. Liu and J. Wang, "A one layer recurrent neural network for constrained nonsmooth optimization", IEEE Trans. Syst. Man Cybern. Part B-Cybern., vol. 41, no. 5, pp. 1323-1333, 2011.

[10] L. Cheng, Z.G. Hou, Y.Z. Lin, M. Tan, W.C. Zhang and F.X. Wu, "Recurrent neural network for non-smooth convex optimization problems with application to identification of genetic regulatory networks", IEEE Trans. Neural Netw., vol. 22, no. 5, pp. 714-726, 2011.

[11] Q. Liu, Z. Guo and J. Wang, "A one-layer recurrent neural network for constrained pseudoconvex optimization and its application for portfolio optimization", Neural Netw., vol. 26, pp. 99-109, 2012.

[12] W. Bian and X. Chen, "Smoothing neural network for constrained non-Lipschitz optimization with applications", IEEE Trans. Neural Netw. Learn. Syst., vol. 23, no. 3, pp. 399-411, 2012.

[13] T. Yamada and T. Yabuta, "Dynamic system identification using neural networks", IEEE Trans. Syst. Man Cybern. Part B-Cybern., vol. 23, no. 1, pp. 204-211, 1993.

[14] W.L. Lu and J. Wang, "Convergence analysis of a class of nonsmooth gradient systems", IEEE Trans. Circuits Syst. I, vol. 55, no. 11, pp. 3514-3527, 2008.

[15] H.D. Qi and L. Qi, "Deriving sufficient conditions for global asymptotic stability of delayed neural networks via nonsmooth analysis", IEEE Trans. Circuits Syst. I, vol. 15, no.1, pp. 99-109, 2004.

[16] M. Forti, "$M$-matrices and global convergence of discontinuous neural networks", Int. J. Circuit Theory Applicat., vol. 35, no. 2, pp. 105-130, 2007.

[17] M. Di Marco, M. Forti, M. Grazzini, P. Nistri and L. Pancioni, "Lyapunov method and convergence of the full-range model of CNNs", IEEE Trans. Circuits Syst. I, vol. 55, no. 11, pp. 3528-3541, 2008.

[18] R. Tibshirani, "Regression shrinkage and selection via the lasso", J. R. Stat. Soc. Ser. B, vol. 58, no. 1, pp. 267-288, 1996.

[19] L.B. Montefusco, D. Lazzaro and S. Papi, "Nonlinear filtering for sparse signal recovery from incomplete measurements", IEEE Trans. Signal Process., vol. 57, no. 7, pp. 2494-2502, 2009.

[20] R. Chartrand and V. Staneva, "Restricted isometry properties and nonconvex compressive sensing", Inverse Probl., vol. 24, no. 3, pp. 1-14, 2008.

[21] W. Bian and X. Chen, "Worst-case complexity of smoothing quadratic regularization methods for non-Lipschitzian optimization", SIAM J. Optim., accepted, 2013.

[22] W. Bian, X. Chen and Y. Ye, "Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization", preprint, 2013.

[23] X. Chen, R. Womersley and J. Ye, "Minimizing the condition number of a Gram matrix", SIAM J. Optim., vol. 21, pp. 127-148, 2011.

[24] C. Zhang and X. Chen, "Smoothing projected gradient method and its application to stochastic linear complementarity problems", SIAM J. Optim., vol. 20, pp. 627-649, 2009.

[25] F.E. Curtis and M.L. Overton, "A sequential quadratic programming algorithm for nonconvex, nonsmooth constrained optimization", SIAM J. Optim., vol. 22, pp. 474-500, 2012.

[26] J. Kreimer and R.Y. Rubinstein, "Nondifferentiable optimization via smooth approximation: general analytical approach", Ann. Oper. Res., vol. 39, pp. 97-119, 1992.

[27] I. Necoara and J.A.K. Suykens, "Application of a smoothing technique to decomposition in convex optimization", IEEE Trans. Autom. Control, vol. 53, no. 11, pp. 2674-2679, 2008.

[28] X. Chen, "Smoothing methods for nonsmooth, nonconvex minimization", Math. Program., vol. 134, pp. 71-99, 2012.

[29] X. Chen, "Smoothing methods for complementarity problems and their applications: a survey", J. Oper. Res. Soc. Jpn., vol. 43, pp. 32-47, 2000.

[30] R.T. Rockafellar and R.J-B Wets, *Variational Analysis*, Springer, Berlin, 1998.

[31] F.H. Clarke, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[32] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer, New York, 1999.

[33] L.O. Chua, C.A. Desoer and E.S. Kuh, *Linear and Nonlinear Circuits,* New York: McGraw-Hill, 1987.

[34] J. Penot and P. Quang, "Generalized convexity of functions and generalized monotonicity of set-valued maps", J. Optim. Theory Appl., vol. 92, no. 2, pp. 343-356, 1997.

[35] M. Gurbuzbalaban and M.L. Overton, "On Nesterov's nonsmooth Chebyshev-Rosenbrock functions", Nonlinear Anal. -Theory Methods Appl., vol. 75, no. 3, pp. 1282-1289, 2012.

[36] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning Data Mining, Inference, and Prediction*, Springer-Verlag, New York, 2009.

[37] D. Betounes, *Differential Equations: Theory and Applications*, Springer, 2009.


**Wei Bian received the B.S. and Ph.D. degrees at Mathematics from Harbin Institute of Technology in 2004 and 2009, respectively. From 2009, she has been a lecturer at the Department of Mathematics, Harbin Institute of Technology. Meantime, she was a post-doctor fellow at the Department of Applied Mathematics, the Hong Kong Polytechnic University from July 2010 to August 2012. Her main scientific interest is in the field of neural network theory and optimization methods.**

**Xiaojun Chen is a Chair Professor at the Department of Applied Mathematics, The Hong Kong Polytechnic University. Previously, she was a Professor at the Department of Mathematical Sciences, Hirosaki University, Japan. Her current research interests include nonsmooth, nonconvex optimization, stochastic variational inequalities and approximations on the sphere. She serves in the editorial boards of six applied mathematical journals including SIAM Journal on Numerical Analysis.**