Nonsmooth convex-concave saddle point problems with cardinality penalties

Wei Bian* and Xiaojun Chen†

Abstract

In this paper, we focus on a class of convexly constrained nonsmooth convex-concave saddle point problems with cardinality penalties. Although such nonsmooth nonconvex-nonconcave and discontinuous min-max problems may not have a saddle point, we show that they have a local saddle point and a global minimax point, and some local saddle points have the lower bound properties. We define a class of strong local saddle points based on the lower bound properties for stability of variable selection. Moreover we give a framework to construct continuous relaxations of the discontinuous min-max problems based on convolution, such that they have the same saddle points with the original problem. We also establish the relations between the continuous relaxation problems and the original problems regarding local saddle points, global minimax points, local minimax points and stationary points. Finally, we illustrate our results with distributionally robust sparse convex regression, sparse robust bond portfolio construction and sparse convex-concave logistic regression saddle point problems.

Keywords: nonsmooth min-max problem, nonconvex-nonconcave, local saddle point, sparse optimization, cardinality functions, smoothing method

MSC Classification: 90C46 , 49K35 , 90C30 , 65K05

1 Introduction

Let $c : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be a Lipschitz continuous function with $c(\mathbf{x}, \mathbf{y})$ convex in $\mathbf{x} \in \mathbb{R}^n$ for $\mathbf{y} \in \mathbb{R}^m$ and concave in $\mathbf{y} \in \mathbb{R}^m$ for $\mathbf{x} \in \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^{\hat{n}}$ and $h : \mathbb{R}^m \to \mathbb{R}^{\hat{m}}$ be continuously differentiable functions. For a vector $\mathbf{a} \in \mathbb{R}^k$, $\|\mathbf{a}_+\|_0$ is the cardinality function for the positive elements in $\mathbf{a}$, that is, $\|\mathbf{a}_+\|_0 = \|\max\{\mathbf{a}, \mathbf{0}\}\|_0 =$

*School of Mathematics, Harbin Institute of Technology, Harbin, China. Email: bianweilvse520@163.com. The research of this author is partially supported by National Key Research and Development Program of China (2021YFA1003500), National Natural Science Foundation of China grants (12271127, 62176073) and Fundamental Research Funds for Central Universities (HIT.OCEF.2024050,2022FRFK060017).

†Department of Applied Mathematics, Hong Kong Polytechnic University, Hong Kong, China. Email: maxjchen@polyu.edu.hk. The research of this author is partially supported by Hong Kong Research Grant Council project PolyU15300022 and CAS-Croucher Funding Scheme for AMSS-PolyU Joint Laboratory. Corresponding author.

Submitted 22 October, 2023; revised 30 April, 2024; accepted 27 June, 2024

$\sum_{i=1}^{k}(\max\{\mathbf{a}_i, 0\})^0$ with $0^0 = 0$. In this paper, we consider the saddle point problems with cardinality penalties in the following form

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) := c(\mathbf{x},\mathbf{y}) + \lambda_1\|g(\mathbf{x})_+\|_0 - \lambda_2\|h(\mathbf{y})_+\|_0, \tag{1.1}$$

where the feasible sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ are convex and compact, and the penalty parameters $\lambda_1, \lambda_2 \in \mathbb{R}$ are positive.

In the last few years, many interesting applications of the min-max problems have been found in machine learning and data science, especially the generative adversarial network (GAN) [23, 24, 31] and adversarial training [8]. Problem (1.1) is a discontinuous and nonconvex-nonconcave min-max problem, i.e. f is discontinuous in $\mathcal{X}\times\mathcal{Y}$, $f(\cdot,\mathbf{y})$ is not convex for some fixed $\mathbf{y}\in\mathcal{Y}$ and $f(\mathbf{x},\cdot)$ is not concave for some fixed $\mathbf{x}\in\mathcal{X}$. A special case of (1.1) is

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} c(\mathbf{x},\mathbf{y})+\lambda_1\|(\mathbf{x}-\underline{\mathbf{a}})_+\|_0+\lambda_1\|(\overline{\mathbf{a}}-\mathbf{x})_+\|_0-\lambda_2\|(\mathbf{y}-\underline{\mathbf{b}})_+\|_0-\lambda_2\|(\overline{\mathbf{b}}-\mathbf{y})_+\|_0, \tag{1.2}$$

where $\underline{\mathbf{a}}$, $\overline{\mathbf{a}} \in \mathbb{R}^n$ and $\underline{\mathbf{b}}$, $\overline{\mathbf{b}} \in \mathbb{R}^m$. In particular, if $\underline{\mathbf{a}} = \overline{\mathbf{a}} = \underline{\mathbf{b}} = \overline{\mathbf{b}} = \mathbf{0}$, then (1.2) reduces to the convex-concave saddle point problem with ℓ_0 penalties as follows

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} c(\mathbf{x},\mathbf{y}) + \lambda_1\|\mathbf{x}\|_0 - \lambda_2\|\mathbf{y}\|_0. \tag{1.3}$$

In 1928, von Neumann [37] proved that when c is a bilinear function, and $\mathcal{X}$, $\mathcal{Y}$ are two finite dimensional simplices,

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} c(\mathbf{x},\mathbf{y}) \tag{1.4}$$

has a saddle point and it holds

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} c(\mathbf{x},\mathbf{y}) = \max_{\mathbf{y}\in\mathcal{Y}}\min_{\mathbf{x}\in\mathcal{X}} c(\mathbf{x},\mathbf{y}). \tag{1.5}$$

This pioneering work has inspired a number of seminal contributions in the existence theory of saddle points of min-max problems in economics and engineering [19–21, 38, 43–46]. In 1949, Shiffman [45] gave a new proof of von Neumann's minimax theorem with a generalization to continuous convex-concave functions. Based on Brouwer's fixed point theorem, Nikaido [38] proved (1.5) for a continuous and quasi-convex-concave function c. Here, we call c is quasi-convex-concave if $c(\mathbf{x},\mathbf{y})$ is quasi-convex in $\mathbf{x}\in\mathbb{R}^n$ for $\mathbf{y}\in\mathbb{R}^m$ and quasi-concave in $\mathbf{y}\in\mathbb{R}^m$ for $\mathbf{x}\in\mathbb{R}^n$. In 1958, Sion [46] generalized von Neumann's result, and showed that if c is quasi-convex-concave and lower semicontinuous-upper semicontinuous, then (1.4) has a nonempty saddle point set whose closedness and convexity were pointed out in [43]. Moreover, we know from [20, Theorem 1.4.1] that

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) = \max_{\mathbf{y}\in\mathcal{Y}}\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y}) \tag{1.6}$$

is a necessary and sufficient condition for the existence of a saddle point of f over $\mathcal{X} \times \mathcal{Y}$.

When f is not convex-concave, (1.6) fails in general. The concept of local saddle points is defined by considering (1.6) locally at a point in $\mathcal{X} \times \mathcal{Y}$. However, a local saddle point also may not exist for a nonconvex-nonconcave min-max problem. In [33], Jin, Netrapalli and Jordan gave the definitions of global minimax points and local minimax points by considering the min-max problem as a two-player sequential game. Necessary and sufficient conditions for the local minimax points were studied in [18, 26, 31, 33]. Recently, Chen and Kelley [12] showed that a min-max problem for robust linear least squares problems does not have a saddle point, a local saddle point and a local minimax point, while it has infinitely first order stationary points and finite global minimax points. However, the set of first order stationary points and the set of global minimax points do not have a common point.

The cardinality functions in problem (1.1) play important roles to ensure the sparsity of the desirable solutions and improve the estimation accuracy by selecting important feature parameters. In the last decades, sparse minimization models with cardinality penalties have been widely used for sparse signal recovery, sparse variable selection, compressed sensing and statistical learning [2, 3, 7, 16, 47]. Advanced mathematical and statistical theory and efficient algorithms have been developed for sparse minimization [3, 9, 13, 32]. Recently, He et al. [28] systematically compared the solutions of a special quadratic minimization problem with ℓ_0 penalty, ℓ_1 penalty and capped-ℓ_1 penalty.

Inspired by the wide applications of saddle point problems and sparse optimization, we consider the sparse min-max problems modeled by (1.1). In section 6, we will use three applications to explain the motivation behind our research on this model and the importance of cardinality penalties in the model. To the best of our knowledge, mathematical theory and numerical algorithms on sparse saddle point problems with cardinality penalties have not been systematically studied.

Approximating cardinality functions by continuous or smooth functions is a promising approach in studying this class of problems. Many continuous relaxations to the cardinality function have been brought forward, such as the ℓ_1 norm [9], SCAD [22], hard thresholding [22], ℓ_p norm $(0 < p < 1)$ [13], MCP [50], capped-ℓ_1 [40], CEL0 [47], etc. In this paper, we construct continuous approximations to the cardinality functions in (1.1) based on convolution [10, 11], which include most popular relaxation functions to the cardinality function.

The main contributions of this paper have four parts.

- We prove the existence of a local saddle point and a global minimax point of (1.1), and define a class of strong local saddle points that have some desirable sparse properties.
- Based on convolution, we introduce two classes of density functions to provide a unified method for constructing the continuous relaxations with different smoothness to the cardinality function, which induce many popular continuous penalties in sparse optimization. Moreover, we propose the continuous relaxation problem of (1.1), which has both the local saddle points and global minimax points.

- We establish the relations between (1.1) and its continuous relaxations regarding the saddle points, local saddle points, local minimax points and global minimax points. Moreover, we define the first order and second order stationary points of the continuous relaxation problem. We show that both the first and second order stationary points of the continuous relaxation problems are not only the strong local saddle points of (1.1), but also have some promising computational tractability.
- We show the gradient consistency of a class of smoothing convex-concave functions to nonsmooth functions c. Moreover we prove that any accumulation point of weak d(irectional)-stationary points of the smoothing relaxation problem is a weak d-stationary point of the nonsmooth relaxation problem as the smoothing parameter goes to zero.

The rest of this paper is organized as follows. In Section 2, we prove the existence of local saddle points and global minimax points of (1.1). In Section 3, we construct the continuous relaxations to (1.1). In Section 4, we establish the relations between (1.1) and its continuous relaxation problems. The smoothing functions of nonsmooth function c are studied at the end of this section. In Section 5, we study the first order and second order stationary points of the continuous relaxation problems for a particular class of (1.1) and their relations with the strong local saddle points of (1.1). In Section 6, we show the applications of problem (1.1).

Notation Let $\mathbb{R}_+ = [0, +\infty)$, $\mathbb{R}_{++} = (0, +\infty)$ and $[n] = \{1, 2, \ldots, n\}$ for a positive integer n. For a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, $\mathbf{A}_{ij}$ means the element of $\mathbf{A}$ at the ith row and jth column. Let $\mathbf{e}_i$ be the vector with 1 at the ith element and 0 for the others and $\mathbf{e}$ be the vector with 1 for all elements. For a Lipschitz continuous function $c : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$, $\partial_{\mathbf{x}} c(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ and $\partial_{\mathbf{y}} c(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ mean the Clarke subgradients of c with respect to $\mathbf{x}$ and $\mathbf{y}$ at point $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, respectively. When c is Lipschitz continuously differentiable, $\partial^2 c(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ means the Clarke generalized Hessian of c at point $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, $\partial^2_{\mathbf{xx}} c(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ and $\partial^2_{\mathbf{yy}} c(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ mean the Clarke generalized Hessian of $c(\mathbf{x}, \mathbf{y})$ with respect to $\mathbf{x}$ and $\mathbf{y}$ at point $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, respectively. For $\mathbf{x} \in \mathbb{R}^n$ and $\delta > 0$, $\mathbf{B}(\mathbf{x}, \delta)$ means the closed ball centered at $\mathbf{x}$ with radius δ, $\mathcal{A}^+(\mathbf{x}) = \{l \in [\hat{n}] : g_l(\mathbf{x}) > 0\}$, $\mathcal{A}^-(\mathbf{x}) = \{l \in [\hat{n}] : g_l(\mathbf{x}) < 0\}$ and $\mathcal{A}^+_\delta(\mathbf{x}) = \{l \in [\hat{n}] : 0 < g_l(\mathbf{x}) < \delta\}$. Similarly, denote $\mathcal{B}^+(\mathbf{y}) = \{k \in [\hat{m}] : h_k(\mathbf{y}) > 0\}$, $\mathcal{B}^-(\mathbf{y}) = \{k \in [\hat{m}] : h_k(\mathbf{y}) < 0\}$ and $\mathcal{B}^+_\delta(\mathbf{y}) = \{k \in [\hat{m}] : 0 < h_k(\mathbf{y}) < \delta\}$. For a set $S \subseteq \mathbb{R}^n$ and $i \in [n]$, $S_i = \{\mathbf{x}_i : \mathbf{x} \in S\}$, $\|S\|_\infty = \sup\{\|\mathbf{x}\|_\infty : \mathbf{x} \in S\}$ and $\mathrm{co}\{S\} = \{\lambda\mathbf{x}^1 + (1-\lambda)\mathbf{x}^2 : \mathbf{x}^1, \mathbf{x}^2 \in S, \lambda \in [0, 1]\}$. For sets $S, \bar{S} \subseteq \mathbb{R}^n$, $S + \bar{S} = \{\mathbf{x}^1 + \mathbf{x}^2 : \mathbf{x}^1 \in S, \mathbf{x}^2 \in \bar{S}\}$. For a closed convex subset $\Omega \subseteq \mathbb{R}^n$ and $\mathbf{x} \in \Omega$, $N_\Omega(\mathbf{x})$ means the normal cone to Ω at $\mathbf{x}$.

2 Existence of local saddle points of problem (1.1)

In this section, we prove the existence of local saddle points and global minimax points of problem (1.1). We also define a class of strong local saddle points of (1.1) and provide its relation with saddle points of problem (1.4) in a certain subset of $\mathcal{X} \times \mathcal{Y}$. First of all, we give some necessary definitions.

Definition 2.1. *A point* $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ *is called a* ***saddle point*** *of problem* (1.1)*, if for all* $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$*, it holds*

$$f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}^*) \leq f(\mathbf{x}, \mathbf{y}^*). \tag{2.1}$$

We call $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ *a* ***local saddle point*** *of problem* (1.1)*, if there exists a* $\delta > 0$ *such that* (2.1) *holds for all* $\mathbf{x} \in \mathcal{X} \cap \mathbf{B}(\mathbf{x}^*, \delta)$ *and* $\mathbf{y} \in \mathcal{Y} \cap \mathbf{B}(\mathbf{y}^*, \delta)$*.*

Definition 2.2. *A point* $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ *is called a* ***global minimax point*** *of* (1.1)*, if for all* $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$*, we have*

$$f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}^*) \leq \max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}').$$

We call $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ *a* ***local minimax point*** *of* (1.1)*, if there exist a* $\delta_0 > 0$ *and a function* $\pi : \mathbb{R}_+ \to \mathbb{R}_+$ *satisfying* $\pi(\delta) \to 0$ *as* $\delta \to 0$ *such that for any* $\delta \in (0, \delta_0]$*,* $\mathbf{x} \in \mathcal{X} \cap \mathbf{B}(\mathbf{x}^*, \delta)$ *and* $\mathbf{y} \in \mathcal{Y} \cap \mathbf{B}(\mathbf{y}^*, \delta)$*, it holds*

$$f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}^*) \leq \max_{\mathbf{y}' \in \mathcal{Y} \cap \mathbf{B}(\mathbf{y}^*, \pi(\delta))} f(\mathbf{x}, \mathbf{y}'). \tag{2.2}$$

A local saddle point is a local minimax point, but a global minimax point is not necessarily a local minimax point. The two inequalities in (2.1) for $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ can be equivalently expressed by

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}^*) \quad \text{and} \quad \mathbf{y}^* \in \arg\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}^*, \mathbf{y}), \tag{2.3}$$

respectively. Since $\mathcal{X}$ and $\mathcal{Y}$ are compact, the lower semicontinuity of $((\cdot)_+)^0$ guarantees the existence of the solutions to the two optimization problems in (2.3), but $(\mathbf{x}^*, \mathbf{y}^*)$ may not be able to solve both the minimization and maximization simultaneously. This means that the saddle point set of problem (1.1) may be empty and $\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}) \neq \max_{\mathbf{y} \in \mathcal{Y}} \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y})$ (see Example 2.1).

