EasyChair Preprint

№ 5224

Neural Network for A Class of Sparse Optimization in Machine Learning Problems

Qingfa Li, Sitian Qin and Wei Bian

March 29, 2021

Neural Network for A Class of Sparse Optimization in Machine Learning Problems

Qingfa Li
Department of Mathematics
Heilongjiang Institute of Technology
Harbin, China
liqingfa_hrb@163.com

Sitian Qin
Department of Mathematics
Harbin Institute of Technology
Weihai, China
qinsitian@163.com

Wei Bian
Department of Mathematics
Harbin Institute of Technology
Harbin, China
bianweilvse520@163.com

***Abstract*—Sparse optimization involving the ℓ_0-norm function in objective function has a wide application in machine learning problems. In this paper, we propose a projected neural network modeled by a differential equation to solve a class of these optimization problems, in which the objective function is the sum of a nonsmooth convex loss function and the regularization defined by the ℓ_0-norm function. This optimization problem is not only nonconvex, but also discontinuous. To simplify the structure of the proposed network and let it own better convergence properties, we use the smoothing method, where the new constructed smoothing function for the regularization term plays a key role. We prove that the solution to the proposed network is globally existent and unique, and any accumulation point of it is a critical point of the continuous relaxation model. Except for a special case, which can be easily justified, any critical point is a local minimizer of the considered sparse optimization problem. It is an interesting thing that all critical points own a promising lower bound property, which is satisfied by all global minimizers of the considered problem, but is not by all local minimizers. Finally, we use some numerical experiments to illustrate the efficiency and good performance of the proposed method for solving this class of sparse optimization problems, which include the most widely used models in feature selection of classification learning.**

***Index Terms*—machine learning, projected neural network, sparse optimization, convergence analysis, critical point.**

I. Introduction

Sparse optimization, which aims to find a solution with most elements of zero and satisfying a system as much as possible, has many applications in various applications, in particular the machine learning, feature selection, image proceeding and finance [1]–[7]. Formally, a class of sparse optimization problems take the form of

$$\begin{aligned} \min \quad & f(x) := l(x) + \lambda \|x\|_0 \\ \text{s.t.} \quad & x \in \mathcal{X}, \end{aligned} \tag{1}$$

where λ is a given positive parameter, $\|x\|_0$ is the ℓ_0 norm function defined by the number of nonzero elements of x, $l : \mathbb{R}^n \to \mathbb{R}$ is the loss function to characterize the data fitting for the system. In this paper, we assume that l is a continuous convex function and $\mathcal{X}$ is a closed convex set of $\mathbb{R}^n$ defined by $\mathcal{X} = \{x : b \le x \le u\}$ with b, $u \in \overline{\mathbb{R}^n}$ and $b \le \mathbf{0} \le u$.

This work was funded by the NSF foundation (11871178,61773136) of China.

The aim of problem (1) is to find a sparse solution in $\mathcal{X}$ which minimizes function l as much as possible. Here, we call a vector sparse, if most of its elements are 0. λ is the parameter to control the tradeoff between the requirement on minimizing l and sparsity. Note that the objective function in (1) is discontinuous. Though the ℓ_0 function is the most desirable function to describe the sparsity, finding a global minimizer of problems with cardinality function is strongly NP-hard in general [8].

Due to the discontinuity of $\|x\|_0$, continuous relaxation is an important method to handle this kind of optimization problems. In the machine learning community, the ℓ_1 function is one of the most popular one, which is also known as Lasso. Since ℓ_1 is a convex function, there are rich algorithms for solving the ℓ_1 regularized optimization problems. The equivalence of the $\ell_2 - \ell_0$ and $\ell_2 - \ell_1$ problem in the sense of global minimizers was first established in [9], [10] and then was improved from many aspects. However, the ℓ_1 relaxation often leads to a biased estimator [11]. Then, some continuous but nonconvex relaxation functions were proposed for the ℓ_0 function, such as the hard thresholding penalty [12], log-sum penalty [13], bridge ℓ_p ($0 < p < 1$) penalty [14], [15], SCAD [11], capped-ℓ_1 penalty [16], [17], MCP [18], etc. Almost at the same time, different algorithms were developed for solving these nonconvex relaxation problems [19]–[22]. The solutions based on these nonconvex relaxations often bring better estimators, which not only have good sparsity but also reduce the deviation on the nonzero elements. However, the literatures on the equivalence between these nonconvex relaxation models and the considered ℓ_0 regularized problem are very few. The authors in [23] gave a CEL0 relaxation model and established its equivalence to the unconstrained $\ell_2 - \ell_0$ problem. In [17], the authors proved that optimization problem (1) is equivalent to a Lipschitz continuous problem with capped-ℓ_1 regularization in the sense of global minimizers. In this paper, we will propose a neural network method to solve (1) based on the capped-ℓ_1 regularized problem, which is an exact continuous nonconvex relaxation model of (1) [17]. Developing new models and methods for sparse optimization problems is always an interesting topic for researchers in optimization and machine learning.

Artificial neural network as a real-time method has a

promising application in optimization. Tank and Hopfield brought forward a neural network to solve a linear programming [24]. This work is an pioneering work on solving optimization by neural network method. From then on, abundant neural network models emerged, such as the network in [25] for nonlinear programming, [26], [27] for nonsmooth convex problems, [28]–[30] for nonsmooth nonconvex problems, [31], [32] for non-Lipschitz problems. Comparing with the iterative methods, the dynamic method on neural network has many advantages. First of all, we do not need to carry about the step size for convergence. Next, neural network can be implemented physically based on circuits, and then it can be run fast at the order of magnitude [26], [33]. Inspired by these reasons, we focus on solving (1) by neural network method.

Most neural network models for solving optimization problems need the regularity on the functions, which is a key condition in convergence analysis. However, we note that ℓ_0 function is discontinuous and the continuous relaxation given in [17] is not regular, which is one of the main difficulties in solving (1) by neural networks. To overcome it, we introduce the smoothing method in this paper. Smoothing methods for solving nonsmooth optimization problems have been used for many decades [34]. The main advantage of smoothing methods is that we can solve the considered nonsmooth optimization problem by a sequence of optimization problems with continuously differentiable objective functions. Another advantage of smoothing method in neural network research for optimization is that we can use the gradient of the smooth function instead of the subgradient of the original nonsmooth function, which promotes the network to be modeled by a differential equation, but no longer a differential inclusion.

Based on the above review and analysis, the main contributions of this paper are as follows. First, we consider a widely used sparse optimization problem and propose a method based on popular neural network model for solving it. To the best of our knowledge, to solve the cardinality regularized sparse optimization problem by neural network was only considered in [35]. However, the model considered in [35] is a special case of (1). On the one hand, loss function l in [35] is smooth, while it can be nonsmooth in this paper. On the other hand, the constraint in [35] is a box in $\mathbb{R}^n_+$, while it can be any box in $\mathbb{R}^n$ in this paper. Though the authors in [35] shew an extension to the box constraint in $\mathbb{R}^n$, the dimension of proposed network will increase to $2n$. And we would like to emphasize that the dimension of proposed network in this paper is just n, which is the same as the dimension of variable in (1). Second, we construct a new smoothing function for the capped-ℓ_1 function, which is totally different from the before ones and owns the necessary properties for the following convergence analysis. Third, thanks to the projection operator and the constructed smoothing function, we propose a projected neural network modeled by a differential equation to solve (1). We prove that any accumulation point of the solution to the proposed network is a critical point of the considered continuous relaxation of problem (1). Moreover, it is a good news that any accumulation point of the solution to the proposed network satisfies a promising lower bound property, which is a necessary condition to the global minimizers of (1), but not to the local minimizers. Though the direct target of the proposed network is to solve its continuous relaxation, we can easily justify whether a critical point of the continuous relaxation problem is a local minimizer of (1) or not.

The remaining parts of this paper are organized as follows. In section II, we introduce some necessary basic results used in this paper. A smoothing function for the capped-ℓ_1 function is constructed in the first part of section III. Then, the proposed neural network model is presented in section III. The convergence analysis and the optimal properties of the network to problem (1) are given in section IV. Some numerical experiments are illustrated in section V to show the effectiveness and good performance of the proposed neural network for solving sparse optimization problem (1). Section VI gives a brief summary of this paper. Section VII is the appendix part, which is used to show all proofs of the results in section IV.

Notation: Denote $\mathbb{R}^n$ the n-dimensional real-valued vector space, $\overline{\mathbb{R}^n} = [-\infty, +\infty]^n$, $\mathbb{R}^n_+ = [0, +\infty)^n$, $\mathbb{R}^n_- = (-\infty, 0]^n$. For $x, y \in \mathbb{R}^n$, $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$, $\|x\| := \|x\|_2 = \left(\sum_{i=1}^n x_i^2\right)^{\frac{1}{2}}$, and $\|x\|_1 = \sum_{i=1}^n |x_i|$. For a closed subset $\Omega \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$, $\mathrm{dist}(x, \Omega) = \inf_{s\in\Omega} \|x - s\|$, $\mathrm{int}(\Omega)$ and $\mathrm{bd}(\Omega)$ mean the interior and boundary of Ω in $\mathbb{R}^n$, respectively. For $x \in \mathbb{R}^n$ and δ, $\mathcal{A}(x) = \{i : x_i \neq 0\}$ and $B(x, \delta)$ indicates the open ball in $\mathbb{R}^n$ centered at x with radius δ. $e_i \in \mathbb{R}^n$ is the ith column of n-dimensional identity matrix.

II. Preliminary Results

In this section, we will first give some necessary basic definitions and results. Then, the exact continuous relaxation to (1) given in [17] is introduced in section II-B.

A. Definitions and properties

For a nonempty, closed and convex set $\Omega \subseteq \mathbb{R}^n$, the projection to Ω at x is well-defined, i.e.

$$P_\Omega(x) = \arg\min_{z\in\Omega} \|z - x\|.$$

And it owns the following properties

$$\langle u - P_\Omega(u), P_\Omega(u) - w\rangle \geq 0, \qquad \forall u \in \mathbb{R}^n, w \in \Omega; \quad (2)$$

$$\|P_\Omega(u) - P_\Omega(w)\| \leq \|u - w\|, \qquad \forall u, w \in \mathbb{R}^n. \quad (3)$$

Since $\mathcal{X}$ in (1) is defined by a simple box constraint, $P_\mathcal{X}$ has a closed-form solution for any $x \in \mathbb{R}^n$, where

$$[P_\mathcal{X}(x)]_i = \max\{b_i, \min\{x_i, u_i\}\},\ i = 1, 2, \ldots, n.$$

For a function $g : \mathbb{R}^n \to \mathbb{R}$, we call g is locally Lipschitz continuous, if for any x, there exist $\delta > 0$ and $L_x > 0$ such that

$$|g(y) - g(z)| \leq L_x \|y - z\|, \quad \forall y, z \in B(x, \delta).$$

Definition 2.1: For a locally Lipschitz continuous function $g:\mathbb{R}^n\to\mathbb{R}$ and $x\in\mathbb{R}^n$, the generalized directional derivative of g at x in direction $v\in\mathbb{R}^n$ is defined by

$$g^0(x;v)=\limsup_{y\to x;t\to 0^+}\frac{g(y+tv)-g(y)}{t}.$$

Then, the Clarke's generalized gradient of g at x is provided by

$$\partial g(x)=\{\xi\in\mathbb{R}^n:\ g^0(x;v)\geq\langle v,\xi\rangle,\ for\ all\ v\in\mathbb{R}^n\}.$$

In particular, if g is convex on convex set $\Omega\subseteq\mathbb{R}^n$, then it holds

$$g(y)-g(x)\geq\langle\xi,y-x\rangle,\quad\forall\xi\in\partial g(x),\ y\in\Omega.$$

Definition 2.2: [36] For nonempty closed convex subset $\Omega\subseteq\mathbb{R}^n$ and $x\in\Omega$, the normal cone to Ω at x is

$$N_\Omega(x)=\{\eta\in\mathbb{R}^n:\langle\eta,y-x\rangle\geq 0,\ \forall y\in\Omega\}.$$

Denote $\mathcal{F}:\mathbb{R}^n\times\mathbb{R}_+\to\mathbb{R}^n$ a continuous function. For a non-autonomous real-time differential equation system:

$$\dot{x}(t)=\mathcal{F}(x(t),t),\tag{4}$$

we call $x:[0,T)\to\mathbb{R}^n$ with $T>0$ a solution of (4) with initial point x_0, if x is absolutely continuous on $[0,T)$ and satisfies (4) for almost all $t\in[0,T)$. Moreover, the following chain rule is often used in the dynamic analysis of (4).

Proposition 2.1: [36] Suppose $g:\mathbb{R}^n\to\mathbb{R}$ is regular[1] and $x:[0,T)\to\mathbb{R}^n$ with $T>0$ is absolutely continuous on any compact set of $[0,T)$, then $g(x(t))$ is differentiable for almost all $t\in[0,T)$ and

$$\frac{d}{dt}g(x(t))=\langle\xi,\dot{x}(t)\rangle,\quad\forall\xi\in\partial g(x(t)).$$

B. Exact continuous relaxation to (1)

To solve (1), the authors in [17] introduced the following Lipschitz continuous optimization model:

$$\begin{aligned}\min\quad & f_r(x):=l(x)+\lambda p(x)\\ \text{s.t.}\quad & x\in\mathcal{X},\end{aligned}\tag{5}$$

where

$$p(x)=\sum_{i=1}^n\min\left\{\frac{1}{\nu}|x_i|,1\right\}$$

with $\nu>0$.

Function p can be formulated by a DC (difference-of-convex) function [17], i.e.

$$p(x)=\sum_{i=1}^n\frac{1}{\nu}|x_i|-\sum_{i=1}^n\max\{\theta_1(x_i),\theta_2(x_i),\theta_3(x_i)\},$$

where $\theta_1(t)=\frac{1}{\nu}t-1$, $\theta_2(t)=-\frac{1}{\nu}t-1$ and $\theta_3(t)=0$.

[1]A Lispchitz continuous function $g:\mathbb{R}^n\to\mathbb{R}$ is said to be regular at x provided the following conditions holds:

(i) for all $v\in\mathbb{R}^n$, the usual one-sided directional derivative $g'(x;v)=\lim_{t\to 0+}\frac{g(x+tv)-g(x)}{t}$ exists;

(ii) for all $v\in\mathbb{R}^n$, $g'(x;v)=g^0(x;v)$.

Though (5) is a nonsmooth nonconvex programming, its local minimizers have a necessary and sufficient condition, which can be defined by the conditions for d-stationary point of it. We call $x^*\in\mathcal{X}$ is a d-stationary point of (5) if

$$\xi_i\in[\partial l(x^*)+N_{\mathcal{X}}(x^*)]_i+\frac{\lambda}{\nu}\partial|x_i^*|,$$

for any $i=1,2,\ldots,n$ and ξ_i satisfying

$$\xi_i\in\begin{cases}[-\frac{\lambda}{\nu},0]\text{sign}(x_i) & \text{if } |x_i|=\nu\\ \{-\text{sign}(x_i)\} & \text{if } |x_i|>\nu\\ \{0\} & \text{if } |x_i|<\nu.\end{cases}$$

The exactness of problem (1) to (5) in the sense of global minimizers is built up in [17].