Note that a nonconvex-nonconcave function may not have a saddle point, a local saddle point, or even a local minimax point. Fortunately, we can prove the existence of global minimax points, local saddle points and local minimax points of problem (1.1) without any additional assumption.

Proposition 2.1. *Min-max problem* (1.1) *always has a global minimax point.*

Proof. By the compactness of $\mathcal{Y}$, we can define $\psi(\mathbf{x}) = \max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y})$. Since function $f(\cdot, \mathbf{y})$ in (1.1) is lower semicontinuous for any fixed $\mathbf{y} \in \mathbb{R}^m$ and $\mathcal{X}$ is compact, ψ is lower semicontinuous on $\mathcal{X}$. Then, there exists a global solution to $\min_{\mathbf{x} \in \mathcal{X}} \psi(\mathbf{x})$, denoted by $\mathbf{x}^*$, i.e.

$$\max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}^*, \mathbf{y}') \leq \max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}'), \quad \forall\, \mathbf{x} \in \mathcal{X}. \tag{2.4}$$

The upper semicontinuity of $f(\mathbf{x}^*,\cdot)$ and the compactness of $\mathcal{Y}$ ensure the existence of the solution to $\max_{\mathbf{y}'\in\mathcal{Y}} f(\mathbf{x}^*,\mathbf{y}')$, denoted by $\mathbf{y}^*$, which implies

$$f(\mathbf{x}^*,\mathbf{y}^*)=\max_{\mathbf{y}'\in\mathcal{Y}} f(\mathbf{x}^*,\mathbf{y}')\geq f(\mathbf{x}^*,\mathbf{y}),\quad \forall \mathbf{y}\in\mathcal{Y}. \tag{2.5}$$

Therefore, (2.4) together with (2.5) implies that $(\mathbf{x}^*,\mathbf{y}^*)$ is a global minimax point of (1.1). □

Proposition 2.2. *Min-max problem* (1.1) *has a local saddle point and any saddle point of* (1.4) *is a local saddle point of* (1.1).

Proof. By Sion's minimax theorem [46], (1.4) has a saddle point $(\mathbf{x}^*,\mathbf{y}^*)\in\mathcal{X}\times\mathcal{Y}$ such that

$$c(\mathbf{x}^*,\mathbf{y})\leq c(\mathbf{x}^*,\mathbf{y}^*)\leq c(\mathbf{x},\mathbf{y}^*),\quad \forall\,\mathbf{x}\in\mathcal{X},\,\mathbf{y}\in\mathcal{Y}. \tag{2.6}$$

By the continuity of functions g_l and h_k, there exists a $\delta>0$ such that

$$\begin{aligned} &g_l(\mathbf{x})>0,\quad \forall l\in\mathcal{A}^+(\mathbf{x}^*),\,\mathbf{x}\in\mathbf{B}(\mathbf{x}^*,\delta)\cap\mathcal{X},\\ &h_k(\mathbf{y})>0,\quad \forall k\in\mathcal{B}^+(\mathbf{y}^*),\,\mathbf{y}\in\mathbf{B}(\mathbf{y}^*,\delta)\cap\mathcal{Y},\end{aligned}$$

which implies $\mathcal{A}^+(\mathbf{x}^*)\subseteq\mathcal{A}^+(\mathbf{x})$ and $\mathcal{B}^+(\mathbf{y}^*)\subseteq\mathcal{B}^+(\mathbf{y})$ in the above neighborhood of $(\mathbf{x}^*,\mathbf{y}^*)$. On one hand,

$$c(\mathbf{x}^*,\mathbf{y}^*)=f(\mathbf{x}^*,\mathbf{y}^*)-\lambda_1\sum_{l\in\mathcal{A}^+(\mathbf{x}^*)}1+\lambda_2\sum_{k\in\mathcal{B}^+(\mathbf{y}^*)}1.$$

On the other hand, for $\mathbf{x}\in\mathbf{B}(\mathbf{x}^*,\delta)\cap\mathcal{X}$ and $\mathbf{y}\in\mathbf{B}(\mathbf{y}^*,\delta)\cap\mathcal{Y}$, it has

$$c(\mathbf{x},\mathbf{y}^*)=f(\mathbf{x},\mathbf{y}^*)-\lambda_1\sum_{l\in\mathcal{A}^+(\mathbf{x})}1+\lambda_2\sum_{k\in\mathcal{B}^+(\mathbf{y}^*)}1\leq f(\mathbf{x},\mathbf{y}^*)-\lambda_1\sum_{l\in\mathcal{A}^+(\mathbf{x}^*)}1+\lambda_2\sum_{k\in\mathcal{B}^+(\mathbf{y}^*)}1,$$

and

$$c(\mathbf{x}^*,\mathbf{y})=f(\mathbf{x}^*,\mathbf{y})-\lambda_1\sum_{l\in\mathcal{A}^+(\mathbf{x}^*)}1+\lambda_2\sum_{k\in\mathcal{B}^+(\mathbf{y})}1\geq f(\mathbf{x}^*,\mathbf{y})-\lambda_1\sum_{l\in\mathcal{A}^+(\mathbf{x}^*)}1+\lambda_2\sum_{k\in\mathcal{B}^+(\mathbf{y}^*)}1.$$

Thus, we can conclude that

$$f(\mathbf{x}^*,\mathbf{y})\leq f(\mathbf{x}^*,\mathbf{y}^*)\leq f(\mathbf{x},\mathbf{y}^*),\quad \forall\,\mathbf{x}\in\mathbf{B}(\mathbf{x}^*,\delta)\cap\mathcal{X},\,\mathbf{y}\in\mathbf{B}(\mathbf{y}^*,\delta)\cap\mathcal{Y},$$

which implies that $(\mathbf{x}^*,\mathbf{y}^*)$ is a local saddle point of (1.1). □

Moreover, by [46], (1.1) also has a local saddle point if c is a continuous quasi-convex-concave function.

The following example shows that the parameters λ_1 and λ_2 play an important role for the existence of saddle points, local saddle points, global minimax points and local minimax points of min-max problem (1.1).

Example 2.1. *Consider the following min-max problem*

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) := (\mathbf{x}-1)(\mathbf{y}-1)+\lambda_1\|\mathbf{x}\|_0-\lambda_2\|\mathbf{y}\|_0, \tag{2.7}$$

where $\mathcal{X}=\mathcal{Y}=[-2,2]$ *and* $\lambda_1,\lambda_2>0$. *It is clear that* $c(\mathbf{x},\mathbf{y})=(\mathbf{x}-1)(\mathbf{y}-1)$ *is convex-concave on* $\mathcal{X}\times\mathcal{Y}$ *and* $(1,1)$ *is the unique saddle point of* $\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} c(\mathbf{x},\mathbf{y})$.

Case 1 (has no saddle point): Let $\lambda_1=3$ *and* $\lambda_2=1$. *By simple calculation, we find*

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y})=\begin{cases}\max\{5-3\mathbf{x},\mathbf{x}+1,4-\mathbf{x}\} & \text{if } \mathbf{x}\neq 0\\ 2 & \text{if } \mathbf{x}=0\end{cases}$$

and

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y})=\begin{cases}\min\{-3\mathbf{y}+5,-\mathbf{y},\mathbf{y}+1\} & \text{if } \mathbf{y}\neq 0\\ 1 & \text{if } \mathbf{y}=0.\end{cases}$$

Hence, we have

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y})=f(0,-2)=2 \quad \textit{and} \quad \max_{\mathbf{y}\in\mathcal{Y}}\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y})=f(0,0)=1.$$

Thus, (1.6) *fails in this case. By [20, Theorem 1.4.1], there is no saddle point to problem* (2.7) *with* $\lambda_1=3$ *and* $\lambda_2=1$. *On the other hand, it has four local saddle points:* $(0,0)$, $(1,1)$, $(0,-2)$ *and* $(2,0)$.

Case 2 (has a saddle point): Let $\lambda_1=\lambda_2=3$. *The similar calculation gives*

$$\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y})=\begin{cases}\max\{\mathbf{x}-1,4-\mathbf{x},-3\mathbf{x}+3\} & \text{if } \mathbf{x}\neq 0\\ 1 & \text{if } \mathbf{x}=0\end{cases}$$

and

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y})=\begin{cases}\min\{-2-\mathbf{y},3-3\mathbf{y},\mathbf{y}-1\} & \text{if } \mathbf{y}\neq 0\\ 1 & \text{if } \mathbf{y}=0.\end{cases}$$

Then, $\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) = \max_{\mathbf{y}\in\mathcal{Y}}\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y}) = 1$. *By [20, Theorem 1.4.1], there exists a saddle point to problem* (2.7) *with* $\lambda_1=\lambda_2=3$, *and* $(0,0)$ *is the unique saddle point.*

Case 3 (any global minimax point is not a local minimax point): Let $\lambda_1=\lambda_2=1$. *We have*

$$\psi(\mathbf{x})=\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y})=\begin{cases}\max\{3-3\mathbf{x},2-\mathbf{x},\mathbf{x}-1\} & \text{if } \mathbf{x}\neq 0\\ 2 & \text{if } \mathbf{x}=0.\end{cases}$$

Then, $\mathbf{x}^*=\arg\min_{\mathbf{x}\in\mathcal{X}}\psi(\mathbf{x})=\{3/2\}$ *and* $\arg\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}^*,\mathbf{y})=\{0,2\}$. *Thus, the set of global minimax points of* (2.7) *with* $\lambda_1=\lambda_2=1$ *contains only two points* $(3/2,0)$ *and* $(3/2,2)$, *and* $f(3/2,0)=f(3/2,2)=1/2$. *Moreover, around* $\mathbf{x}^*=3/2$, *for any* $0<\delta<1/2$, $\max_{\mathbf{y}'\in\{\mathbf{y}\in\mathcal{Y}:|\mathbf{y}|\leq\delta\}} f(\mathbf{x},\mathbf{y}')=f(\mathbf{x},0)=2-\mathbf{x}$ *and*

$\max_{\mathbf{y}'\in\{\mathbf{y}\in\mathcal{Y}:|\mathbf{y}-2|\le\delta\}} f(\mathbf{x},\mathbf{y}') = f(\mathbf{x},2) = \mathbf{x}-1$, *which means that neither* $(3/2,0)$ *nor* $(3/2,2)$ *is a local minimax point of* (2.7) *with* $\lambda_1=\lambda_2=1$.

To study sparse saddle points, we introduce a class of strong local saddle points of (1.1).

Definition 2.3. *For a given* $\nu>0$, *we call* $(\mathbf{x}^*,\mathbf{y}^*)\in\mathcal{X}\times\mathcal{Y}$ *a* ***ν-strong local saddle point*** *of problem* (1.1), *if it is a local saddle point of* (1.1) *and satisfies the lower bound property as follows*

$$g_l(\mathbf{x}^*)\notin(0,\nu),\ \forall l\in[\hat{n}]\quad \textit{and}\quad h_k(\mathbf{y}^*)\notin(0,\nu),\ \forall k\in[\hat{m}]. \tag{2.8}$$

On one hand, for any local saddle point $(\mathbf{x}^*,\mathbf{y}^*)$ of (1.1), there exists a $\nu>0$ such that (2.8) holds, where we can set $\nu=\min\{1,g_l(\mathbf{x}^*),h_k(\mathbf{y}^*):l\in\mathcal{A}^+(\mathbf{x}^*),k\in\mathcal{B}^+(\mathbf{y}^*)\}$. Hence, (1.1) has a ν-strong local saddle point with a certain value of ν. On the other hand, for a given $\nu>0$, not all local saddle points of (1.1) satisfy (2.8) (see Example 2.2). In particular, if $(\mathbf{x}^*,\mathbf{y}^*)$ is a ν-strong local saddle point of (1.3), then

$$|\mathbf{x}_i^*|\notin(0,\nu)\text{ and }|\mathbf{y}_j^*|\notin(0,\nu),\ \forall i\in[n],\ j\in[m],$$

which not only helps us distinguish the zero and nonzero elements efficiently, but also provides a solution with certain stability [3, 13]. Therefore, the study on ν-strong local saddle points of (1.1) is interesting and important in sparse problems.

Example 2.2. *Consider the following min-max problem*

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) := |\mathbf{x}_1+\mathbf{x}_2-1|(\mathbf{y}+1)+\|\mathbf{x}\|_0-3\|\mathbf{y}\|_0 \tag{2.9}$$

with $\mathcal{X}=[-1,1]^2$ *and* $\mathcal{Y}=[-1,1]$. *By simple calculation, we can find that* (2.9) *has three saddle points, i.e.* $(0,0,0)^\top$, $(1,0,0)^\top$ *and* $(0,1,0)^\top$ *with*

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) = \max_{\mathbf{y}\in\mathcal{Y}}\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y}) = 1.$$

The local saddle point set of (2.9) *is*

$$\mathcal{SL} := \{(\mathbf{x}_1,1-\mathbf{x}_1,\mathbf{y})^\top:\mathbf{x}_1\in[-1,1],\mathbf{y}\in[-1,1]\}\cup\{(0,0,0)^\top,(0,0,1)^\top\},$$

while the ν*-strong local saddle point set of* (2.9) *with* $0<\nu<1$ *is*

$$\mathcal{SL}\cap\{(\mathbf{x}_1,\mathbf{x}_2,\mathbf{y})^\top:\ |\mathbf{x}_1|\notin(0,\nu),\ |\mathbf{x}_2|\notin(0,\nu)\ \textit{and}\ |\mathbf{y}|\notin(0,\nu)\}.$$

Notice that the ν*-strong local saddle point set of* (2.9) *is a proper subset of its local saddle point set, and contains all saddle points of* (2.9).

For a given $\delta>0$, by [20, Theorem 1.4.1], we know that $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ is a saddle point of problem (1.1) on $(\mathcal{X}\cap\mathbf{B}(\bar{\mathbf{x}},\delta))\times(\mathcal{Y}\cap\mathbf{B}(\bar{\mathbf{y}},\delta))$ if and only if

$$\max_{\mathbf{y}\in\mathcal{Y}\cap\mathbf{B}(\bar{\mathbf{y}},\delta)}\min_{\mathbf{x}\in\mathcal{X}\cap\mathbf{B}(\bar{\mathbf{x}},\delta)} f(\mathbf{x},\mathbf{y}) = f(\bar{\mathbf{x}},\bar{\mathbf{y}}) = \min_{\mathbf{x}\in\mathcal{X}\cap\mathbf{B}(\bar{\mathbf{x}},\delta)}\max_{\mathbf{y}\in\mathcal{Y}\cap\mathbf{B}(\bar{\mathbf{y}},\delta)} f(\mathbf{x},\mathbf{y}). \tag{2.10}$$

Hence $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$ is a local saddle point of (1.1) if and only if there is a $\delta > 0$ such that (2.10) holds. In what follows, we provide the relation between ν-strong local saddle points of (1.1) and local saddle points of c restricted to a certain set.