Proposition 2.2: [17] Suppose

$$\nu<\min\{\frac{\lambda}{L_l},|b_i|,|u_j|:i,j=1,2,\ldots,n,b_i\neq 0,u_j\neq 0\},$$

where L_l is a Lipschitz constant of l on $\mathcal{X}$. Then, $x^*\in\mathcal{X}$ is a global minimizer of (1) if and only if it is a global minimizer of (5) and x^* owns a lower bound property

$$\text{if } |x_i^*|\leq\nu,\text{ then } x_i^*=0.\tag{6}$$

Moreover, if $\hat{x}$ is a d-stationary point of problem (5), then $\hat{x}$ is a local minimizer of problem (1).

How to provide other sufficient condition to ensure the exactness of problem (5) to problem (1) in the sense of global minimizers is an interesting topic. Since the objective function in problem (5) is nonsmooth and nonconvex, we focus on the critical points of it.

Definition 2.3: [37] We call $x^*\in\mathcal{X}$ a critical point of problem (5), if

$$\mathbf{0}\in\partial l(x^*)+\lambda\partial p(x^*)+N_{\mathcal{X}}(x^*).\tag{7}$$

Though finding the global minimizers of problem (1) is strongly NP-hard in general, its local minimizers have a sufficient and necessary condition.

Proposition 2.3: $x^*\in\mathcal{X}$ is a local minimizer of problem (1) if and only if it satisfies

$$0\in[\partial l(x^*)+N_{\mathcal{X}}(x^*)]_{\mathcal{A}(x^*)},\tag{8}$$

where $\mathcal{A}(x^*)=\{i:x_i^*\neq 0\}$.

Thanks to the convexity of l, $x^*\in\mathcal{X}$ satisfies (8) if and only if x^* is a local minimizer of l on $\mathcal{X}_{\mathcal{A}(x^*)}$. For $x^*\in\mathcal{X}$, if $|x_i^*|\neq\nu$, $\forall i$, we can easily find that x^* is a critical point of (5) if and only if x^* is a local minimizer of problem (1). Thus, finding the critical points of problem (5) is an interesting work and of importance to the solving of problem (1).

III. Proposed neural network

To propose a network with better convergence properties, we will first construct a new smoothing function for the given continuous relaxation p in (5), where some properties of the smoothing function are also analyzed in section III-A. Based on the relationships between (1) and (5) shown in section

II-B and the constructed smoothing function in section III-A, we will propose a neural network modeled by a differential equation to solve (5) in section III-B.

A. Smoothing approximations

Since convex function l in (5) can be nonsmooth, to simplify the model of proposed network from differential inclusion to differential equation, we introduce a smoothing function of it defined as follows.

Definition 3.1: [17] For convex function l in (1), we call $\tilde{l}:\mathbb{R}^n\times(0,1]\to\mathbb{R}$ a smoothing function of it, if $\tilde{l}$ satisfies the following conditions:

(i) for any fixed $\mu\in(0,1]$, $\tilde{l}(\cdot,\mu)$ is continuously differentiable on $\mathcal{X}$;
(ii) for any fixed $\mu\in(0,1]$, $\tilde{l}(\cdot,\mu)$ is also convex on $\mathcal{X}$;
(iii) for any $x\in\mathcal{X}$, $\{\lim_{z\to x,\mu\downarrow 0}\nabla_z\tilde{l}(z,\mu)\}\subseteq\partial l(x)$;
(iv) for any fixed $x\in\mathcal{X}$, $\tilde{l}(x,\cdot)$ is differentiable on $(0,1]$ and there exists a positive constant κ such that

$$|\tilde{l}(x,\mu_2)-\tilde{l}(x,\mu_1)|\le\kappa|\mu_1-\mu_2|,\ \forall\,\mu_1,\mu_2\in(0,1].$$

Item (iv) in Definition 3.1 implies that for any $x\in\mathcal{X}$ and $\mu\in(0,1]$, it holds

$$|\nabla_\mu\tilde{l}(x,\mu)|\le\kappa\quad\text{and}\quad|\tilde{l}(x,\mu)-l(x)|\le\kappa\mu, \tag{9}$$

then

$$\lim_{\mu\downarrow 0}\tilde{l}(x,\mu)=l(x),\quad\forall x\in\mathcal{X}. \tag{10}$$

How to construct a smoothing function for different l satisfying the conditions in Definition 3.1 can be consulted to [31], [34], [38], [39]. For example, if $l(x)=\|Ax-b\|_1$ with $A\in\mathbb{R}^{m\times n}$ and $b\in\mathbb{R}^m$, we can set

$$\tilde{l}(x,\mu)=\sum_{i=1}^m\psi(A_ix-b_i,\mu) \tag{11}$$

with

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

Moreover, if $l(x)=\max\{Ax-b,\mathbf{0}\}$ with $A\in\mathbb{R}^{m\times n}$ and $b\in\mathbb{R}^m$, we can let

$$\tilde{l}(x,\mu)=\sum_{i=1}^m\varphi(A_ix-b_i,\mu) \tag{12}$$

with

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

Note that function p in (5) is also nonsmooth. Though we can give a continuously differentiable function to approximate p based on some existing smoothing functions, it brings us some difficulties in analyzing the convergence of the proposed network. For example, the monotone decreasing with respect to the smoothing parameter is often not satisfied. So, we will propose a continuously differentiable function to approximate p not only satisfying conditions (i), (iii), (iv) in Definition 3.1, but also owning some other properties, which will be much helpful for the further convergence analysis. Set

$$\tilde{p}(x,\mu)=\sum_{i=1}^n\theta(x_i,\mu), \tag{13}$$

where

$$\theta(s,\mu)=\begin{cases}\dfrac{1}{2\nu a\mu}s^2+\dfrac{a}{2\nu}\mu & \text{if } |s|\le a\mu\\ \dfrac{1}{\nu}|s| & \text{if } a\mu<|s|\le\nu\\ -\dfrac{(|s|-\nu-\mu)^2}{2\nu\mu}+1+\dfrac{1}{2\nu}\mu & \text{if } \nu<|s|\le\nu+\mu\\ 1+\dfrac{1}{2\nu}\mu & \text{if } |s|>\nu+\mu.\end{cases} \tag{14}$$

with fixed positive parameters a and μ satisfying $a\mu<\nu$. For any fixed $x\in\mathbb{R}^n$, we can easily verify that

$$\lim_{\mu\downarrow 0}\tilde{p}(x,\mu)=p(x). \tag{15}$$

By simple calculation, the first and second derivative of $\theta(s,\mu)$ with respect to s for any fixed $\mu\in\mathbb{R}_{++}$ can be expressed by

$$\theta'_s(s,\mu)=\begin{cases}\dfrac{1}{\nu a\mu}s & \text{if } |s|\le a\mu\\ \dfrac{1}{\nu}\mathrm{sign}(s) & \text{if } a\mu<|s|\le\nu\\ \dfrac{(\nu+\mu-|s|)\mathrm{sign}(s)}{\nu\mu} & \text{if } \nu<|s|\le\nu+\mu\\ 0 & \text{if } |s|>\nu+\mu\end{cases} \tag{16}$$

and

$$\theta''_s(s,\mu)=\begin{cases}\dfrac{1}{\nu a\mu} & \text{if } |s|<a\mu\\ 0 & \text{if } a\mu<|s|<\nu\\ -\dfrac{1}{\nu\mu} & \text{if } \nu<|s|<\nu+\mu\\ 0 & \text{if } |s|>\nu+\mu,\end{cases}$$

which implies that $\theta(\cdot,\mu)$ is Lipschitz continuously differentiable for any $\mu\in\mathbb{R}_{++}$. Thus, $p(\cdot,\mu)$ is also Lipschitz continuously differentiable for any $\mu\in\mathbb{R}_{++}$ and

$$\nabla_x p(x,\mu)=\sum_{i=1}^n\theta'(x_i,\mu)e_i.$$

For the gradient consistence, for any $x\in\mathcal{X}$, we also obtain

$$\{\lim_{z\to x,\mu\downarrow 0}\nabla_z\tilde{p}(z,\mu)\}\subseteq\partial p(x). \tag{17}$$

In a similar way, for any fixed $x\in\mathbb{R}^n$, we obtain

$$\theta'_\mu(s,\mu)=\begin{cases}\dfrac{a}{2\nu}-\dfrac{1}{2\nu a\mu^2}s^2 & \text{if } |s|\le a\mu\\ 0 & \text{if } a\mu<|s|\le\nu\\ \dfrac{\mu^2-(\mu+\nu-|s|)(\mu-\nu+|s|)}{2\nu\mu^2} & \text{if } \nu<|s|\le\nu+\mu\\ \dfrac{1}{2\nu} & \text{if } |s|>\nu+\mu,\end{cases}$$

which shows the continuously differentiability of $p(x,\cdot)$ for fixed $x \in \mathbb{R}^n$. Moreover, we find that

$$\theta'_\mu(s,\mu) \geq 0, \quad \forall s \in \mathbb{R},\, \mu \in (0,1], \tag{18}$$

which implies that $\tilde{p}(x,\cdot)$ is non-decreasing on $(0,1]$.

B. Neural network model

Inspired by the previous analysis, we propose a neural network modeled by a differential equation as follows

$$\begin{cases} \dot{x}(t) = -x(t) + P_{\mathcal{X}}\left[x(t) - \left(\nabla_x \tilde{l}(x(t),\mu(t)) + \lambda \nabla_x p(x(t),\mu(t))\right)\right] \\ x(0) = x_0, \end{cases} \tag{19}$$

where $\mu : [0,+\infty) \to (0,1]$ is a monotone decreasing function and converges to 0 as t tends to $+\infty$. For example, we can set it with the formulation

$$\mu(t) = (t+1)^{-\sigma} \quad \text{or} \quad \mu(t) = e^{-\sigma t} \tag{20}$$

with $\sigma > 0$. We refer that $\mu(t)$ in (20) can also be implemented by circuits. For example, the first formulation in (20) is the solution of the following system

$$\begin{cases} \dot{\mu}(t) = -\sigma^{-1}(t+1)^{-\sigma-1}, \\ \mu(0) = 1. \end{cases}$$

IV. Convergence analysis of network (19)

In this section, we will give some theoretical analysis on network (19) including the global existence and uniqueness of its solution, and then the convergence of the proposed network for solving (5) is analyzed. For readability, we put the proof of all results in this section to Appendix part.

Theorem 4.1: For any initial point $x_0 \in \mathcal{X}$, there exists a global solution to network (19). And any solution $x : [0,+\infty) \to \mathbb{R}^n$ to network (19) satisfies $x(t) \in \mathcal{X}$, $\forall t \in [0,+\infty)$. Moreover, if $\nabla \tilde{l}(\cdot,\mu)$ is locally Lipschitz continuous on $\mathcal{X}$ for any fixed $\mu \in (0,1]$, the solution to (19) with initial point $x_0 \in \mathcal{X}$ is unique.

Next proposition shows a basic convergence of the solution to network (19).

Proposition 4.1: Denote $x : [0,+\infty) \to \mathbb{R}^n$ the solution of (19) with initial point $x_0 \in \mathcal{X}$. Then, it holds

(i) $\lim_{t\to+\infty} f_r(x(t)) = \lim_{t\to+\infty} f(x(t))$ exists;
(ii) $\int_0^{+\infty} \|\dot{x}(t)\|^2 dt < +\infty$.

It is a good news that the critical points of problem (5) satisfy a similar lower bound property as in (6).

Proposition 4.2: If $\bar{x} \in \mathcal{X}$ is a critical point of problem (5), then

$$\text{if } |\bar{x}_i| < \nu, \text{ then } \bar{x}_i = 0. \tag{21}$$

And $p(\bar{x}) = \|\bar{x}\|_0$.

It is an interesting thing the lower bound in Proposition 4.2 is a necessary optimality condition for the global minimizers but not for the local minimizers of (1), which is indicated in the following example.

Example 4.1: Consider

$$\min_{|x_1|\leq 1, |x_2|\leq 1} |x_1 + x_2 - 1| + \|x\|_0. \tag{22}$$

By simple calculation, we verify that the global optimal solution set of (22) is

$$\mathcal{M} = \{(0,1),(1,0),(0,0)\},$$

while the local optimal solution set is

$$\mathcal{LM} = \{x \in \mathbb{R}^2 : x_1 + x_2 - 1 = 0,\, -1 \leq x_1 \leq 1\} \bigcap \{(0,0)\}.$$

We see that all points in $\mathcal{M}$ satisfy (6) with $\nu < \frac{\sqrt{2}}{2}$, while the points in $\mathcal{LM}$ are not all. And the critical point set of corresponding problem (5) for (22) is

$$\mathcal{LM} \bigcap \{|x_1| \geq \nu,\, |x_2| \geq \nu\} \bigcup \{(0,\nu),(\nu,\nu),(\nu,0),(0,0)\},$$

which is a proper subset of $\mathcal{LM}$ as removing $\{(0,\nu),(\nu,\nu),(\nu,0)\}$.

To prove some further properties of network (19), we need to illustrate an important property of the solution to network (19) at first.

Proposition 4.3: Denote $x : [0,+\infty) \to \mathbb{R}^n$ the solution of (19) with initial point $x_0 \in \mathcal{X}$. Then, there exists a $T > 0$ such that once there exist $i \in \{1,2\ldots,n\}$ and $T_1 \geq T$ such that $|x_i(T_1)| \leq \nu$, then $\lim_{t\to+\infty} x_i(t) = 0$.

Theorem 4.2: Denote $x : [0,+\infty) \to \mathbb{R}^n$ the solution of (19) with initial point $x_0 \in \mathcal{X}$. Then, any accumulation point $x(t)$ is a critical point of problem (5). Moreover, if x^* and $\hat{x}$ are two accumulation points of $x(t)$ as $t \to +\infty$, then $\mathcal{A}(x^*) = \mathcal{A}(\hat{x})$.

V. Numerical experiments

In this section, we illustrate the performance of proposed network (19) for solving (1) by three examples. We use ode45 in Matlab 2016b on a Lenovo PC to run the codes.

Example 5.1: In this example, we test the effectiveness of the proposed network for a simple linear regression problem, a special case of which is used in Example 4.1

$$\min_{|x_1|\leq 1, |x_2|\leq 1} |x_1 + x_2 - 1| + \lambda \|x\|_0. \tag{23}$$

Since $L_l = \sqrt{2}$, by simple calculation, we can let $\nu = 0.7\lambda < \lambda\frac{\sqrt{2}}{2}$ in (5) when $0.7\lambda < 1$. Define the smoothing function of l in (23) as in (11). Set $a = 1$ in (13) and

$$\mu(t) = 10^{-3}/\sqrt{t+1}.$$

For different values of λ, the global minimizers of (23) and the corresponding ν satisfying the conditions in Proposition 2.2 are listed in Table I. For these different cases of λ, the limit points of the solution to (19) with different initial points are shown in Table II. From Table II, we see that though this class of problem is NP-hard in general, we can also find its global minimizers by network (19). Fix $\lambda = 1$ and $\lambda = 1.2$, the trajectories of network (19) with initial point $x^0 = (0.2, 0.8)^T$ are pictured in Fig. 1 and Fig. 2, respectively.