Theorem 2.1. *For $\nu > 0$ and $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}$, let $\hat{\mathcal{X}}(\bar{\mathbf{x}}) = \{\mathbf{x} \in \mathcal{X} : g_l(\mathbf{x}) \leq 0, \forall l \notin \mathcal{A}^+(\bar{\mathbf{x}})\}$ and $\hat{\mathcal{Y}}(\bar{\mathbf{y}}) = \{\mathbf{y} \in \mathcal{Y} : h_k(\mathbf{y}) \leq 0, \forall k \notin \mathcal{B}^+(\bar{\mathbf{y}})\}$. Then the following statements are equivalent.*

(i) *$(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a ν-strong local saddle point of (1.1);*

(ii) *$(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a local saddle point of c on $\hat{\mathcal{X}}(\bar{\mathbf{x}}) \times \hat{\mathcal{Y}}(\bar{\mathbf{y}})$ and satisfies (2.8).*

Proof. (i)⇒(ii). Suppose $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a ν-strong local saddle point of (1.1), then there exists a $\delta > 0$ such that

$$f(\bar{\mathbf{x}}, \mathbf{y}) \leq f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \leq f(\mathbf{x}, \bar{\mathbf{y}}), \quad \forall\, \mathbf{x} \in \mathcal{X} \cap \mathbf{B}(\bar{\mathbf{x}}, \delta),\, \mathbf{y} \in \mathcal{Y} \cap \mathbf{B}(\bar{\mathbf{y}}, \delta). \tag{2.11}$$

For any $\mathbf{x} \in \hat{\mathcal{X}}(\bar{\mathbf{x}})$, it holds that $\|g(\mathbf{x})_+\|_0 \leq \|g(\bar{\mathbf{x}})_+\|_0$. Rearranging the second inequality in (2.11) gives

$$c(\bar{\mathbf{x}}, \bar{\mathbf{y}}) + \lambda_1 \|g(\bar{\mathbf{x}})_+\|_0 \leq c(\mathbf{x}, \bar{\mathbf{y}}) + \lambda_1 \|g(\mathbf{x})_+\|_0, \quad \forall\, \mathbf{x} \in \mathcal{X} \cap \mathbf{B}(\bar{\mathbf{x}}, \delta).$$

Thus, $\bar{\mathbf{x}}$ is a local minimizer of $c(\cdot, \bar{\mathbf{y}})$ on $\hat{\mathcal{X}}(\bar{\mathbf{x}})$. Following the same way, the first inequality in (2.11) gives that $\bar{\mathbf{y}}$ is a local maximizer of $c(\bar{\mathbf{x}}, \cdot)$ on $\hat{\mathcal{Y}}(\bar{\mathbf{y}})$. Thus, (ii) holds.

(ii)⇒(i). Since $\bar{\mathbf{x}}$ is a local minimizer of $c(\cdot, \bar{\mathbf{y}})$ on $\hat{\mathcal{X}}(\bar{\mathbf{x}})$ and satisfies (2.8), there exists a $\delta_1 > 0$ such that

$$c(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \leq c(\mathbf{x}, \bar{\mathbf{y}}), \quad \forall\, \mathbf{x} \in \hat{\mathcal{X}}(\bar{\mathbf{x}}) \cap \mathbf{B}(\bar{\mathbf{x}}, \delta_1)$$

and

$$g_l(\bar{\mathbf{x}}) \notin (0, \nu),\ \forall l \in [\hat{n}]. \tag{2.12}$$

Based on (2.12), there exists a $\delta_2 \in (0, \delta_1]$ such that $g_l(\mathbf{x}) > 0$, $\forall\, \mathbf{x} \in \mathbf{B}(\bar{\mathbf{x}}, \delta_2)$, $l \in \mathcal{A}^+(\bar{\mathbf{x}})$, which implies

$$\|g(\bar{\mathbf{x}})_+\|_0 \leq \|g(\mathbf{x})_+\|_0, \quad \forall\, \mathbf{x} \in \mathbf{B}(\bar{\mathbf{x}}, \delta_2). \tag{2.13}$$

Then,

$$f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \leq f(\mathbf{x}, \bar{\mathbf{y}}), \quad \forall\, \mathbf{x} \in \hat{\mathcal{X}}(\bar{\mathbf{x}}) \cap \mathbf{B}(\bar{\mathbf{x}}, \delta_2). \tag{2.14}$$

Due to the continuity of $c(\cdot, \bar{\mathbf{y}})$, there is a $\delta_3 \in (0, \delta_2]$ such that

$$c(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \leq c(\mathbf{x}, \bar{\mathbf{y}}) + \lambda_1, \quad \forall\, \mathbf{x} \in \mathbf{B}(\bar{\mathbf{x}}, \delta_3). \tag{2.15}$$

When $\mathbf{x} \in \mathcal{X} \cap \mathbf{B}(\bar{\mathbf{x}}, \delta_3)$ but $\mathbf{x} \notin \hat{\mathcal{X}}(\bar{\mathbf{x}})$, there exists an $\hat{l} \notin \mathcal{A}^+(\bar{\mathbf{x}})$ such that $g_{\hat{l}}(\mathbf{x}) > 0$, which together with (2.13) further gives

$$\|g(\bar{\mathbf{x}})_+\|_0 + 1 \leq \|g(\mathbf{x})_+\|_0. \tag{2.16}$$

Thanks to (2.14)-(2.16), we have

$$f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \leq f(\mathbf{x}, \bar{\mathbf{y}}), \quad \forall\, \mathbf{x} \in \mathcal{X} \cap \mathbf{B}(\bar{\mathbf{x}}, \delta_3). \tag{2.17}$$

We can ensure $f(\bar{\mathbf{x}}, \mathbf{y}) \leq f(\bar{\mathbf{x}}, \bar{\mathbf{y}})$, $\forall \mathbf{y} \in \mathcal{Y} \cap \mathbf{B}(\bar{\mathbf{y}}, \delta_4)$ with $\delta_4 > 0$ in the same way. Thus, $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a ν-strong local saddle point of (1.1). □

For problem (1.3), $\hat{\mathcal{X}}(\bar{\mathbf{x}}) = \{\mathbf{x} \in \mathcal{X} : \mathbf{x}_i \geq 0, \forall i \notin \mathcal{A}^-(\bar{\mathbf{x}}) \text{ and } \mathbf{x}_i \leq 0, \forall i \notin \mathcal{A}^+(\bar{\mathbf{x}})\}$ and $\hat{\mathcal{Y}}(\bar{\mathbf{y}}) = \{\mathbf{y}_j \geq 0, \forall j \notin \mathcal{B}^-(\bar{\mathbf{y}}) \text{ and } \mathbf{y}_j \leq 0, \forall j \notin \mathcal{B}^+(\bar{\mathbf{y}})\}$. Thus, together the formulations of $\hat{\mathcal{X}}(\bar{\mathbf{x}})$ and $\hat{\mathcal{Y}}(\bar{\mathbf{y}})$ with the local optimality conditions, we obtain that $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a local saddle point of (1.3) if and only if $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a saddle point of c on $\mathcal{X}^0(\bar{\mathbf{x}}) \times \mathcal{Y}^0(\bar{\mathbf{y}})$ with $\mathcal{X}^0(\bar{\mathbf{x}}) = \{\mathbf{x} \in \mathcal{X} : \mathbf{x}_i = 0 \text{ if } \bar{\mathbf{x}}_i = 0\}$ and $\mathcal{Y}^0(\bar{\mathbf{y}}) = \{\mathbf{y} \in \mathcal{Y} : \mathbf{y}_j = 0 \text{ if } \bar{\mathbf{y}}_j = 0\}$, i.e. the subspace corresponding to nonzero components of $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$.

By "pull-down" the discontinuity of the objective at an $\bar{\mathbf{x}} \in \mathcal{X}$ to the constraints, the authors in [16] brought forward the notion of "pseudo stationary" problem and the corresponding pseudo local minimizer for the minimization problem with $((\cdot)_+)^0$ in the objective and constraints. By [16, Proposition 4], the special structure of $f(\mathbf{x}, \bar{\mathbf{y}})$ with $\bar{\mathbf{y}} \in \mathcal{Y}$ and the Lipschitz continuity of g_l, we find that $\bar{\mathbf{x}}$ is a local minimizer of $\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \bar{\mathbf{y}})$ if and only if it is a pseudo local minimizer of it, i.e. $\bar{\mathbf{x}}$ is a local minimizer of $\min_{\mathbf{x} \in \hat{\mathcal{X}}(\bar{\mathbf{x}})} c(\mathbf{x}, \bar{\mathbf{y}})$ with $\hat{\mathcal{X}}(\bar{\mathbf{x}}) = \{\mathbf{x} \in \mathcal{X} : g_l(\mathbf{x}) \leq 0, \forall l \notin \mathcal{A}^+(\bar{\mathbf{x}})\}$. It is stated in [27] that a pseudo B-stationary solution is necessary to be a pseudo local minimizer, where we call $\bar{\mathbf{x}}$ a pseudo B-stationary solution of $\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \bar{\mathbf{y}})$, if it is a B-stationary solution of $\min_{\mathbf{x} \in \hat{\mathcal{X}}(\bar{\mathbf{x}})} c(\mathbf{x}, \bar{\mathbf{y}})$ [16]. Similar ideas are also employed in [27, 34]. Combining Theorem 2.1 with [16, Proposition 4], $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a ν-strong local saddle point of (1.1), if and only if $\bar{\mathbf{x}}$ is a pseudo local minimizer of $\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \bar{\mathbf{y}})$, $\bar{\mathbf{y}}$ is a pseudo local minimizer of $\min_{\mathbf{y} \in \mathcal{Y}} -f(\bar{\mathbf{x}}, \mathbf{y})$ and $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ satisfies the lower bounds in (2.8). In general, a pseudo B-stationary solution to the min-max problem (1.1) defined by a similar way is not necessary to be a local saddle point. However, when c is convex-concave, and g_l, $\forall l \in [\hat{n}]$, h_k, $\forall k \in [\hat{m}]$ are convex, by Theorem 2.1, $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is local saddle point of (1.1), if and only if $\bar{\mathbf{x}}$ is a pseudo B-stationary solution of $\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \bar{\mathbf{y}})$ and $\bar{\mathbf{y}}$ is a pseudo B-stationary solution of $\min_{\mathbf{y} \in \mathcal{Y}} -f(\bar{\mathbf{x}}, \mathbf{y})$.

3 Continuous relaxations

In this section, we propose a class of continuous relaxations to the cardinality function in min-max problem (1.1) based on convolution [10], which include the capped-ℓ_p function [40], SCAD function [22], MCP function [50] and hard thresholding penalty function [22] as special cases. Then, we show the existence of local saddle points to the continuous relaxations of (1.1).

3.1 Density functions

Let $\rho : \mathbb{R} \to \mathbb{R}_+$ be a piecewise continuous density function satisfying

$$\rho(s) = 0, \ \forall s \notin [0, \alpha] \tag{3.1}$$

with a positive number α, which means that $\int_0^\alpha \rho(s)ds = 1$. Then, for any fixed $\mu > 0$,

$$\begin{aligned} r(t,\mu) &:= \int_{-\infty}^{+\infty} ((t-\mu s)_+)^0 \rho(s)ds \\ &= \int_{-\infty}^{\frac{t}{\mu}} \rho(s)ds = (t_+)^0 + \begin{cases} 0 & \text{if } t \le 0 \\ -\int_{\frac{t}{\mu}}^{+\infty} \rho(s)ds & \text{if } t > 0 \end{cases} \end{aligned} \tag{3.2}$$

is well-defined, and when $r(\cdot,\mu)$ is Lipschitz continuous around t, it holds

$$\partial_t r(t,\mu) = \mathrm{co}\left\{ \lim \frac{\rho(t_i/\mu)}{\mu} : t_i \to t,\ \rho \text{ is continuous at } t_i/\mu \right\}. \tag{3.3}$$

The continuous relaxation in (3.2) is inspired by the smoothing function to t_+ in [11, 41, 44]. We can use formulation (3.2) to construct a continuous relaxation r by a density function ρ.

By (3.2), for any $\mu > 0$, we have

$$r(t,\mu) = (t_+)^0, \quad \forall\, t \notin (0, \alpha\mu), \tag{3.4}$$
$$r(t,\mu) - (t_+)^0 \le 0, \quad \forall\, t \in \mathbb{R}, \tag{3.5}$$
$$\lim_{\mu \downarrow 0} r(t,\mu) = (t_+)^0, \quad \forall\, t \in \mathbb{R}, \tag{3.6}$$
$$\lim_{a \to t, \mu \downarrow 0} r(a,\mu) = (t_+)^0, \quad \forall\, t \neq 0. \tag{3.7}$$

For any $t \in \mathbb{R}$, we see from (3.5) and (3.6) that $r(t,\mu)$ approximates $(t_+)^0$ from below as μ tends to 0. In what follows, we give four examples of the function r with ρ satisfying (3.1).

Example 3.1. *Choose a density function with $\alpha = 1$ and $0 < p \le 1$ as*

$$\rho(s) = \begin{cases} ps^{p-1} & \text{if } 0 < s < 1 \\ 0 & \text{otherwise} \end{cases} \quad \Rightarrow \quad r(t,\mu) = \begin{cases} 0 & \text{if } t \le 0 \\ \dfrac{t^p}{\mu^p} & \text{if } 0 < t \le \mu \\ 1 & \text{if } t > \mu. \end{cases}$$

Here, $r(\cdot,\mu)$ with $p = 1$ is the ***capped-ℓ_1 function*** $\varphi_{\mathtt{cap}}$[1] *on $\mathbb{R}_+$.*

Example 3.2. *For any $\alpha > 1$, choose a density function as*

$$\rho(s) = \begin{cases} \dfrac{2}{\alpha+1} & \text{if } 0 \le s \le 1 \\ \dfrac{2\alpha - 2s}{(\alpha-1)(\alpha+1)} & \text{if } 1 < s \le \alpha \\ 0 & \text{otherwise} \end{cases} \quad \Rightarrow \quad r(t,\mu) = \begin{cases} 0 & \text{if } t \le 0 \\ \dfrac{2t}{(\alpha+1)\mu} & \text{if } 0 < t \le \mu \\ \dfrac{2\alpha\mu t - t^2 - \mu^2}{(\alpha-1)(\alpha+1)\mu^2} & \text{if } \mu < t \le \alpha\mu \\ 1 & \text{if } t > \alpha\mu. \end{cases}$$

[1] capped-ℓ_1 function: $\varphi_{\mathtt{cap}}(t) = \min\{1, |t|/\mu\}$.

Here, $r(\cdot,\mu)$ is a scaled ***SCAD function*** $\varphi_{\texttt{SCAD}}$[2] *on $\mathbb{R}_+$, i.e. $r(t,\mu)=\frac{2}{(\alpha+1)\mu}\varphi_{\texttt{SCAD}}(t)$, $\forall t\geq 0$.*

Example 3.3. *For any $\alpha>0$, choose a density function as*

$$\rho(s)=\begin{cases}\dfrac{2}{\alpha}-\dfrac{2s}{\alpha^2} & \text{if } 0\leq s\leq\alpha\\ 0 & \text{otherwise}\end{cases}\quad\Rightarrow\quad r(t,\mu)=\begin{cases}0 & \text{if } t\leq 0\\ \dfrac{2t}{\alpha\mu}-\dfrac{t^2}{\alpha^2\mu^2} & \text{if } 0<t\leq\alpha\mu\\ 1 & \text{if } t>\alpha\mu.\end{cases}$$

Here, $r(\cdot,\mu)$ is a scaled ***MCP function*** $\varphi_{\texttt{MCP}}$[3] *on $\mathbb{R}_+$, i.e. $r(t,\mu)=\frac{2}{\alpha\mu}\varphi_{\texttt{MCP}}(t)$, $\forall t\geq 0$.*

Example 3.4. *Choose a density function with $\alpha=1$ as*

$$\rho(s)=\begin{cases}2(1-s) & \text{if } 0<s<1\\ 0 & \text{otherwise}\end{cases}\quad\Rightarrow\quad r(t,\mu)=\begin{cases}0 & \text{if } t\leq 0\\ 1-(1-t/\mu)^2 & \text{if } 0<t\leq\mu\\ 1 & \text{if } t>\mu.\end{cases}$$

Here, $r(\cdot,\mu)$ is the ***hard thresholding penalty function*** $\varphi_{\texttt{hard}}$[4] *on $\mathbb{R}_+$.*

For further analysis, we bring forward two assumptions on density function ρ.