Example 5.2: Linear regression is one of the most well-known models in statistics and machine learning. Linear regression in machine learning is a supervised learning technique that comes from classical statistics. One method to

TABLE I
DIFFERENT VALUES OF λ AND ν FOR PROBLEM (23)

λ	global minimizers	ν
0.9	$(1,0)^T$, $(0,1)^T$	0.6
1	$(1,0)^T$, $(0,1)^T$, $(0,0)^T$	0.7
1.2	$(0,0)^T$	0.8

TABLE II
LIMIT POINTS OF SOLUTIONS TO NETWORK (19) WITH DIFFERENT λ AND INITIAL POINTS

λ	initial points	limit points
0.9	$(0.8,0.2)^T/(0.2,0.8)^T$	$(1,0)^T/(0,1)^T$
1	$(0.8,0.2)^T/(1,1)^T/(0.2,0.8)^T$	$(1,0)^T/(0,0)^T/(0,1)^T$
1.2	$(0.8,0.2)^T/(1,1)^T/(0.2,0.8)^T$	$(0,0)^T/(0,0)^T/(0,0)^T$

characterize the linear fitting is to minimize the ℓ_1 function. So, we consider the linear regression problem with ℓ_1 loss function and cardinality regularization in this example, which is modeled by

$$\min_{x\in\mathcal{X}} \|Ax-b\|_1 + \lambda\|x\|_0, \tag{24}$$

where $\mathcal{X} = \{x : -\mathbf{1} \le x \le \mathbf{1}\}$, $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$ are generated randomly by the following codes for given positive integers $(n,m,s) = (20,10,5)$:

```
I=randperm(n); I=I(1:s); x*=zeros(n,1);
B=randn(n,m); x*(I)=unifrnd(-1,1,[s,1]);
A=orth(B)'; b = A*x*.
```

We calculate L_l by $L_l = \|A\|_\infty$ and define $\nu = \min\{\lambda/(L_l+1), 1\}$. Set $\lambda = 0.5$. To show the efficiency, we use the mean square error (MSE) to evaluate it, where

$$\text{MSE}(x) = \frac{\|x-x^\star\|^2}{n}.$$

The MSE of the solution to (19) with a random initial point in $\mathcal{X}$ is pictured in Fig. 3.

Example 5.3: Feature selection is one of the popular problems in machine learning. The main goal is to select a subset of main features based on the given data which preserving the right ability of the classification. In this example, we test the prostate cancer data by network (19), in which

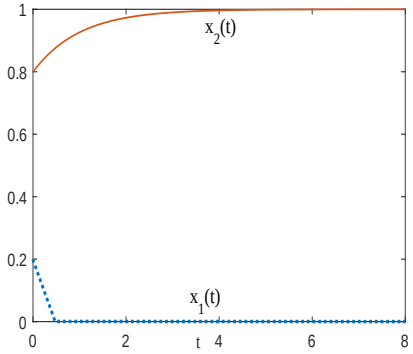

Fig. 1. Trajectory of (19) for (23) with $\lambda = 1$

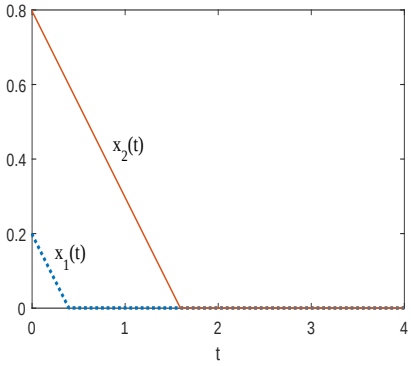

Fig. 2. Trajectory of (19) for (23) with $\lambda = 1.2$

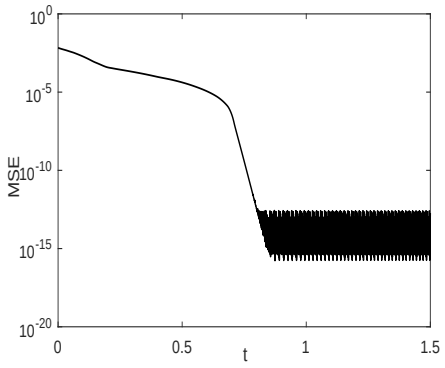

Fig. 3. MSE of trajectory to (19) for (24)

the goal is to find the main predictors as fewer as better while preserving the right justification. The date set is from https://web.stanford.edu/ hastie/ElemStatLearn/data.html. This set contains the medical records of 97 men who were about to receive a radical prostatectomy. There are eight clinical measures as the predictors, which are lcavol, lweight, age, lbph, svi, lcp, pleason and pgg45. To test the effect of the proposed method, we divide the set into two classes, in which one is the training set with 67 data and the other is the test set with 30 data. The model is as follows

$$\min_{\mathbf{0}\le x\le\mathbf{1}} \log(\|Ax-b\|^2+1) + 100\|x\|_0, \tag{25}$$

where $A \in \mathbb{R}^{67\times 8}$ and $b \in \mathbb{R}^{67}$ are defined by the codes in training set. Set $L_l = 2\|A^TA\|_\infty$ and then choose $\nu = 0.1814$. Define $\mu(t) = \frac{1}{10(t+1)}$.

We evaluate the performance by the prediction error, which is defined by the mean square error over the 30 data in test set. A smaller prediction error is better. The solution to network (19) with initial point $x^0 = (1,1,\ldots,1)^T$ is shown in Fig. 4. In addition, the convergence of smoothing function of objective function along the solution is pictured in Fig. 5, where

$$\tilde{f}_r(x,\mu) = \log(\|Ax-b\|^2+1) + 100p(x,\mu).$$

Meantime, the prediction error values along the solution are shown in Fig. 6. The results obtained are listed in Table III, where the best results of FOIPA [40], SSQP [41] and Lasso [42] are also listed. From the results in Table III, we see that neural network (19) can find the right 3 main predictors with the smallest prediction error.

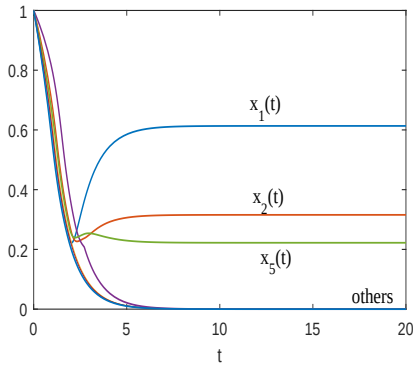

Fig. 4. Trajectory of (19) for Example 5.3

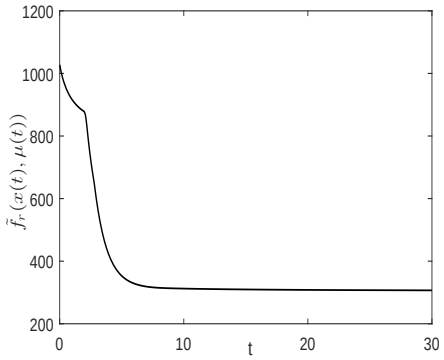

Fig. 5. Convergence on the objective function values

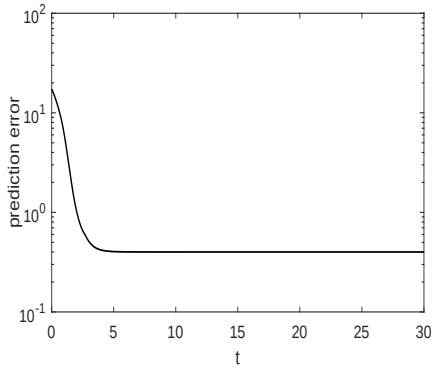

Fig. 6. Convergence on prediction error values

VI. Conclusions

In this paper, we considered the sparse optimization problem with a nonsmooth convex loss function, cardinality regularization and box constraints. Based on its continuous relaxation and a new given smoothing function, we proposed a neural network modeled by a differential equation to solve it. Thanks to the smoothing method, we proved that the solution to the proposed network is global existent and unique. Further, we proved that all accumulation points of the solution to proposed network are critical points of the used continuous relaxation problem, which are local minimizers of the considered sparse optimization problem except for some special and easily justified cases. Moreover, all accumulation points own a common support set and a desirable lower bound for the nonzero elements. Finally, numerical results shew the promising effect of the proposed network for solving the considered sparse optimization problems.

TABLE III
Numerical results of network (19) and some other methods

	(19)	FOIPA	SSQP	LASSO
$\bar{x}_1$	0.6134	0.6479	0.6437	0.533
$\bar{x}_2$	0.3145	0.2941	0.2765	0.169
$\bar{x}_3$	0	0	0	0
$\bar{x}_4$	0	0	0	0.002
$\bar{x}_5$	0.2222	0.1498	0.1327	0.094
$\bar{x}_6$	0	0	0	0
$\bar{x}_7$	0	0	0	0
$\bar{x}_8$	0	0	0	0
Prediction error	0.4002	0.4194	0.4262	0.479

VII. Appendix

A. Proof of Theorem 4.1

By the preliminary analysis in sections II-A and III-A, the right-hand function of (19) is continuous with respect to x and μ. Then, by [43, Theorem 1.1], network (19) exists at least one solution $x(\cdot)$ defined on $[0,T)$ with $T>0$. To prove its global existence, we argue it by contradiction and assume the maximal existence interval of it is $[0,T)$.

Reformulate (19) as

$$\dot{x}(t)+x(t)=\eta(x(t),\mu(t)), \tag{26}$$

where

$$\eta(x,\mu)=P_{\mathcal{X}}\left[x-\left(\nabla_x\tilde{l}(x,\mu)+\lambda\nabla_x p(x,\mu)\right)\right].$$

By (26), we have

$$x(t)=e^{-t}x_0+(1-e^{-t})\int_0^t\frac{e^s}{e^t-1}\eta(x(s),\mu(s))ds. \tag{27}$$

Owning to the continuity of function η, convexity of $\mathcal{X}$ and $\int_0^t\frac{e^s}{e^t-1}ds=1$, since $\eta(x(t),\mu(t))\in\mathcal{X}$, $\forall t\in[0,T)$, we obtain

$$\int_0^t\frac{e^s}{e^t-1}\eta(x(s),\mu(s))ds\in\mathcal{X},\ \forall t\in[0,T).$$

Using the convexity of $\mathcal{X}$ again, (27) implies

$$x(t)\in\mathcal{X},\quad \forall t\in[0,T).$$

Recalling the boundedness of $\mathcal{X}$, we get that x is bounded on $[0,T)$. This together with the structure of (19) implies the boundedness of $\dot{x}$ on $[0,T)$. Thanks to [43, Lemma 2.1], this solution x to (19) can be extended, which leads to a contradiction. Thus, there exists a global solution x to (19) with initial point $x_0 \in \mathcal{X}$, which is defined on $[0,+\infty)$.

Following a similar analysis, we derive that

$$x(t) \in \mathcal{X}, \quad \forall t \in [0,+\infty).$$

Next, we will argue the uniqueness of the solution to (19) by contradiction. Assume both x and y be two different solutions of (19) with the given initial point $x_0 \in \mathcal{X}$. Since x and y are absolutely continuous on $[0,+\infty)$, there exist $\bar{t} \in (0,+\infty)$ and $\tau > 0$ such that

$$x(t) \neq y(t), \quad \forall t \in [\bar{t}, \bar{t}+\tau].$$

When $\nabla\tilde{l}(\cdot,\mu)$ is locally Lipschitz continuous on $\mathcal{X}$ for any fixed $\mu > 0$, since $\mu(t) \geq \mu(\bar{t}+\tau) > 0$, $\forall t \in [\bar{t}, \bar{t}+\tau]$, there exists $L_1 > 0$ such that for all $t \in [\bar{t}, \bar{t}+\tau]$,

$$\|\nabla\tilde{l}(x(t),\mu(t)) - \nabla\tilde{l}(y(t),\mu(t))\| \leq L_1\|x(t)-y(t)\|.$$

Similarly, by (16), there exists $L_2 > 0$ such that for all $t \in [\bar{t}, \bar{t}+\tau]$,

$$\|\nabla p(x(t),\mu(t)) - \nabla p(y(t),\mu(t))\| \leq L_2\|x(t)-y(t)\|.$$

Combining the above results with the Lispchitz property of projection operator $P_{\mathcal{X}}$ given in (3), it implies the existence of $L > 0$ such that for all $t \in [\bar{t}, \bar{t}+\tau]$,

$$\|-x(t)+\eta(x(t),\mu(t))+y(t)-\eta(y(t),\mu(t))\| \leq L\|x(t)-y(t)\|.$$

Then, differentiating $\|x(t)-y(t)\|^2$ along the two solutions of (19) gives

$$\begin{aligned}\frac{d}{dt}\|x(t)-y(t)\|^2 =& 2\langle x(t)-y(t), \dot{x}(t)-\dot{y}(t)\rangle \\ \leq& 2L\|x(t)-y(t)\|^2.\end{aligned}$$

Integrating the above inequality from 0 to $t \in (0, \bar{t}+\tau]$ shows

$$\|x(t)-y(t)\|^2 \leq 2L\int_0^t \|x(s)-y(s)\|^2 ds.$$

Since $x(0) = y(0) = x_0$, using the Gronwall's inequality [36] to it gives

$$x(t) = y(t), \quad \forall t \in (0, \bar{t}+\tau],$$

which leads to a contraction to the hypothesis. Thus, the solution to network (19) with initial point $x_0 \in \mathcal{X}$ is unique.

B. Proof of Proposition 4.1

(i) Since $x(t) \in \mathcal{X}$, $\forall t \in [0,+\infty)$ and $\mathcal{X}$ is a closed convex set, by setting $w = x(t)$ and $u = x(t) - \alpha\left(\nabla_x\tilde{l}(x(t),\mu(t)) + \lambda\nabla_x p(x(t),\mu(t))\right)$ in (2) and from (19), we obtain

$$\langle \nabla_x\tilde{l}(x(t),\mu(t)) + \lambda\nabla_x p(x(t),\mu(t)), \dot{x}(t)\rangle \leq -\|\dot{x}(t)\|^2. \tag{28}$$

Then, by (9), (18) and (28), we have

$$\begin{aligned}&\frac{d}{dt}\left[\tilde{l}(x(t),\mu(t)) + \lambda\tilde{p}(x(t),\mu(t))\right] \\ =&\langle \nabla_x\tilde{l}(x(t),\mu(t)) + \lambda\nabla_x\tilde{p}(x(t),\mu(t)), \dot{x}(t)\rangle \\ &+ \left(\nabla_\mu\tilde{l}(x(t),\mu(t)) + \nabla_\mu\tilde{p}(x(t),\mu(t))\right)\dot{\mu}(t) \\ \leq& -\|\dot{x}(t)\|^2 - \kappa\dot{\mu}(t),\end{aligned} \tag{29}$$

where the last inequality uses the non-decreasing of $\tilde{p}(x,\cdot)$ with respect to μ and the non-increasing of $\mu(t)$ on $[0,+\infty)$.