Assumption 3.1. *There exists a positive number $\underline{\rho}$ such that the density function $\rho:\mathbb{R}\to\mathbb{R}_+$ satisfies*

$$\rho(s)\geq\underline{\rho},\quad\forall s\in(0,\alpha).$$

Assumption 3.2. *The density function ρ is Lipschitz continuous on $\mathbb{R}_{++}$ and there exist $\underline{\rho}_2>0$ and $\check{\rho}_2>0$ such that for any $s\in(0,\alpha)$,*

$$\textit{either } \rho(s)\geq\underline{\rho}_2 \quad \textit{or} \quad \sup\{a:a\in\partial\rho(s)\}\leq-\check{\rho}_2.$$

Notice that if ρ is Lipschitz continuous on $\mathbb{R}_{++}$ and $\rho(s)=0$, $\forall s\notin[0,\alpha]$, then Assumption 3.1 fails. Thus, ρ can not satisfy Assumption 3.1 and Assumption 3.2 at the same time.

When the density function ρ satisfies Assumption 3.1 and $r(\cdot,\mu)$ is Lipschitz continuous around t, we have

$$\inf\{\xi:\xi\in\partial_t r(t,\mu)\}\geq\underline{\rho}/\mu,\quad\forall\, t\in(0,\alpha\mu). \tag{3.8}$$

---

[2] SCAD function: $\varphi_{\texttt{SCAD}}(t)=\begin{cases}t & \text{if } t\leq\mu\\ \dfrac{2\alpha\mu t-t^2-\mu^2}{2(\alpha-1)\mu} & \text{if } \mu<t\leq\alpha\mu\\ \dfrac{(\alpha+1)\mu}{2} & \text{if } t>\alpha\mu.\end{cases}$

[3] MCP function: $\varphi_{\texttt{MCP}}(t)=\begin{cases}\dfrac{\alpha\mu}{2} & \text{if } t\geq\alpha\mu\\ t-\dfrac{t^2}{2\alpha\mu} & \text{if } t<\alpha\mu.\end{cases}$

[4] hard thresholding penalty function: $\varphi_{\texttt{hard}}(t)=1-(1-t/\mu)_+^2$.

When the density function ρ satisfies Assumption 3.2, inspired by (3.3), we have that for any $\mu > 0$, $r(\cdot, \mu)$ is Lipschitz continuously differentiable on $\mathbb{R}_{++}$ and satisfies

$$\partial_t^2 r(t, \mu) = \partial\rho(t/\mu)/\mu^2, \quad \forall\, t > 0, \tag{3.9}$$

which implies for any $t \in (0, \alpha\mu)$ such that $\sup\{a : a \in \partial\rho(t/\mu)\} \le -\check{\rho}_2$, it holds

$$\sup\{\xi : \xi \in \partial_t^2 r(t, \mu)\} \le -\check{\rho}_2/\mu^2.$$

Since all the four density functions ρ in Examples 3.1-3.4 satisfy (3.1), $r(t, \mu)$ in these examples satisfy (3.4)-(3.6). To end this subsection, we use Table 1 to conclude the different properties of the density functions and the corresponding continuous functions r in these four examples.

Example	differentiability of $r(\cdot, \mu)$	Assumption 3.1	Assumption 3.2
3.1	not at 0, μ	$\underline{\rho} = p$	$\times$
3.2	not at 0	$\times$	$\underline{\rho}_2 = \frac{2}{\alpha+1}$, $\check{\rho}_2 = \frac{2}{(\alpha+1)(\alpha-1)}$
3.3	not at 0	$\times$	$\underline{\rho}_2 > 0$, $\check{\rho}_2 = \frac{2}{\alpha^2}$
3.4	not at 0	$\times$	$\underline{\rho}_2 > 0$, $\check{\rho}_2 = 2$

Table 1: Properties of the density functions ρ and corresponding functions r in Examples 3.1-3.4

3.2 Continuous relaxation models to (1.1)

In what follows, we will use the continuous function r defined in (3.2) to approximate the cardinality function in (1.1). For $l \in [\hat{n}]$ and $k \in [\hat{m}]$, denote

$$\phi_l(\mathbf{x}) = (g_l(\mathbf{x})_+)^0 \quad \text{and} \quad \psi_k(\mathbf{y}) = (h_k(\mathbf{y})_+)^0,$$

and define their continuous relaxations by

$$\phi_l^R(\mathbf{x}, \mu) = r(g_l(\mathbf{x}), \mu) \quad \text{and} \quad \psi_k^R(\mathbf{y}, \mu) = r(h_k(\mathbf{y}), \mu). \tag{3.10}$$

For any $\mu > 0$, $l \in [\hat{n}]$ and $k \in [\hat{m}]$, by (3.5), we have

$$\phi_l^R(\mathbf{x}, \mu) \le \phi_l(\mathbf{x}) \quad \text{and} \quad \psi_k^R(\mathbf{y}, \mu) \le \psi_k(\mathbf{y}), \qquad \forall\, \mathbf{x} \in \mathcal{X},\, \mathbf{y} \in \mathcal{Y}. \tag{3.11}$$

We propose the continuous relaxation of (1.1) as follows

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f^R(\mathbf{x}, \mathbf{y}, \mu) := c(\mathbf{x}, \mathbf{y}) + \lambda_1 \sum_{l\in[\hat{n}]} \phi_l^R(\mathbf{x}, \mu) - \lambda_2 \sum_{k\in[\hat{m}]} \psi_k^R(\mathbf{y}, \mu), \tag{3.12}$$

where μ is a given positive number. Here, $f^R(\cdot, \cdot, \mu)$ in (3.12) is continuous on $\mathcal{X} \times \mathcal{Y}$, and it is clear by (3.6) that, $\lim_{\mu\downarrow 0} f^R(\mathbf{x}, \mathbf{y}, \mu) = f(\mathbf{x}, \mathbf{y})$ for any $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$.

Notice that (3.12) is a nonconvex-nonconcave min-max problem and may not have a saddle point. However, similar to Proposition 2.2, we can have the existence results for the local saddle points of (3.12).

Proposition 3.1. *There exists a $\tilde{\mu} > 0$ such that (3.12) has a local saddle point for any $\mu \in (0, \tilde{\mu})$.*

Proof. Let $(\mathbf{x}^*, \mathbf{y}^*)$ be a saddle point of $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} c(\mathbf{x}, \mathbf{y})$, i.e. (2.6) holds. Denote $\vartheta = \min\{1, g_l(\mathbf{x}^*), h_k(\mathbf{y}^*) : l \in \mathcal{A}^+(\mathbf{x}^*), k \in \mathcal{B}^+(\mathbf{y}^*)\}$ and set $\tilde{\mu} = \vartheta/2\alpha$, then $g_l(\mathbf{x}^*) \geq 2\alpha\tilde{\mu}$, $h_k(\mathbf{y}^*) \geq 2\alpha\tilde{\mu}$, $\forall l \in \mathcal{A}^+(\mathbf{x}^*)$, $k \in \mathcal{B}^+(\mathbf{y}^*)$. Choose $\mu \in (0, \tilde{\mu})$. By the continuity of g and h, there exists a $\delta > 0$ such that for any $l \in \mathcal{A}^+(\mathbf{x}^*)$, $k \in \mathcal{B}^+(\mathbf{y}^*)$, $\mathbf{x} \in \mathbf{B}(\mathbf{x}^*, \delta)$ and $\mathbf{y} \in \mathbf{B}(\mathbf{y}^*, \delta)$, it holds $g_l(\mathbf{x}) \geq \alpha\mu$ and $h_k(\mathbf{y}) \geq \alpha\mu$, by (3.4), which further implies

$$\phi_l^R(\mathbf{x}, \mu) = 1 \quad \text{and} \quad \psi_k^R(\mathbf{y}, \mu) = 1.$$

This means that, for any $\mathbf{x} \in \mathbf{B}(\mathbf{x}^*, \delta)$ and $\mathbf{y} \in \mathbf{B}(\mathbf{y}^*, \delta)$,

$$\sum_{l\in[\hat{n}]} \phi_l^R(\mathbf{x}^*, \mu) \leq \sum_{l\in[\hat{n}]} \phi_l^R(\mathbf{x}, \mu) \quad \text{and} \quad \sum_{k\in[\hat{m}]} \psi_k^R(\mathbf{y}^*, \mu) \leq \sum_{k\in[\hat{m}]} \psi_k^R(\mathbf{y}, \mu). \tag{3.13}$$

Together (3.13) with (2.6), we obtain

$$f^R(\mathbf{x}^*, \mathbf{y}, \mu) \leq f^R(\mathbf{x}^*, \mathbf{y}^*, \mu) \leq f^R(\mathbf{x}, \mathbf{y}^*, \mu), \quad \mathbf{x} \in \mathbf{B}(\mathbf{x}^*, \delta) \cap \mathcal{X}, \, \mathbf{y} \in \mathbf{B}(\mathbf{y}^*, \delta) \cap \mathcal{Y},$$

which means that $(\mathbf{x}^*, \mathbf{y}^*)$ a local saddle point of (3.12). □

From the compactness of $\mathcal{X}$ and $\mathcal{Y}$, (3.12) has a global minimax point for any $\mu > 0$.

4 Theoretical analysis on exact continuous relaxations

In this section, we will consider the consistence of problem (1.1) and its continuous relaxation problem (3.12) with the density function ρ satisfying (3.1). Moreover, the smoothing approximation to a nonsmooth function c is defined and discussed in subsection 4.3.

4.1 Relations on saddle points

To proceed the discussion on the saddle points and local saddle points between problem (1.1) and its continuous relaxation model (3.12), we need the following Assumption 4.1, which will be discussed and verified in Section 5.

Assumption 4.1. *For a given $\mu > 0$, the following conditions hold.*

(i) *For any $\bar{\mathbf{y}} \in \mathcal{Y}$, if $\mathbf{x}^*$ is a local minimizer of $f^R(\mathbf{x}, \bar{\mathbf{y}}, \mu)$ on $\mathcal{X}$, then*

$$g_l(\mathbf{x}^*) \notin (0, \alpha\mu), \quad \forall l \in [\hat{n}]. \tag{4.1}$$

(ii) *For any* $\bar{\mathbf{x}} \in \mathcal{X}$*, if* $\mathbf{y}^*$ *is a local maximizer of* $f^R(\bar{\mathbf{x}}, \mathbf{y}, \mu)$ *on* $\mathcal{Y}$*, then*

$$h_k(\mathbf{y}^*) \notin (0, \alpha\mu), \quad \forall k \in [\hat{m}]. \tag{4.2}$$

If Assumption 4.1 holds for $\hat{\mu}$, then it holds for any $\mu \in (0, \hat{\mu}]$. Assumption 4.1 is to put the lower bound properties on the local solutions of $f^R(\mathbf{x}, \mathbf{y}, \mu)$ with respect to $\mathbf{x}$ and $\mathbf{y}$, respectively. If $(\mathbf{x}^*, \mathbf{y}^*)$ satisfies the lower bounds in (4.1) and (4.2), by (3.4), then

$$\phi_l^R(\mathbf{x}^*, \mu) = \phi_l(\mathbf{x}^*), \ \forall l \in [\hat{n}] \quad \text{and} \quad \psi_k^R(\mathbf{y}^*, \mu) = \psi_k(\mathbf{y}^*), \ \forall k \in [\hat{m}]. \tag{4.3}$$

Together this with the definition of (local) saddle points, we can find that the values of function $f(\cdot, \mu)$ and its continuous relaxation function $f^R(\cdot, \cdot, \mu)$ coincide at all (local) saddle points of (3.12), which is the main idea behind assuming the two lower bounds in Assumption 4.1 and the key motivation behind defining the continuous relaxation function as in (3.2). In Section 5, we will consider two ways to guarantee these two central properties in (4.1) and (4.2), one based on a first order necessary optimality condition and another based on a second order necessary optimality condition of (3.12). In what follows, we first derive the relations on the saddle points and local saddle points between problems (1.1) and (3.12) based on Assumption 4.1.

Theorem 4.1. *Suppose problem* (3.12) *satisfies Assumption 4.1, then*

(i) $(\mathbf{x}^*, \mathbf{y}^*)$ *is a saddle point of* (1.1) *if and only if it is a saddle point of* (3.12)*;*

(ii) $(\mathbf{x}^*, \mathbf{y}^*)$ *is a local saddle point of* (1.1)*, if it is a local saddle point of* (3.12)*.*

Proof. Suppose $(\mathbf{x}^*, \mathbf{y}^*)$ is a global (local) saddle point of problem (3.12). By the relations in (3.11) and (4.3), it holds $f^R(\mathbf{x}^*, \mathbf{y}^*, \mu) = f(\mathbf{x}^*, \mathbf{y}^*)$, $f^R(\mathbf{x}^*, \mathbf{y}, \mu) \geq f(\mathbf{x}^*, \mathbf{y})$ and $f^R(\mathbf{x}, \mathbf{y}^*, \mu) \leq f(\mathbf{x}, \mathbf{y}^*)$. Then, $(\mathbf{x}^*, \mathbf{y}^*)$ is a global (local) saddle point of problem (1.1). Thus we only need to prove that if $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point of problem (1.1), then it is a saddle point of (3.12).

Assume on contradiction that $(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point of problem (1.1), but it is not a saddle point of (3.12). Then

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}, \mathbf{y}^*) \quad \text{and} \quad \mathbf{y}^* \in \arg\max_{\mathbf{y} \in \mathcal{Y}} f(\mathbf{x}^*, \mathbf{y}), \tag{4.4}$$

but either $\mathbf{x}^*$ is not a global minimizer of $f^R(\mathbf{x}, \mathbf{y}^*, \mu)$ on $\mathcal{X}$ or $\mathbf{y}^*$ is not a global maximizer of $f^R(\mathbf{x}^*, \mathbf{y}, \mu)$ on $\mathcal{Y}$. As a possible situation, if $\mathbf{x}^*$ is not a global minimizer of $f^R(\mathbf{x}, \mathbf{y}^*, \mu)$ on $\mathcal{X}$, then there exists $\bar{\mathbf{x}} \in \arg\min_{\mathbf{x} \in \mathcal{X}} f^R(\mathbf{x}, \mathbf{y}^*, \mu)$ such that

$$f^R(\bar{\mathbf{x}}, \mathbf{y}^*, \mu) < f^R(\mathbf{x}^*, \mathbf{y}^*, \mu). \tag{4.5}$$

By (4.1) in Assumption 4.1, either $g_l(\bar{\mathbf{x}}) \geq \alpha\mu$ or $g_l(\bar{\mathbf{x}}) \leq 0$, which means $\phi_l^R(\bar{\mathbf{x}}, \mu) = \phi_l(\bar{\mathbf{x}})$, $\forall l \in [\hat{n}]$, and then

$$f^R(\bar{\mathbf{x}}, \mathbf{y}^*, \mu) = f(\bar{\mathbf{x}}, \mathbf{y}^*) - \lambda_2 \sum_{k \in [\hat{m}]} \psi_k^R(\mathbf{y}^*, \mu) + \lambda_2 \sum_{k \in [\hat{m}]} \psi_k(\mathbf{y}^*). \tag{4.6}$$

While by $\phi_l^R(\mathbf{x}^*,\mu) \le \phi_l(\mathbf{x}^*)$, $\forall l \in [\hat{n}]$, we obtain

$$f^R(\mathbf{x}^*,\mathbf{y}^*,\mu) \le f(\mathbf{x}^*,\mathbf{y}^*) - \lambda_2 \sum_{k\in[\hat{m}]} \psi_k^R(\mathbf{y}^*,\mu) + \lambda_2 \sum_{k\in[\hat{m}]} \psi_k(\mathbf{y}^*). \tag{4.7}$$

Combining (4.5)-(4.7), we find that $f(\bar{\mathbf{x}},\mathbf{y}^*) < f(\mathbf{x}^*,\mathbf{y}^*)$, which contradicts to the first relation in (4.4). Thus, $\mathbf{x}^*$ is a global minimizer of $f^R(\mathbf{x},\mathbf{y}^*,\mu)$ on $\mathcal{X}$. And we can verify that $\mathbf{y}^*$ is a global maximizer of $f^R(\mathbf{x}^*,\mathbf{y},\mu)$ on $\mathcal{Y}$ by a similar way. Therefore, $(\mathbf{x}^*,\mathbf{y}^*)$ is a saddle point of problem (3.12). □

Remark 4.1. *Following the proof of Theorem 4.1, besides the lower bound properties assumed in Assumption 4.1, we see that the property of continuous relaxation function r that $r(t,\mu) \le (t_+)^0$, $\forall t \in \mathbb{R}$ and $\mu > 0$, is used to guarantee the equivalence between the saddle points of (1.1) and (3.12) from sufficiency and necessity. Moreover, under Assumption 4.1, we confirm by Theorem 4.1 that any saddle point $(\mathbf{x}^*,\mathbf{y}^*)$ of (1.1), if it exists, satisfies the lower bounds in (4.1) and (4.2).*

4.2 Relations on minimax points

To establish the equivalent relation on global minimax points between problem (1.1) and problem (3.12), we need the following assumption, which will be discussed and verified in Section 5.