Reformulating (29) gives

$$\frac{d}{dt}\left[\tilde{l}(x(t),\mu(t)) + \lambda\tilde{p}(x(t),\mu(t)) + \kappa\mu(t)\right] \leq -\|\dot{x}(t)\|^2. \tag{30}$$

Consequently, $\tilde{l}(x(t),\mu(t)) + \lambda\tilde{p}(x(t),\mu(t)) + \kappa\mu(t)$ is non-increasing on $[0,+\infty)$. Combining this with

$$\tilde{l}(x,\mu) + \lambda\tilde{p}(x,\mu) + \kappa\mu \geq l(x) + \lambda p(x) \geq \min_{\mathcal{X}} l \tag{31}$$

derived from (9) and (18), we find the existence of

$$\begin{aligned}&\lim_{t\to+\infty} f_r(x(t)) \\ =&\lim_{t\to+\infty}\left[\tilde{l}(x(t),\mu(t)) + \lambda\tilde{p}(x(t),\mu(t)) + \kappa\mu(t)\right] \\ =&\lim_{t\to+\infty}\left[l(x(t)) + \lambda p(x(t))\right] = \lim_{t\to+\infty} f(x(t)),\end{aligned}$$

where the first equality follows from (10) and (15).

(ii) Integrating (30) from 0 to $+\infty$ and by (31), we obtain the estimation in item (ii).

C. Proof of Proposition 4.2

Suppose $\bar{x}$ is a critical point of problem (5) and (21) does not hold for $\bar{x}$. Then there exists $\hat{i} \in \{1,2,\ldots,n\}$ such that $0 < |x_{\hat{i}}| < \nu$. From the definition of critical point, we have

$$0 \in [\partial l(\bar{x}) + N_{\mathcal{X}}(\bar{x})]_i + \frac{\lambda}{\nu}\mathrm{sign}(x_i^*),$$

which combining with $[N_{\mathcal{X}}(\bar{x})]_i = 0$ implies $\frac{\lambda}{\nu} \leq L_l$. This leads to a contradiction to $\nu < \frac{\lambda}{L_l}$ and then (21) holds. Thus,

$$p(\bar{x}) = \|\bar{x}\|_0.$$

D. Proof of Proposition 4.3

Set

$$\epsilon = \min\{\frac{\lambda - \nu L_l}{2\nu}, |b_i| - \nu, u_j - \nu, i,j = 1,\ldots,n,\, b_i u_j \neq 0\}.$$

Recalling $\lim_{t\to+\infty}\mu(t) = 0$ and Definition 3.1-(iii), there exists $T > 0$ such that $a\mu(t) < \nu/2$ and $\|\nabla_x\tilde{l}(x(t),\mu(t))\| < L_l + \epsilon$, $\forall t \geq T$, which indicates

$$\frac{\lambda}{\nu} > L_l + 2\epsilon \geq \|\nabla_x\tilde{l}(x(t),\mu(t))\| + \epsilon, \ \forall t \geq T.$$

Suppose there exists a $T_1 \geq T$ such that $0 < x_i(T) \leq \nu$. In view of (16), if $a\mu(t) \leq x_i(t) \leq \nu$, then

$$[\nabla_x\tilde{p}(x(t),\mu(t))]_i = \theta_s'(x_i(t_k),\mu(t_k)) = \frac{1}{\nu},$$

which implies

$$w_i(t) := \left[\nabla_x \tilde{l}(x(t),\mu(t)) + \lambda \nabla_x \tilde{p}(x(t),\mu(t))\right]_i > \epsilon.$$

Then,

$$\dot{x}_i(t) = -x_i(t) + P_{[b_i,u_i]}[x_i(t) - w_i(t)] \leq -\epsilon, \tag{32}$$

which indicates that $x_i(t) \leq x_i(T_1) - \epsilon(t - T_1)$ as long as $a\mu(t) < x_i(t) < \nu$ and $t \geq T_1$. Similarly, if there exists $T_1 \geq T$ such that $x_i(T_1) < 0$, then $x_i(t) \geq x_i(T_1) + \epsilon(t - T_1)$ as long as $-\nu < x_i(t) < -a\mu(t)$. Therefore, we can conclude that $\lim_{t\to+\infty} x_i(t) = 0$.

E. Proof of Theorem 4.2

Denote $\mathcal{C}$ the critical point set of problem (5), i.e.

$$\mathcal{C} = \{x \in \mathcal{X} : \mathbf{0} \in \partial l(x^*) + \lambda \partial p(x^*) + N_{\mathcal{X}}(x^*)\},$$

which is a closed set due to the upper semicontinuity of ∂l and ∂p. We will prove this result by contradiction.

If not, there exists a subsequence $\{t_k\}$ and $x^* \in \mathcal{X}$ such that

$$\lim_{k\to+\infty} t_k = +\infty, \quad \lim_{k\to+\infty} x(t_k) = x^* \notin \mathcal{C}.$$

Set

$$\lim_{k\to+\infty} \operatorname{dist}(x(t_k), \mathcal{C}) = 2\iota > 0.$$

Then, there exists K such that

$$\lim_{k\to+\infty} \operatorname{dist}(x(t_k), \mathcal{C}) \geq \iota, \quad \forall k \geq K.$$

Using the boundedness of $\mathcal{X}$, $\dot{x}$ is bounded on $[0, +\infty)$. Denote $\sigma > 0$ such that

$$\|\dot{x}(t)\| \leq \sigma, \quad \forall t \in [0, +\infty). \tag{33}$$

Next, we will show that there exists $\varepsilon > 0$ such that

$$\|\dot{x}(t)\| \geq \varepsilon, \, \forall t \in [t_k, t_k + \iota/2\sigma]. \tag{34}$$

If not, there exists a sequence $s_k \in [t_k, t_k + \iota/2\sigma]$ such that

$$\lim_{k\to+\infty} s_k = +\infty \quad \text{and} \quad \lim_{k\to+\infty} \dot{x}(s_k) = \mathbf{0}.$$

On the one hand, since $\{x(s_k)\} \subseteq \mathcal{X}$ is bounded, there exist a subsequence of $\{x(s_k)\}$ (also denoted as $\{x(s_k)\}$) and $\tilde{x} \in \mathcal{X}$, such that $\lim_{k\to+\infty} x(s_k) = \tilde{x}$. Invoking of (19), by $\lim_{k\to+\infty} \mu(s_k) = 0$, Definition 3.1-(iii) and (17), it gives

$$\tilde{x} \in P_{\mathcal{X}}\left[\tilde{x} - \partial l(\tilde{x}) - \lambda \partial p(\tilde{x})\right], \tag{35}$$

which indicates that $\tilde{x}$ is a critical point of problem (5).

On the other hand, for any $t \in [t_k, t_k + \frac{\iota}{2\sigma}]$ and $\bar{x} \in \mathcal{C}$, we find that

$$\|x(t) - \bar{x}\| \geq \|x(t_k) - \bar{x}\| - \|x(t) - x(t_k)\| \geq \iota - \sigma(t - t_k) \geq \iota/2.$$

Thus, any accumulation point of $x(t)$ for $t \in [t_k, t_k + \iota/2\sigma]$ is not a critical point of (5), which leads to a contradiction to (35) and then (34) holds.

Notice that

$$\int_0^{+\infty} \|\dot{x}(t)\|^2 dt \leq \sum_{k=1}^{+\infty} \int_{t_k}^{t_k+\iota/2\sigma} \|\dot{x}(t)\|^2 dt = +\infty,$$

which also leads to a contradiction to the result in Proposition 4.1-(ii). Therefore, we can conclude that any accumulation point of $x(t)$ is a critical point of problem (5).

To prove $\mathcal{A}(x^*) = \mathcal{A}(\hat{x})$, we argue it by contradiction. Without loss of generality, we suppose that there exists $i \in \{1, 2, \ldots, n\}$ such that $x_i^* > 0$ but $\hat{x}_i = 0$. Denote $\{t_k\}$ and $\{s_k\}$ be the sequences converging to x^* and $\hat{x}$, respectively. By (21), we have $x_i^* \geq \nu$, where $\nu < u_i$. Since $\lim_{k\to+\infty} x_i(t_k) = x_i^* > 0$ and $\lim_{k\to+\infty} x_i(s_k) = 0$, there exists $T_2 \geq T$ such that $0 < x_i(T_2) < \nu$. By Proposition 4.3, we have $\lim_{t\to+\infty} x_i(t) = 0$, which leads to a contradiction to the supposition.

REFERENCES

[1] E.J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(1):489–509, 2006.

[2] P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. *Ann. Rev. Stat. Appl.*, 1(1):255–278, 2014.

[3] Y.F. Liu and Y.C. Wu. Variable selection via a combination of the L_0 and L_1 penalties. *J. Comput. Graph. Statist.*, 16(4):782–798, 2007.

[4] P. Buhlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. *Ann. Rev. Stat. Appl.*, 1:255–278, 2014.

[5] J. Fan, L.Z. Xue, and H. Zou. Strong oracle optimization of folded concave penalized estimation. *Ann. Statist.*, 42:819–849, 2014.

[6] J. Weston and A.B. Elisseeff and. Use of the zero-norm with linear models and kernel methods. *J. Mach. Learn. Res.*, 3:1439–1461, 2003.

[7] Hoai An Le Thi, Hoai Minh Le, and Tao Pham Dinh. Feature selection in machine learning: an exact penalty approach using a difference of convex function algorithm. *Mach. Learn.*, 101(1-3):163–186, 2015.

[8] B.K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, 1995.

[9] D.L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52:1289–1306, 2006.

[10] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52:489–509, 2006.

[11] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 9:1348–1360, 2001.

[12] Z. Zhang, Y.Y. Fan, and J.C. Lv. High dimensional thresholded regression and shrinkage effect. *J. R. Stat. Soc. Ser. B. Sta. Methodol.*, 76(3):627–649, 2014.

[13] E.J. Candès, M.B. Walkin, and S. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *J. Fourier Anal. Appl.*, 14(5-6):877–905, 2008.

[14] X. Chen, D. Ge, Z. Wang, and Y. Ye. Complexity of unconstrained ℓ_2-ℓ_p minimization. *Math. Progrm.*, 143:371–383, 2014.

[15] S. Foucart and M.J. Lai. Sparsest solutions of underdetermined linear system via ℓ_q-minimization for $0 < q \leq 1$. *Appl. Comput. Harmon. Anal.*, 26(3):395–407, 2009.

[16] C.S. Ong and L.T.H. An. Learning sparse classifiers with difference of convex functions algorithms. *Optim. Method Softw.*, 28(4):830–854, 2013.

[17] W. Bian and X. Chen. A smoothing proximal gradient algorithm for nonsmooth convex regression with cardinality penalty. *SIAM J. Numer. Anal.*, 58(1):858–883, 2020.

[18] C. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, 38:894–942, 2010.

[19] H.A. Le Thi, T. Pham Dinh, H.M. Le, and X.T. Vo. DC approximation approaches for sparse optimization. *Eur. J. Oper. Res.*, 244(1):26–46, 2015.

[20] J.S. Pan, Z. Hu, Z.X. Su, and M.H. Yang. L_0-regularized intensity and gradient prior for deblurring text images and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(2):342–355, 2017.

[21] J.C., W. Dan, and X.W. Zhang. L_0-based sparse canonical correlation analysis with application to cross-language document retrieval. *Neurocomputing*, 329:32–45, 2019.

[22] F.C. Xiong, J. Zhou, and Y.T. Qian. Hyperspectral restoration via L_0 gradient regularized low-rank tensor factorization. *IEEE Trans. Geosci. Remote Sens*, 57(12):10410–10425, 2019.

[23] E. Soubies, L. Blanc-Féraud, and G. Aubert. A continuous exact ℓ_0 penalty (CEL0) for least squares regularized problem. *SIAM J. Imaging Sci.*, 8(3):1607–1639, 2015.

[24] J.J. Hopfield and D.W. Tank. "neural" computation of decisions in optimization problems. *Biol. Cybern.*, 52(3):141–152, 1985.

[25] M.P. Kennedy and L.O. Chua. Neural networks for nonlinear programming. *IEEE Trans. Circuits Syst.*, 35(5):554–562, 1988.

[26] Y. Xia, G. Feng, and J. Wang. A novel recurrent neural network for solving nonlinear optimization problems with inequality constraints. *IEEE Trans. Neural Netw.*, 19(8):1340–1353, 2008.

[27] M. Forti, P. Nistri, and M. Quincampoix. Generalized neural network for nonsmooth nonlinear programming problems. *IEEE Trans. Circuits Syst. I - Regul. Pap.*, 51(9):1741–1754, 2004.

[28] W. Bian and X.P. Xue. Subgradient-based neural networks for nonsmooth nonconvex optimization problems. *IEEE Trans. Neural Netw.*, 20(6):1024–1038, 2009.

[29] N. Liu and S.T. Qin. A neurodynamic approach to nonlinear optimization problems with affine equality and convex inequality constraints. *Neural Netw.*, 109:147–158, 2019.

[30] 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.*, 22(4):601–613, 2011.

[31] W. Bian and X. Chen. Neural network for nonsmoth, nonconvex constrained minimization via smooth approximation. *IEEE Trans. Neural Netw. Learn. Syst.*, 25(3):545–556, 2014.

[32] W.J. Li, W. Bian, and X.P. Xue. Projected neural network for a class of non-lipschitz optimization problems with linear constraints. *IEEE Trans. Neural Netw. Learn. Syst.*, 31(9):3361–3373, 2020.

[33] X.B. Gao and L.L. Liao. A new projection-based neural network for constrained variational inequalities. *IEEE Trans. Neural Netw.*, 20(3):373–388, 2009.

[34] X. Chen. Smoothing methods for nonsmooth, nonconvex minimization. *Math. Program.*, 134(1):71–99, 2012.

[35] W.J. Li and W. Bian. Projection neural network for a class of sparse regression problems with cardinality penalty. *Neurocomputing*, 431:188–200, 2021.

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

[37] J.S. Pang, M. Razaviyayn, and A. Alvarado. Computing B-stationary points of nonsmooth DC programs. *Math. Oper. Res.*, 42(1):95–118, 2017.

[38] J. Kreimer and R.Y. Rubinstein. Nondifferentiable optimization via smooth approximation: general analytical approach. *Ann. Oper. Res.*, 39(1):97–119, 1992.

[39] R.T. Rockafellar and R. Wets. *Variational Analysis*. Springer-Verlag, Berlin, Germany, 1998.

[40] W. Bian, X. Chen, and Y. Ye. Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization. *Math. Program.*, 149(1-2):301–327, 2015.

[41] W. Bian and X. Chen. Smoothing neural network for constrained non-Lipschitz optimization with applications. *IEEE Trans. Neural Netw. Learn. Syst.*, 23(3):399–411, 2012.

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

[43] J.K. Hale. *Ordinary Differential Equations*. New York: Wiley, 1980.