Assumption 4.2. *For a given $\mu > 0$, the following conditions hold.*

(i) *If $\mathbf{x}^*$ is a global minimizer of $\max_{\mathbf{y}\in\mathcal{Y}} f^R(\mathbf{x},\mathbf{y},\mu)$ on $\mathcal{X}$, then (4.1) holds.*

(ii) *For any $\bar{\mathbf{x}} \in \mathcal{X}$, if $\mathbf{y}^*$ is a global maximizer of $f^R(\bar{\mathbf{x}},\mathbf{y},\mu)$ on $\mathcal{Y}$, then (4.2) holds.*

Assumption 4.2 implies that any global minimax point $(\mathbf{x}^*,\mathbf{y}^*)$ of (3.12) satisfies the lower bounds in (4.1) and (4.2), and subsequentially the function values of $f(\cdot,\cdot)$ and its continuous relaxation function $f^R(\cdot,\cdot,\mu)$ coincide at all global minimax points of (3.12). Then, we can establish the following relations on the global minimax points between problems (1.1) and (3.12).

Theorem 4.2. *Under Assumption 4.2, $(\mathbf{x}^*,\mathbf{y}^*)$ is a global minimax point of problem (3.12) if and only if it is a global minimax point of problem (1.1).*

Proof. Let $(\mathbf{x}^*,\mathbf{y}^*)$ be a global minimax point of problem (3.12), i.e.

$$f^R(\mathbf{x}^*,\mathbf{y},\mu) \le f^R(\mathbf{x}^*,\mathbf{y}^*,\mu) \le \max_{\mathbf{y}'\in\mathcal{Y}} f^R(\mathbf{x},\mathbf{y}',\mu), \quad \forall\, \mathbf{x}\in\mathcal{X},\ \mathbf{y}\in\mathcal{Y}, \tag{4.8}$$

which implies $f^R(\mathbf{x}^*,\mathbf{y}^*,\mu) = \max_{\mathbf{y}'\in\mathcal{Y}} f^R(\mathbf{x}^*,\mathbf{y}',\mu) \le \max_{\mathbf{y}'\in\mathcal{Y}} f^R(\mathbf{x},\mathbf{y}',\mu)$. Then, Assumption 4.2 gives $\phi_l(\mathbf{x}^*) = \phi_l^R(\mathbf{x}^*,\mu)$ and $\psi_k(\mathbf{y}^*) = \psi_k^R(\mathbf{y}^*,\mu)$, and we further have

$$f^R(\mathbf{x}^*,\mathbf{y}^*,\mu) = f(\mathbf{x}^*,\mathbf{y}^*). \tag{4.9}$$

Invoking (3.11), we have

$$f(\mathbf{x}^*,\mathbf{y}) = c(\mathbf{x}^*,\mathbf{y}) + \lambda_1 \sum_{l\in[\hat{n}]} \phi_l(\mathbf{x}^*) - \lambda_2 \sum_{k\in[\hat{m}]} \psi_k(\mathbf{y}) \le f^R(\mathbf{x}^*,\mathbf{y},\mu), \quad \forall\, \mathbf{y}\in\mathcal{Y}. \tag{4.10}$$

For any $\mathbf{x} \in \mathcal{X}$, denote $\mathbf{y}_{\mathbf{x}}$ a maximizer of $f^R(\mathbf{x}, \mathbf{y}, \mu)$ on $\mathcal{Y}$. Recalling Assumption 4.2, we have that $\mathbf{y}_{\mathbf{x}}$ satisfies the lower bound property in (4.2) and $\psi_k(\mathbf{y}_{\mathbf{x}}) = \psi_k^R(\mathbf{y}_{\mathbf{x}}, \mu)$. Together it with (3.11), we have that

$$\begin{aligned}\max_{\mathbf{y}' \in \mathcal{Y}} f^R(\mathbf{x}, \mathbf{y}', \mu) =& c(\mathbf{x}, \mathbf{y}_{\mathbf{x}}) + \lambda_1 \sum_{l \in [\hat{n}]} \phi_l^R(\mathbf{x}, \mu) - \lambda_2 \sum_{k \in [\hat{m}]} \psi_k^R(\mathbf{y}_{\mathbf{x}}, \mu) \\ \leq& c(\mathbf{x}, \mathbf{y}_{\mathbf{x}}) + \lambda_1 \sum_{l \in [\hat{n}]} \phi_l(\mathbf{x}) - \lambda_2 \sum_{k \in [\hat{m}]} \psi_k(\mathbf{y}_{\mathbf{x}}) \leq \max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}').\end{aligned} \tag{4.11}$$

Thus, $(\mathbf{x}^*, \mathbf{y}^*)$ is a global minimax point of problem (1.1) by (4.8)-(4.11).

Conversely, let $(\mathbf{x}^*, \mathbf{y}^*)$ be a global minimax point of problem (1.1), i.e.

$$f(\mathbf{x}^*, \mathbf{y}) \leq f(\mathbf{x}^*, \mathbf{y}^*) \leq \max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}'), \quad \forall\, \mathbf{x} \in \mathcal{X},\, \mathbf{y} \in \mathcal{Y}. \tag{4.12}$$

By the first inequality in (4.12), similar to the proof of Theorem 4.1, we have

$$f^R(\mathbf{x}^*, \mathbf{y}, \mu) \leq f^R(\mathbf{x}^*, \mathbf{y}^*, \mu), \quad \forall\, \mathbf{y} \in \mathcal{Y}. \tag{4.13}$$

Next, recalling Assumption 4.2, $\mathbf{y}^* \in \arg\max_{\mathbf{y}' \in \mathcal{Y}} f^R(\mathbf{x}^*, \mathbf{y}', \mu)$ implies

$$\psi_k(\mathbf{y}^*) = \psi_k^R(\mathbf{y}^*, \mu), \quad \forall\, k \in [\hat{m}], \tag{4.14}$$

which together with $\phi_l^R(\mathbf{x}^*, \mu) \leq \phi_l(\mathbf{x}^*)$, $\forall l \in [\hat{n}]$ gives

$$f^R(\mathbf{x}^*, \mathbf{y}^*, \mu) \leq f(\mathbf{x}^*, \mathbf{y}^*). \tag{4.15}$$

Denote $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ a global minimax point of (3.12), then

$$f^R(\bar{\mathbf{x}}, \mathbf{y}, \mu) \leq f^R(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \mu) \leq \max_{\mathbf{y}' \in \mathcal{Y}} f^R(\mathbf{x}, \mathbf{y}', \mu), \quad \forall\, \mathbf{x} \in \mathcal{X},\, \mathbf{y} \in \mathcal{Y}, \tag{4.16}$$

by the first part of this theorem, which implies

$$f(\bar{\mathbf{x}}, \mathbf{y}) \leq f(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \leq \max_{\mathbf{y}' \in \mathcal{Y}} f(\mathbf{x}, \mathbf{y}'), \quad \forall\, \mathbf{x} \in \mathcal{X},\, \mathbf{y} \in \mathcal{Y}. \tag{4.17}$$

By Assumption 4.2, we further have

$$\phi_l^R(\bar{\mathbf{x}}, \mu) = \phi_l(\bar{\mathbf{x}}),\ \forall l \in [\hat{n}],\ \psi_k^R(\bar{\mathbf{y}}, \mu) = \psi_k(\bar{\mathbf{y}}),\ \forall k \in [\hat{m}] \text{ and } f^R(\bar{\mathbf{x}}, \bar{\mathbf{y}}, \mu) = f(\bar{\mathbf{x}}, \bar{\mathbf{y}}). \tag{4.18}$$

Letting $\mathbf{x} = \bar{\mathbf{x}}$ in the second inequality of (4.12), we have

$$f(\mathbf{x}^*, \mathbf{y}^*) \leq \max_{\mathbf{y}' \in \mathcal{Y}} f(\bar{\mathbf{x}}, \mathbf{y}'). \tag{4.19}$$

The first inequality in (4.17) gives $\max_{\mathbf{y}'\in\mathcal{Y}} f(\bar{\mathbf{x}},\mathbf{y}') = f(\bar{\mathbf{x}},\bar{\mathbf{y}})$, which together with (4.19) implies $f(\mathbf{x}^*,\mathbf{y}^*) \le f(\bar{\mathbf{x}},\bar{\mathbf{y}})$. This together with (4.15) and the third equality of (4.18) gives

$$f^R(\mathbf{x}^*,\mathbf{y}^*,\mu) \le f^R(\bar{\mathbf{x}},\bar{\mathbf{y}},\mu). \tag{4.20}$$

By virtue of (4.20) and the second inequality in (4.16), we have

$$f^R(\mathbf{x}^*,\mathbf{y}^*,\mu) \le \max_{\mathbf{y}'\in\mathcal{Y}} f^R(\mathbf{x},\mathbf{y}',\mu), \quad \forall\, \mathbf{x}\in\mathcal{X},$$

which together with (4.13) guarantees that $(\mathbf{x}^*,\mathbf{y}^*)$ is a global minimax point of problem (3.12). □

4.3 Smoothing functions to a nonsmooth convex-concave function c

If the function c in problem (3.12) is nonsmooth, the smoothing approximation of it is often needed in the algorithms [6, 11]. In what follows, we introduce a class of smoothing functions of c defined in [11].

Definition 4.1. *We call $\tilde{c}:\mathcal{X}\times\mathcal{Y}\times(0,1]\to\mathbb{R}$ a smoothing function of a nonsmooth function c on $\mathcal{X}\times\mathcal{Y}$, if $\tilde{c}(\cdot,\cdot,\varepsilon)$ is continuously differentiable on $\mathcal{X}\times\mathcal{Y}$ for any fixed $\varepsilon\in(0,1]$ and for any $(\bar{\mathbf{x}},\bar{\mathbf{y}})\in\mathcal{X}\times\mathcal{Y}$, it satisfies*

$$\lim_{\mathbf{x}\to\bar{\mathbf{x}},\mathbf{y}\to\bar{\mathbf{y}},\varepsilon\downarrow 0} \tilde{c}(\mathbf{x},\mathbf{y},\varepsilon) = c(\bar{\mathbf{x}},\bar{\mathbf{y}}). \tag{4.21}$$

The gradient consistence between the Clarke subgradient of the nonsmooth function and the gradients associated with its smoothing function sequence is important for the efficiency of the smoothing method, i.e. for any $\bar{\mathbf{x}}\in\mathcal{X}$ and $\bar{\mathbf{y}}\in\mathcal{Y}$,

$$\{\lim_{\mathbf{x}\to\bar{\mathbf{x}},\mathbf{y}\to\bar{\mathbf{y}},\varepsilon\downarrow 0} \nabla\tilde{c}(\mathbf{x},\mathbf{y},\varepsilon)\} \subseteq \partial c(\bar{\mathbf{x}},\bar{\mathbf{y}}). \tag{4.22}$$

The partial gradient consistences with respect to the update of two variables are often necessary for the algorithm analysis of the min-max problems, i.e. for any $\bar{\mathbf{x}}\in\mathcal{X}$ and $\bar{\mathbf{y}}\in\mathcal{Y}$,

$$\{\lim_{\mathbf{x}\to\bar{\mathbf{x}},\mathbf{y}\to\bar{\mathbf{y}},\varepsilon\downarrow 0} \nabla_{\mathbf{x}}\tilde{c}(\mathbf{x},\mathbf{y},\varepsilon)\} \subseteq \partial_{\mathbf{x}} c(\bar{\mathbf{x}},\bar{\mathbf{y}}), \tag{4.23}$$

$$\{\lim_{\mathbf{x}\to\bar{\mathbf{x}},\mathbf{y}\to\bar{\mathbf{y}},\varepsilon\downarrow 0} \nabla_{\mathbf{y}}\tilde{c}(\mathbf{x},\mathbf{y},\varepsilon)\} \subseteq \partial_{\mathbf{y}} c(\bar{\mathbf{x}},\bar{\mathbf{y}}). \tag{4.24}$$

However, neither $\partial c(\mathbf{x},\mathbf{y})$ nor $\partial_{\mathbf{x}} c(\mathbf{x},\mathbf{y})\times\partial_{\mathbf{y}} c(\mathbf{x},\mathbf{y})$ are contained in each other generally. See [14, Example 2.5.2]. When c is Clarke regular with respect to $(\mathbf{x},\mathbf{y})$, by [14, Proposition 2.3.15], it holds

$$\partial c(\mathbf{x},\mathbf{y}) \subseteq \partial_{\mathbf{x}} c(\mathbf{x},\mathbf{y})\times\partial_{\mathbf{y}} c(\mathbf{x},\mathbf{y}). \tag{4.25}$$

In what follows, we will show that (4.25) holds for any convex-concave function c, though the convexity-concavity of c cannot give the regularity of it. For example, $c(\mathbf{x},\mathbf{y}) = |\mathbf{x}|-|\mathbf{y}|$ is convex-concave on $\mathbb{R}\times\mathbb{R}$, but not Clarke regular in $(\mathbf{x},\mathbf{y})$ at $(1,0)$.

Proposition 4.1. *For any convex-concave function c, (4.25) holds for any $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$.*

Proof. Let $(\xi, \eta) \in \partial c(\mathbf{x}, \mathbf{y})$. We will prove that $\xi \in \partial_{\mathbf{x}} c(\mathbf{x}, \mathbf{y})$ and $\eta \in \partial_{\mathbf{y}} c(\mathbf{x}, \mathbf{y})$.

Since $c(\cdot, \mathbf{y})$ is convex on $\mathbb{R}^n$ for any $\mathbf{y}$, by [14, Proposition 2.5.3], it has $\xi \in \partial_{\mathbf{x}} c(\mathbf{x}, \mathbf{y})$. Inspired by the result in [14, Proposition 2.3.1], $(-\xi, -\eta) \in \partial(-c(\mathbf{x}, \mathbf{y}))$. Using the concavity of $c(\mathbf{x}, \cdot)$ on $\mathbb{R}^m$ for any $\mathbf{x} \in \mathbb{R}^n$, $-c(\mathbf{x}, \mathbf{y})$ is convex with respect to $\mathbf{y}$ and then $-\eta \in \partial_{\mathbf{y}}(-c(\mathbf{x}, \mathbf{y})) = -\partial_{\mathbf{y}} c(\mathbf{x}, \mathbf{y})$, which uses [14, Proposition 2.5.3] again. Thus, $\eta \in \partial_{\mathbf{y}} c(\mathbf{x}, \mathbf{y})$. □

For a nonsmooth function c, we can construct a smoothing function of c by convolution [11, 41, 44] as follows

$$\tilde{c}(\mathbf{z}, \varepsilon) = \int_{\mathbb{R}^{n+m}} c(\mathbf{z} - \mathbf{u}) \psi_\varepsilon(\mathbf{u}) d\mathbf{u}, \tag{4.26}$$

where $\mathbf{z} := (\mathbf{x}, \mathbf{y})$, $\psi_\varepsilon : \mathbb{R}^{n+m} \to \mathbb{R}_+$ is a sequence of bounded, measurable functions satisfying $\int_{\mathbb{R}^{n+m}} \psi_\varepsilon(\mathbf{u}) d\mathbf{u} = 1$ and $\lim_{\varepsilon \downarrow 0} B^\varepsilon = \{\mathbf{0}\}$ with $B^\varepsilon := \{\mathbf{u} : \psi_\varepsilon(\mathbf{u}) > 0\}$.

Proposition 4.2. *Let $\tilde{c} : \mathbb{R}^n \times \mathbb{R}^m \times (0, 1] \to \mathbb{R}$ be defined as in (4.26). Then $\tilde{c}$ is a smoothing function of c on $\mathcal{X} \times \mathcal{Y}$ and satisfies the following properties:*

(i) *for any $\bar{\mathbf{x}} \in \mathcal{X}$ and $\bar{\mathbf{y}} \in \mathcal{Y}$, (4.22) and (4.23)-(4.24) hold;*

(ii) *for any $\varepsilon > 0$, $\tilde{c}(\mathbf{x}, \mathbf{y}, \varepsilon)$ is convex in $\mathbf{x} \in \mathbb{R}^n$ and concave in $\mathbf{y} \in \mathbb{R}^m$.*

Proof. From [44, Theorem 9.67], $\tilde{c}$ in (4.26) is a smoothing function of c on $\mathcal{X} \times \mathcal{Y}$ and satisfies $\{\lim_{\mathbf{x} \to \bar{\mathbf{x}}, \mathbf{y} \to \bar{\mathbf{y}}, \varepsilon \downarrow 0} \nabla \tilde{c}(\mathbf{x}, \mathbf{y}, \varepsilon)\} \subseteq \partial c(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ for any $\bar{\mathbf{x}} \in \mathcal{X}$ and $\bar{\mathbf{y}} \in \mathcal{Y}$. Recalling the convexity-concavity of c, by Proposition 4.1, (4.23)-(4.24) hold.

In what follows, we first verify that $\tilde{c}(\mathbf{x}, \mathbf{y}, \varepsilon)$ is convex in $\mathbf{x} \in \mathbb{R}^n$ for any $\mathbf{y} \in \mathbb{R}^m$ and $\varepsilon > 0$. For any $\bar{\mathbf{x}}, \hat{\mathbf{x}} \in \mathcal{X}$ and $\eta \in [0, 1]$, observe that

$$\begin{aligned}
&\tilde{c}(\eta\bar{\mathbf{x}} + (1-\eta)\hat{\mathbf{x}}, \mathbf{y}, \varepsilon) = \int_{\mathbb{R}^{n+m}} c(\eta\bar{\mathbf{x}} + (1-\eta)\hat{\mathbf{x}} - \mathbf{v}, \mathbf{y} - \mathbf{w}) \psi_\varepsilon(\mathbf{u}) d\mathbf{u} \\
\leq &\eta \int_{\mathbb{R}^{n+m}} c(\bar{\mathbf{x}} - \mathbf{v}, \mathbf{y} - \mathbf{w}) \psi_\varepsilon(\mathbf{u}) d\mathbf{u} + (1-\eta) \int_{\mathbb{R}^{n+m}} c(\hat{\mathbf{x}} - \mathbf{v}, \mathbf{y} - \mathbf{w}) \psi_\varepsilon(\mathbf{u}) d\mathbf{u} \\
= &\eta\tilde{c}(\bar{\mathbf{x}}, \mathbf{y}, \varepsilon) + (1-\eta)\tilde{c}(\hat{\mathbf{x}}, \mathbf{y}, \varepsilon),
\end{aligned}$$

where $\mathbf{u} = (\mathbf{v}, \mathbf{w}) \in \mathbb{R}^{n+m}$, the inequality uses the convexity of $c(\cdot, \mathbf{y})$ and the nonnegativity of ψ_ε on $\mathbb{R}^{n+m}$. Thus, $\tilde{c}(\mathbf{x}, \mathbf{y}, \varepsilon)$ is convex in $\mathbf{x} \in \mathbb{R}^n$ for any $\mathbf{y} \in \mathbb{R}^m$ and $\varepsilon > 0$. By similar calculation, $\tilde{c}(\mathbf{x}, \mathbf{y}, \varepsilon)$ is concave in $\mathbf{y} \in \mathbb{R}^m$ for any $\mathbf{x} \in \mathbb{R}^n$ and $\varepsilon > 0$. □

Then, denote the smoothing model of (3.12) by

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \tilde{f}^R(\mathbf{x}, \mathbf{y}, \mu, \varepsilon) := \tilde{c}(\mathbf{x}, \mathbf{y}, \varepsilon) + \lambda_1 \sum_{l \in [\hat{n}]} \phi_l^R(\mathbf{x}, \mu) - \lambda_2 \sum_{k \in [\hat{m}]} \psi_k^R(\mathbf{y}, \mu), \tag{4.27}$$

where $\tilde{c}(\mathbf{x}, \mathbf{y}, \varepsilon)$ is a smoothing function of c. Since $\tilde{c}(\cdot, \cdot, \varepsilon)$ is a convex-concave function, it always has a saddle point over $\mathcal{X} \times \mathcal{Y}$. Following the results in previous sections,

problem (4.27) with an $\varepsilon > 0$ has a local saddle point and a global minimax point for some $\mu > 0$.

5 A particular case of problem (1.1)

In Section 4, we showed the consistency on the saddle point sets and the inclusion on the local saddle point sets of problems (1.1) and (3.12) under Assumption 4.1, and the consistence on their global minimax point sets under Assumption 4.2. In this section, we verify that Assumptions 4.1 and 4.2 hold for the continuous relaxation of a particular case of problem (1.1). Moreover, with the specific structure of r under Assumption 3.1 or Assumption 3.2, we establish the relations between the first order or second order stationary points of (3.12) and the local saddle points for the particular case of (1.1).

Denote

$$\begin{aligned} &\mathcal{N}_i := \{l \in [\hat{n}] : g_l(\mathbf{x}) := g_l(\mathbf{x}_i),\ \forall \mathbf{x} \in \mathbb{R}^n\}, && \text{for } i \in [n],\\ &\mathcal{M}_j := \{k \in [\hat{m}] : h_k(\mathbf{y}) := h_k(\mathbf{y}_j),\ \forall \mathbf{y} \in \mathbb{R}^m\}, && \text{for } j \in [m], \end{aligned}$$

and suppose

$$\bigcup_{i\in[n]} \mathcal{N}_i = [\hat{n}],\ \mathcal{N}_{\tilde{i}} \cap \mathcal{N}_{\hat{i}} = \emptyset, \forall \tilde{i} \neq \hat{i}; \quad \bigcup_{j\in[m]} \mathcal{M}_j = [\hat{m}],\ \mathcal{M}_{\tilde{j}} \cap \mathcal{M}_{\hat{j}} = \emptyset, \forall \tilde{j} \neq \hat{j},$$

which means that for any $l \in [\hat{n}]$, there is an $i \in [n]$ such that g_l is only dependent on $\mathbf{x}_i$, and for any $k \in [\hat{m}]$, there is a $j \in [m]$ such that h_k is only dependent on $\mathbf{y}_j$. In this section, we consider a case of (1.1) as follows

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x},\mathbf{y}) := c(\mathbf{x},\mathbf{y}) + \lambda_1 \sum_{i\in[n]} \sum_{l\in\mathcal{N}_i} (g_l(\mathbf{x}_i)_+)^0 - \lambda_2 \sum_{j\in[m]} \sum_{k\in\mathcal{M}_j} (h_k(\mathbf{y}_j)_+)^0. \quad (5.1)$$

Moreover, we assume that $\mathcal{X}$ and $\mathcal{Y}$ satisfy Slater's condition, i.e.

$$\mathrm{int}(\mathcal{X}) \neq \emptyset, \quad \mathrm{int}(\mathcal{Y}) \neq \emptyset, \quad (5.2)$$

and have the following structures

$$\mathcal{X} := \tilde{\mathcal{X}} \cap \bar{\mathcal{X}}, \quad \mathcal{Y} := \tilde{\mathcal{Y}} \cap \bar{\mathcal{Y}}, \quad (5.3)$$

where

$$\tilde{\mathcal{X}} = \{\mathbf{x} \in \mathbb{R}^n : \underline{\mathbf{u}} \leq \mathbf{x} \leq \overline{\mathbf{u}}\}, \quad \tilde{\mathcal{Y}} = \{\mathbf{y} \in \mathbb{R}^m : \underline{\mathbf{v}} \leq \mathbf{y} \leq \overline{\mathbf{v}}\} \quad (5.4)$$

with $\underline{\mathbf{u}}$, $\overline{\mathbf{u}} \in \mathbb{R}^n$, $\underline{\mathbf{v}}$, $\overline{\mathbf{v}} \in \mathbb{R}^m$, $\underline{\mathbf{u}} < \overline{\mathbf{u}}$, $\underline{\mathbf{v}} < \overline{\mathbf{v}}$, and

$$\bar{\mathcal{X}} = \{\mathbf{x} \in \mathbb{R}^n :\ u_t(\mathbf{x}) \leq 0,\ t \in [\bar{n}]\}, \quad \bar{\mathcal{Y}} = \{\mathbf{y} \in \mathbb{R}^m : v_s(\mathbf{y}) \leq 0,\ s \in [\bar{m}]\} \quad (5.5)$$

with Lipschitz continuous convex functions $u_t : \mathbb{R}^n \to \mathbb{R}$ for $t \in [\bar{n}]$ and $v_s : \mathbb{R}^m \to \mathbb{R}$ for $s \in [\bar{m}]$.

Denote

$$\mathcal{T}_0(\mathbf{x}) = \{t \in [\bar{n}] : u_t(\mathbf{x}) = 0\}, \quad \mathcal{S}_0(\mathbf{y}) = \{s \in [\bar{m}] : v_s(\mathbf{y}) = 0\}.$$

Since $\mathrm{int}(\bar{\mathcal{X}}) \supseteq \mathrm{int}(\mathcal{X}) \neq \emptyset$, by [20, Theorems 6.8.2 and 6.8.3] and [29, Proposition 5.3.1 and Remark 5.3.2],

$$N_{\bar{\mathcal{X}}}(\mathbf{x}) = \sum\nolimits_{t\in\mathcal{T}_0(\mathbf{x})}[0,+\infty)\partial u_t(\mathbf{x}), \quad \forall\, \mathbf{x} \in \bar{\mathcal{X}}. \tag{5.6}$$

Using $\mathrm{int}(\mathcal{X}) \neq \emptyset$ again, we have

$$N_{\mathcal{X}}(\mathbf{x}) = N_{\tilde{\mathcal{X}}}(\mathbf{x}) + N_{\bar{\mathcal{X}}}(\mathbf{x}). \tag{5.7}$$

Similar calculation can be put forward to $\mathcal{Y}$.

For any $i \in [n]$ and $j \in [m]$, denote

$$\phi_i(\mathbf{x}_i) = \sum_{l\in\mathcal{N}_i}(g_l(\mathbf{x}_i)_+)^0, \quad \psi_j(\mathbf{y}_j) = \sum_{k\in\mathcal{M}_j}(h_k(\mathbf{y}_j)_+)^0$$

and define their continuous relaxations by

$$\phi_i^R(\mathbf{x}_i,\mu) = \sum_{l\in\mathcal{N}_i} r(g_l(\mathbf{x}_i),\mu), \quad \psi_j^R(\mathbf{y}_j,\mu) = \sum_{k\in\mathcal{M}_j} r(h_k(\mathbf{y}_j),\mu) \tag{5.8}$$

with r in (3.2) and $\mu > 0$. We consider the continuous relaxation of (5.1) as follows

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} f^R(\mathbf{x},\mathbf{y},\mu) := c(\mathbf{x},\mathbf{y}) + \lambda_1\sum_{i\in[n]}\phi_i^R(\mathbf{x}_i,\mu) - \lambda_2\sum_{j\in[m]}\psi_j^R(\mathbf{y}_j,\mu). \tag{5.9}$$

We impose the following assumption on functions g and h related to the sets $\tilde{\mathcal{X}}$ and $\bar{\mathcal{X}}$ in (5.4)-(5.5).

Assumption 5.1. *There exist positive numbers τ and σ such that the following conditions hold.*

(i) *For any $\mathbf{x} \in \mathcal{X}$, if there exist $\hat{\imath} \in [n]$ and $\hat{l} \in \mathcal{N}_{\hat{\imath}}$ such that $g_{\hat{l}}(\mathbf{x}_{\hat{\imath}}) \in (0,\tau)$, then*

$$|g'_{\hat{l}}(\mathbf{x}_{\hat{\imath}})| \geq \sigma, \quad \mathbf{x}_{\hat{\imath}} \in \mathrm{int}(\tilde{\mathcal{X}}_{\hat{\imath}}), \tag{5.10}$$

$$g_l(\mathbf{x}_{\hat{\imath}}) \notin [0,\tau],\ \forall l \in \mathcal{N}_{\hat{\imath}},\ l \neq \hat{l}, \tag{5.11}$$

$$\xi^t_{\hat{\imath}}(\mathbf{x})g'_{\hat{l}}(\mathbf{x}_{\hat{\imath}}) \geq 0,\ \forall \xi^t(\mathbf{x}) \in \partial u_t(\mathbf{x}),\ t \in \mathcal{T}_0(\mathbf{x}). \tag{5.12}$$

(ii) *For any $\mathbf{y} \in \mathcal{Y}$, if there exist $\hat{\jmath} \in [m]$ and $\hat{k} \in \mathcal{M}_{\hat{\jmath}}$ such that $h_{\hat{k}}(\mathbf{y}_{\hat{\jmath}}) \in (0,\tau)$, then*

$$|h'_{\hat{k}}(\mathbf{y}_{\hat{\jmath}})| \geq \sigma, \quad \mathbf{y}_{\hat{\jmath}} \in \mathrm{int}(\tilde{\mathcal{Y}}_{\hat{\jmath}}), \tag{5.13}$$

$$h_k(\mathbf{y}_{\hat{\jmath}}) \notin [0,\tau],\ \forall k \in \mathcal{M}_{\hat{\jmath}},\ k \neq \hat{k}, \tag{5.14}$$

$$\eta_{\hat{j}}^s(\mathbf{y})h'_{\hat{k}}(\mathbf{y}_{\hat{j}}) \geq 0,\ \forall \eta^s(\mathbf{y}) \in \partial v_s(\mathbf{y}),\ s \in \mathcal{S}_0(\mathbf{y}). \tag{5.15}$$

When $\alpha\mu < \tau$, (5.11) is used to guarantee that for any $\mathbf{x} \in \mathcal{X}$ and $i \in [n]$, there is at most one $l \in \mathcal{N}_i$ such that $r(g_l(\mathbf{x}_i), \mu) \neq (g_l(\mathbf{x}_i)_+)^0$. If there exist $\hat{\imath} \in [n]$ and $\hat{l} \in \mathcal{N}_{\hat{\imath}}$ such that $r(g_{\hat{l}}(\mathbf{x}_{\hat{\imath}}), \mu) \neq (g_{\hat{l}}(\mathbf{x}_{\hat{\imath}})_+)^0$, by (5.10) and Assumption 3.1, if $r(\cdot, \mu)$ is Lipschitz continuous around $g_{\hat{l}}(\mathbf{x}_{\hat{\imath}})$, we obtain

$$\inf\{\xi : \xi \in \partial_{\mathbf{x}_{\hat{\imath}}} r(g_{\hat{l}}(\mathbf{x}_{\hat{\imath}}), \mu)\} \geq \underline{\rho}\sigma/\mu, \tag{5.16}$$

and by (5.6), (5.7) and (5.10), $[N_{\mathcal{X}}(\mathbf{x})]_{\hat{\imath}} = 0$ or $[N_{\mathcal{X}}(\mathbf{x})]_{\hat{\imath}} = [N_{\bar{\mathcal{X}}}(\mathbf{x})]_{\hat{\imath}} = [\sum_{t\in\mathcal{T}_0(\mathbf{x})}[0, +\infty)\partial u_t(\mathbf{x})]_{\hat{\imath}}$, which together with (5.12) implies that

$$\xi_{\hat{\imath}}(\mathbf{x})g'_{\hat{l}}(\mathbf{x}_{\hat{\imath}}) \geq 0, \quad \forall\, \xi(\mathbf{x}) \in N_{\mathcal{X}}(\mathbf{x}). \tag{5.17}$$

In particular, if $\mathcal{N}_i = \{i\}$ and $\mathcal{M}_j = \{j\}$ for all $i \in [n]$ and $j \in [m]$, then (5.1) reduces to

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) := c(\mathbf{x}, \mathbf{y}) + \lambda_1 \sum_{i=1}^{n} (g_i(\mathbf{x}_i)_+)^0 - \lambda_2 \sum_{j=1}^{m} (h_j(\mathbf{y}_j)_+)^0. \tag{5.18}$$

Remark 5.1. *Consider*

$$f(\mathbf{x}, \mathbf{y}) := c(\mathbf{x}, \mathbf{y}) + \lambda_1\|(\mathbf{x} - \underline{\mathbf{a}})_+\|_0 + \lambda_1\|(\overline{\mathbf{a}} - \mathbf{x})_+\|_0 - \lambda_2\|(\mathbf{y} - \underline{\mathbf{b}})_+\|_0 - \lambda_2\|(\overline{\mathbf{b}} - \mathbf{y})_+\|_0, \tag{5.19}$$

with $\underline{\mathbf{a}}$, $\overline{\mathbf{a}} \in \mathbb{R}^n$ *and* $\underline{\mathbf{b}}$, $\overline{\mathbf{b}} \in \mathbb{R}^m$. *Then* (5.19) *is a special case of problem* (5.1) *with*

$$g_i(\mathbf{x}_i) = \mathbf{x}_i - \underline{\mathbf{a}}_i,\ g_{n+i}(\mathbf{x}_i) = \overline{\mathbf{a}}_i - \mathbf{x}_i,\ i \in [n];\ h_j(\mathbf{y}_j) = \mathbf{y}_j - \underline{\mathbf{b}}_j,\ h_{m+j}(\mathbf{y}_j) = \overline{\mathbf{b}}_j - \mathbf{y}_j,\ j \in [m].$$

In particular, the following three cases satisfy Assumption 5.1-(i), while the judgment on Assumption 5.1-(ii) is the same.

- Case 1: *Let* $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \underline{\mathbf{u}} \leq \mathbf{x} \leq \overline{\mathbf{u}}\}$ *and* f *be defined as in* (5.19) *with* $\overline{\mathbf{a}}$, $\underline{\mathbf{a}} \in \mathcal{X}$. *Then, Assumption 5.1-(i) holds with*

$$\sigma = 1 \quad \textit{and} \quad \tau = \min\{\tau_{\mathbf{x},1}, \tau_{\mathbf{x},2}\}, \tag{5.20}$$

where

$$\begin{aligned}
\tau_{\mathbf{x},1} &= \min\{1, \overline{\mathbf{u}}_i - \underline{\mathbf{a}}_i, \overline{\mathbf{a}}_i - \underline{\mathbf{u}}_i : \overline{\mathbf{u}}_i > \underline{\mathbf{a}}_i, \overline{\mathbf{a}}_i > \underline{\mathbf{u}}_i, i \in [n]\},\\
\tau_{\mathbf{x},2} &= \min\{1, |\underline{\mathbf{a}}_i - \overline{\mathbf{a}}_i|/2 : \underline{\mathbf{a}}_i \neq \overline{\mathbf{a}}_i,\ i \in [n]\}.
\end{aligned}$$

- Case 2: *Let* f *be defined as in* (5.19) *with* $\underline{\mathbf{a}} \geq \mathbf{0}$, $\overline{\mathbf{a}} \leq \mathbf{0}$, $\underline{\mathbf{b}} \geq \mathbf{0}$, $\overline{\mathbf{b}} \leq \mathbf{0}$, $\tilde{\mathcal{X}} = \{\mathbf{x} \in \mathbb{R}^n : \underline{\mathbf{u}} \leq \mathbf{x} \leq \overline{\mathbf{u}}\}$ *and* $\bar{\mathcal{X}}$ *be*

$$\bar{\mathcal{X}} = \{\mathbf{x} :\ \|\mathbf{x}\|_2 \leq \delta\} \quad \textit{or} \quad \bar{\mathcal{X}} = \{\mathbf{x} :\ \|\mathbf{x}\|_1 \leq \delta\}$$

with $\delta > 0$. *Then Assumption 5.1-(i) also holds with* σ *and* τ *in* (5.20).

- Case 3: *Let* $\tilde{\mathcal{X}} = \{\mathbf{x} \in \mathbb{R}^n : \underline{\mathbf{u}} \le \mathbf{x} \le \overline{\mathbf{u}}\}$, $\bar{\mathcal{X}} = \{\mathbf{x} : \boldsymbol{A}\mathbf{x} \le \mathbf{c}\}$ *with* $\boldsymbol{A} \in \mathbb{R}_+^{r_1 \times n}$, $\mathbf{c} \in \mathbb{R}^{r_1}$ *and* f *be specialized to*

$$f(\mathbf{x}, \mathbf{y}) := c(\mathbf{x}, \mathbf{y}) + \lambda_1 \|(\mathbf{x} - \underline{\mathbf{a}})_+\|_0 - \lambda_2 \|(\mathbf{y} - \underline{\mathbf{b}})_+\|_0$$

with $\underline{\mathbf{a}}_i \in [\underline{\mathbf{u}}_i, \overline{\mathbf{u}}_i)$ *and* $\underline{\mathbf{b}}_j \in [\underline{\mathbf{v}}_j, \overline{\mathbf{v}}_j)$ *for* $i \in [n]$ *and* $j \in [m]$. *Then Assumption 5.1-(i) also holds with* $\sigma = 1$ *and* $\tau = \min\{1, \overline{\mathbf{u}}_i - \underline{\mathbf{a}}_i : \overline{\mathbf{u}}_i > \underline{\mathbf{a}}_i, i \in [n]\}$.

In particular, case 2 in Remark 5.1 indicates that problem (1.3) with $\mathcal{X} = \{\mathbf{x} : \underline{\mathbf{u}} \le \mathbf{x} \le \overline{\mathbf{u}}, \|\mathbf{x}\|_1 \le 1\}$ and $\mathcal{Y} = \{\mathbf{y} : \underline{\mathbf{v}} \le \mathbf{x} \le \overline{\mathbf{v}}, \|\mathbf{y}\|_1 \le 1\}$ satisfies Assumption 5.1, and the problems in case 3 satisfying Assumption 5.1 include

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) := c(\mathbf{x}, \mathbf{y}) + \lambda_1 \|\mathbf{x}_+\|_0 - \lambda_2 \|\mathbf{y}_+\|_0$$

with $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_\infty \le 1, \mathbf{e}^\top \mathbf{x} \le 1\}$ and $\mathcal{Y} = \{\mathbf{y} : \|\mathbf{y}\|_\infty \le 1, \mathbf{e}^\top \mathbf{y} \le 1\}$ as a special case. Moreover, (5.1) satisfying Assumption 5.1 is not limited to problem (5.19). For example, the following problem

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f(\mathbf{x}, \mathbf{y}) := c(\mathbf{x}, \mathbf{y}) + \lambda_1 \|\mathbf{x}_+\|_0 - \lambda_2 \sum\nolimits_{j\in[m]} ((\sin(4\mathbf{y}_j))_+)^0$$

with $\mathcal{X} = \{\mathbf{x} : \|\mathbf{x}\|_\infty \le 5\}$ and $\mathcal{Y} = \{\mathbf{y} : \|\mathbf{y}\|_\infty \le 5\}$ also satisfies Assumption 5.1 with $\tau = \frac{1}{2}$ and $\sigma = 1$.

5.1 Density function ρ under Assumption 3.1

In this subsection, we will show that when problem (5.1) satisfies Assumption 5.1 and density function ρ satisfies Assumption 3.1, Assumptions 4.1 and 4.2 hold for the continuous relaxation of (5.1) formulated by (5.9). Moreover, we need assume that ρ has an upper bound on its support set, i.e. there exists $\bar{\rho} > 0$ such that

$$\rho(s) \le \bar{\rho}, \quad \forall s \in (0, \alpha). \tag{5.21}$$

Here, the density function ρ in Example 3.1 with $p = 1$ satisfies this condition with $\bar{\rho} = 1$. In this situation, $r(\cdot, \mu)$ is Lipschitz continuous on $\mathbb{R}$, which implies $f^R(\cdot, \cdot, \mu)$ is Lipschitz continuous on $\mathbb{R}^n \times \mathbb{R}^m$ for any fixed $\mu > 0$. Moreover, when the density function ρ is as in Example 3.1 with $p = 1$, we will give more discussion on the weak-d stationary points of (5.9) and its smoothing version in (4.27).

5.1.1 Relations on saddle points and minimax points

From the boundedness of $\mathcal{X}$ and $\mathcal{Y}$, there exists a positive constant $L_{c,1}$ such that for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$, it holds

$$\|\partial_{\mathbf{x}} c(\mathbf{x}, \mathbf{y})\|_\infty \le L_{c,1} \quad \text{and} \quad \|\partial_{\mathbf{y}} c(\mathbf{x}, \mathbf{y})\|_\infty \le L_{c,1}. \tag{5.22}$$

For a given $\mu \in \mathbb{R}_{++}$ and $\bar{\mathbf{y}} \in \mathcal{Y}$, if $\mathbf{x}^*$ is a local solution of $\min_{\mathbf{x}\in\mathcal{X}} f^R(\mathbf{x}, \bar{\mathbf{y}}, \mu)$, then

$$\mathbf{0} \in \partial_{\mathbf{x}} f^R(\mathbf{x}^*, \bar{\mathbf{y}}, \mu) + N_{\mathcal{X}}(\mathbf{x}^*). \tag{5.23}$$

Similarly, for $\bar{\mathbf{x}} \in \mathcal{X}$, if $\mathbf{y}^*$ is a local solution of $\max_{\mathbf{y}\in\mathcal{Y}} f^R(\bar{\mathbf{x}}, \mathbf{y}, \mu)$, then

$$\mathbf{0} \in -\partial_{\mathbf{y}} f^R(\bar{\mathbf{x}}, \mathbf{y}^*, \mu) + N_{\mathcal{Y}}(\mathbf{y}^*). \tag{5.24}$$

For $i \in [n]$ and $j \in [m]$, by [14, Proposition 2.3.9], we have

$$\begin{aligned} \partial_{\mathbf{x}_i} \phi_i^R(\mathbf{x}_i, \mu) &\subseteq \tilde{\partial}_{\mathbf{x}_i} \phi_i^R(\mathbf{x}_i, \mu) := \sum\nolimits_{l\in\mathcal{N}_i} \partial_t r(t, \mu)_{t=g_l(\mathbf{x}_i)} g_l'(\mathbf{x}_i), \\ \partial_{\mathbf{y}_j} \psi_j^R(\mathbf{y}_j, \mu) &\subseteq \tilde{\partial}_{\mathbf{y}_j} \psi_j^R(\mathbf{y}_j, \mu) := \sum\nolimits_{k\in\mathcal{M}_j} \partial_t r(t, \mu)_{t=h_k(\mathbf{y}_j)} h_k'(\mathbf{y}_j). \end{aligned} \tag{5.25}$$

By [14, Corollary 2] and recalling (5.25), one has

$$\partial_{\mathbf{x}} f^R(\mathbf{x}^*, \bar{\mathbf{y}}, \mu) \subseteq \tilde{\partial}_{\mathbf{x}} f^R(\mathbf{x}^*, \bar{\mathbf{y}}, \mu) := \partial_{\mathbf{x}} c(\mathbf{x}^*, \bar{\mathbf{y}}) + \lambda_1 \textstyle\sum_{i=1}^n \tilde{\partial}_{\mathbf{x}_i} \phi_i^R(\mathbf{x}_i^*, \mu)\mathbf{e}_i, \tag{5.26}$$

$$\partial_{\mathbf{y}} f^R(\bar{\mathbf{x}}, \mathbf{y}^*, \mu) \subseteq \tilde{\partial}_{\mathbf{y}} f^R(\bar{\mathbf{x}}, \mathbf{y}^*, \mu) := \partial_{\mathbf{y}} c(\bar{\mathbf{x}}, \mathbf{y}^*) - \lambda_2 \textstyle\sum_{j=1}^m \tilde{\partial}_{\mathbf{y}_j} \psi_j^R(\mathbf{y}_j^*, \mu)\mathbf{e}_j. \tag{5.27}$$

Combining (5.23) with (5.26), and (5.24) with (5.27), we obtain that

- if $\mathbf{x}^*$ is a local solution of $\min_{\mathbf{x}\in\mathcal{X}} f^R(\mathbf{x}, \bar{\mathbf{y}}, \mu)$, then

$$\mathbf{0} \in \tilde{\partial}_{\mathbf{x}} f^R(\mathbf{x}^*, \bar{\mathbf{y}}, \mu) + N_{\mathcal{X}}(\mathbf{x}^*); \tag{5.28}$$

- if $\mathbf{y}^*$ is a local solution of $\max_{\mathbf{y}\in\mathcal{Y}} f^R(\bar{\mathbf{x}}, \mathbf{y}, \mu)$, then

$$\mathbf{0} \in -\tilde{\partial}_{\mathbf{y}} f^R(\bar{\mathbf{x}}, \mathbf{y}^*, \mu) + N_{\mathcal{Y}}(\mathbf{y}^*). \tag{5.29}$$

If $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ is a local saddle point of (5.9), then (5.28) and (5.29) hold at $\bar{\mathbf{x}} = \mathbf{x}^*$ and $\bar{\mathbf{y}} = \mathbf{y}^*$.

In the rest of this paper, we denote $\lambda = \min\{\lambda_1, \lambda_2\}$ and set

$$\bar{\mu}_1 = \min\left\{\frac{\tau}{\alpha}, \frac{\lambda\sigma\underline{\rho}}{L_{c,1}}\right\} \tag{5.30}$$

with τ, σ in Assumption 5.1, α in (3.1), $\underline{\rho}$ in Assumption 3.1, and $L_{c,1}$ a constant satisfying the two inequalities in (5.22). Here, $0 < \mu < \bar{\mu}_1 \le \tau/\alpha$ gives $\mu\alpha < \tau$, which together with (3.4) implies

$$\begin{aligned} &\text{if } g_l(\mathbf{x}_i) \notin (0, \tau), \text{ then } r(g_l(\mathbf{x}_i), \mu) = (g_l(\mathbf{x}_i)_+)^0, \ \forall l \in \mathcal{N}_i,\ i \in [n]; \\ &\text{if } h_k(\mathbf{y}_j) \notin (0, \tau), \text{ then } r(h_k(\mathbf{y}_j), \mu) = (h_k(\mathbf{y}_j)_+)^0, \ \forall k \in \mathcal{M}_j,\ j \in [m]. \end{aligned}$$

Next, we will derive the lower bounds in (4.1) and (4.2) based on (5.28) and (5.29), respectively.

Proposition 5.1. *Suppose problem (5.1) satisfies Assumption 5.1. When density function ρ satisfies Assumption 3.1 and $0 < \mu < \bar{\mu}_1$ with $\bar{\mu}_1$ defined in (5.30), then the continuous relaxation model in (5.9) owns the following properties:*

$$\text{if (5.28) holds at } (\mathbf{x}^*, \bar{\mathbf{y}}) \in \mathcal{X} \times \mathcal{Y}, \text{ then } g_l(\mathbf{x}_i^*) \notin (0, \alpha\mu),\ \forall l \in \mathcal{N}_i,\ i \in [n], \tag{5.31}$$

$$\text{if (5.29) holds at } (\bar{\mathbf{x}}, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}, \text{ then } h_k(\mathbf{y}_j^*) \notin (0, \alpha\mu),\ \forall k \in \mathcal{M}_j,\ j \in [m]. \tag{5.32}$$

Proof. We argue the above statements by contradiction.

If there exist $\tilde{i} \in [n]$ and $\tilde{l} \in \mathcal{N}_{\tilde{i}}$ such that $0 < g_{\tilde{l}}(\mathbf{x}_{\tilde{i}}^*) < \alpha\mu$, by $\alpha\mu < \alpha\bar{\mu}_1 \leq \tau$, Assumption 5.1, (5.25) and (5.26), we obtain $\mathbf{x}_{\tilde{i}}^* \in \mathrm{int}(\tilde{\mathcal{X}}_{\tilde{i}})$, and

$$[\tilde{\partial}_{\mathbf{x}} f^R(\mathbf{x}^*, \bar{\mathbf{y}}, \mu)]_{\tilde{i}} = [\partial_{\mathbf{x}} c(\mathbf{x}^*, \bar{\mathbf{y}})]_{\tilde{i}} + \lambda_1 \partial_{\mathbf{x}_{\tilde{i}}} r(g_{\tilde{l}}(\mathbf{x}_{\tilde{i}}^*), \mu).$$

From Assumption 5.1-(i), (5.16), (5.17) and (5.28), we further have that

$$\lambda_1 \sigma \underline{\rho} / \mu \leq L_{c,1}, \tag{5.33}$$

which contradicts $\mu < \bar{\mu}_1 \leq \lambda_1 \sigma \underline{\rho} / L_{c,1}$. Thus, (5.31) holds. Similar analysis can be derived to (5.32). □

Thus, under the conditions in Proposition 5.1, Assumption 4.1 holds naturally for any $\mu \in (0, \bar{\mu}_1)$ with $\bar{\mu}_1$ defined in (5.30). In what follows, we will verify Assumption 4.2 in this situation.

Proposition 5.2. *Suppose problem (5.1) satisfies Assumption 5.1. Then Assumption 4.2 holds for (5.9) when density function ρ satisfies Assumption 3.1 and $0 < \mu < \bar{\mu}_1$ with $\bar{\mu}_1$ in (5.30). Moreover, all global minimax points of (5.9) satisfy the lower bounds in (4.1) and (4.2).*

Proof. For $\bar{\mathbf{x}} \in \mathcal{X}$, if $\mathbf{y}^*$ is a global maximizer of $f^R(\bar{\mathbf{x}}, \mathbf{y}, \mu)$ on $\mathcal{Y}$, then (5.29) holds. By Proposition 5.1, we have $h_k(\mathbf{y}_j^*) \notin (0, \alpha\mu)$, $\forall\, k \in \mathcal{M}_j$, $j \in [m]$, which means Assumption 4.2-(ii) holds. Now we prove Assumption 4.2-(i) holds. Denote

$$f_1^R(\mathbf{x}, \mathbf{y}, \mu) := c(\mathbf{x}, \mathbf{y}) - \lambda_2 \sum_{j \in [m]} \psi_j^R(\mathbf{y}_j, \mu), \quad \vartheta(\mathbf{x}, \mu) := \max_{\mathbf{y}' \in \mathcal{Y}} f_1^R(\mathbf{x}, \mathbf{y}', \mu).$$

For an $\mathbf{x} \in \mathcal{X}$, let $\mathbf{y}_{\mathbf{x}}$ be a maximizer of $f_1^R(\mathbf{x}, \mathbf{y}, \mu)$ on $\mathcal{Y}$, then it is also a maximizer of $f^R(\mathbf{x}, \mathbf{y}, \mu)$ on $\mathcal{Y}$. For any $\tilde{\mathbf{x}}$, $\hat{\mathbf{x}} \in \mathcal{X}$, if $\vartheta(\tilde{\mathbf{x}}, \mu) \geq \vartheta(\hat{\mathbf{x}}, \mu)$, then

$$\vartheta(\tilde{\mathbf{x}}, \mu) - \vartheta(\hat{\mathbf{x}}, \mu) \leq f_1^R(\tilde{\mathbf{x}}, \mathbf{y}_{\tilde{\mathbf{x}}}, \mu) - f_1^R(\hat{\mathbf{x}}, \mathbf{y}_{\tilde{\mathbf{x}}}, \mu) \leq L_{c,1} \|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\|.$$

Similarly, if $\vartheta(\tilde{\mathbf{x}}, \mu) < \vartheta(\hat{\mathbf{x}}, \mu)$,

$$\vartheta(\hat{\mathbf{x}}, \mu) - \vartheta(\tilde{\mathbf{x}}, \mu) \leq f_1^R(\hat{\mathbf{x}}, \mathbf{y}_{\hat{\mathbf{x}}}, \mu) - f_1^R(\tilde{\mathbf{x}}, \mathbf{y}_{\hat{\mathbf{x}}}, \mu) \leq L_{c,1} \|\tilde{\mathbf{x}} - \hat{\mathbf{x}}\|.$$

Thus, $\vartheta(\cdot, \mu)$ is Lipschitz continuous on $\mathcal{X}$ with constant $L_{c,1}$.

Let $\mathbf{x}^*$ be a global minimizer of $\vartheta(\mathbf{x},\mu) + \lambda_1 \sum_{i\in[n]} \phi_i^R(\mathbf{x}_i,\mu) = \max_{\mathbf{y}'\in\mathcal{Y}} f^R(\mathbf{x},\mathbf{y}',\mu)$ on $\mathcal{X}$. The first order necessary optimality condition gives

$$0 \in [\partial_{\mathbf{x}}\vartheta(\mathbf{x}^*,\mu)]_i + \lambda_1 \partial_{\mathbf{x}_i}\phi_i^R(\mathbf{x}_i^*,\mu) + [N_{\mathcal{X}}(\mathbf{x}^*)]_i, \quad \forall i \in [n]. \tag{5.34}$$

Assume there exist $\tilde{i} \in [n]$ and $\hat{l} \in \mathcal{N}_{\tilde{i}}$ such that $0 < g_{\hat{l}}(\mathbf{x}_{\tilde{i}}^*) < \alpha\mu$. Similar to the derivation in Proposition 5.1, we also obtain (5.33) and a contradiction to the value of μ. Thus, for all $i \in [n]$ and $l \in \mathcal{N}_i$, $g_l(\mathbf{x}_i^*) \notin (0, \alpha\mu)$, which together with the above analysis gives that any gloabl minimax point of (5.9) satisfies the lower bounds in (4.1) and (4.2). Hence Assumption 4.2-(i) holds. □

Similar to the proof of Proposition 5.2, we can show that all local minimax points of (5.9) satisfy the lower bounds in (4.1) and (4.2). Moreover, similar to the proof of Theorem 4.2, we can have the relation on the local minimax points between problems (5.1) and (5.9). Combining this with the above discussion, we conclude the relations on (5.1) and (5.9) in the following theorem.

Theorem 5.1. *Suppose problem (5.1) satisfies Assumption 5.1, density function ρ satisfies Assumption 3.1 and $0 < \mu < \bar{\mu}_1$ with $\bar{\mu}_1$ defined in (5.30), then the following statements hold:*

(i) *$(\mathbf{x}^*, \mathbf{y}^*)$ is a saddle point (global minimax point) of problem (5.1) if and only if it is a saddle point (global minimax point) of (5.9);*

(ii) *$(\mathbf{x}^*, \mathbf{y}^*)$ is an $\alpha\mu$-strong local saddle point of (5.1), if it is a local saddle point of problem (5.9);*

(iii) *$(\mathbf{x}^*, \mathbf{y}^*)$ is a local minimax point of (5.1), if it is a local minimax point of problem (5.9).*

5.1.2 Stationary points of (5.9) with ρ in Example 3.1 with $p = 1$

In this subsection, we focus on the relations of (5.1) and (5.9) when the density function ρ is defined as in Example 3.1 with $p = 1$, which makes the corresponding function r satisfy Assumption 3.1 with $\alpha = \underline{\rho} = 1$ and (5.21) with $\overline{\rho} = 1$. Theorem 5.1 has established the relations on the (local) saddle points and global minimax points between problems (5.1) and (5.9). In this subsection, we suppose problem (5.1) satisfies Assumption 5.1, and functions g_l, h_k in (5.1) are convex for all $l \in \mathcal{N}_i$, $i \in [n]$ and $k \in \mathcal{M}_j$, $j \in [m]$. We will study the relations on a class of stationary points of (5.9) with the μ-strong local saddle points of (5.1) in what follows.

For a locally Lipschitz continuous function $\varphi : \mathbb{R}^n \to \mathbb{R}$, the generalized (Clarke) directional derivative [14] of φ at point $\mathbf{x}$ in direction $\mathbf{v}$ is well-defined, i.e.

$$\varphi^\circ(\mathbf{x},\mathbf{v}) = \limsup_{\mathbf{z}\to\mathbf{x}, t\downarrow 0} \frac{\varphi(\mathbf{z}+t\mathbf{v}) - \varphi(\mathbf{z})}{t}.$$

Function φ is said to be Bouligand-differentiable (B-differentiable) at $\mathbf{x}$, if φ is locally Lipschitz continuous around $\mathbf{x}$ and directionally differentiable at $\mathbf{x}$, i.e. for any $\mathbf{v} \in \mathbb{R}^n$,

$$\varphi'(\mathbf{x}, \mathbf{v}) = \limsup_{t \downarrow 0} \frac{\varphi(\mathbf{x} + t\mathbf{v}) - \varphi(\mathbf{x})}{t} \quad \text{exists.}$$

It is well-known that $\varphi^\circ(\mathbf{x}, \mathbf{v}) \geq \varphi'(\mathbf{x}, \mathbf{v})$ in general and these two directional derivatives are the same if function φ is (Clarke) regular [14]. However, most nonconvex functions are not regular and a nonsmooth nonconvex function is not always directionally differentiable. Notice that convex functions and differentiable functions are directionally differentiable, then a DC (difference-of-convex) function is directionally differentiable [44], where we call function φ a DC function, if it can be formulated by the difference of two convex functions. This promotes some kinds of stationary points for the DC programming [39], such as the d(irectional)-stationary point and the weak d-stationary point, both of which are generally stronger than the Clarke stationary point.

Note that $r(t, \mu)$ in Example 3.1 with $p = 1$ can be expressed by the following DC function

$$r(t, \mu) = t_+/\mu - (t - \mu)_+/\mu. \tag{5.35}$$

From the definitions of ϕ_i^R and ψ_j^R in (5.8), the objective function in (5.9) has the formulation of

$$\begin{aligned} f^R(\mathbf{x}, \mathbf{y}, \mu) = c(\mathbf{x}, \mathbf{y}) &+ \lambda_1 \sum_{i \in [n]} \sum_{l \in \mathcal{N}_i} (g_l(\mathbf{x}_i)_+/\mu - (g_l(\mathbf{x}_i) - \mu)_+/\mu) \\ &- \lambda_2 \sum_{j \in [m]} \sum_{k \in \mathcal{M}_j} (h_k(\mathbf{y}_j)_+/\mu - (h_k(\mathbf{y}_j) - \mu)_+/\mu)\,. \end{aligned} \tag{5.36}$$

For fixed $\mathbf{x}^* \in \mathcal{X}$, $\mathbf{y}^* \in \mathcal{Y}$ and $\mu \in \mathbb{R}_{++}$, denote

$$f^R_{\mathbf{y}^*,\mu}(\mathbf{x}) \triangleq f^R(\mathbf{x}, \mathbf{y}^*, \mu), \quad f^R_{\mathbf{x}^*,\mu}(\mathbf{y}) \triangleq f^R(\mathbf{x}^*, \mathbf{y}, \mu),$$

and consider the following two optimization problems

$$\min_{\mathbf{x} \in \mathcal{X}} f^R_{\mathbf{y}^*,\mu}(\mathbf{x}) \quad \text{and} \quad \max_{\mathbf{y} \in \mathcal{Y}} f^R_{\mathbf{x}^*,\mu}(\mathbf{y}). \tag{5.37}$$

By (5.36) and the convexity-concavity of c, the two objective functions in (5.37) are DC functions with respect to $\mathbf{x}$ and $\mathbf{y}$, and then they are B-differentiable on $\mathcal{X}$ and $\mathcal{Y}$, respectively. For the sake of completeness, we recall the definition of d-stationary point in DC programming. We call $\mathbf{x}^* \in \mathcal{X}$ a d-stationary point [17, Definition 6.1.1] of the minimization problem in (5.37), if

$$(f^R_{\mathbf{y}^*,\mu})'(\mathbf{x}^*; \mathbf{x} - \mathbf{x}^*) \geq 0, \quad \forall\, \mathbf{x} \in \mathcal{X}, \tag{5.38}$$

which is a necessary optimality condition to the minimization program in (5.37).

Define

$$\varpi_1(t) = t, \quad \varpi_2(t) = 0, \quad \varpi(t) = \max\{\varpi_1(t), \varpi_2(t)\}$$
$$\text{and} \quad \mathcal{D}(t) = \{d \in \{1,2\} : \varpi(t) = \varpi_d(t)\}.$$

It is clear that

$$\varpi_1'(t) = 1,\ \varpi_2'(t) = 0 \text{ and } \partial\varpi(t) = \begin{cases} 1 & \text{if } t > 0 \\ [0,1] & \text{if } t = 0 \\ 0 & \text{if } t < 0. \end{cases}$$

(5.38) is equivalent to that, for any $q_l^* \in \mathcal{D}(g_l(\mathbf{x}_i^*) - \mu)$, it holds

$$\frac{\lambda_1}{\mu} \sum_{i\in[n]} \sum_{l\in\mathcal{N}_i} \varpi_{q_l^*}'(t)_{t=g_l(\mathbf{x}_i^*)-\mu} g_l'(\mathbf{x}_i^*)\mathbf{e}_i \in \partial_{\mathbf{x}} c(\mathbf{x}^*, \mathbf{y}^*) + \frac{\lambda_1}{\mu}\partial\left(\sum_{i\in[n]} \sum_{l\in\mathcal{N}_i} \varpi(g_l(\mathbf{x}_i^*))\right) + N_{\mathcal{X}}(\mathbf{x}^*), \tag{5.39}$$

in which by [14, Proposition 2.3.10],

$$\partial\left(\sum_{i\in[n]} \sum_{l\in\mathcal{N}_i} \varpi(g_l(\mathbf{x}_i^*))\right) = \sum_{i\in[n]} \sum_{l\in\mathcal{N}_i} \partial\varpi(t)_{t=g_l(\mathbf{x}_i^*)} g_l'(\mathbf{x}_i^*)\mathbf{e}_i.$$

Similarly, we call $\mathbf{y}^* \in \mathcal{Y}$ a d-stationary point of the maximization problem in (5.37), if

$$(f^R_{\mathbf{x}^*,\mu})'(\mathbf{y}^*; \mathbf{y} - \mathbf{y}^*) \le 0, \quad \forall \mathbf{y} \in \mathcal{Y},$$

which is equivalent to that for any $p_k^* \in \mathcal{D}(h_k(\mathbf{y}_j^*) - \mu)$, it holds

$$\begin{aligned} &\frac{\lambda_2}{\mu} \sum_{j\in[m]} \sum_{k\in\mathcal{M}_j} \varpi_{p_k^*}'(t)_{t=h_k(\mathbf{y}_j^*)-\mu} h_k'(\mathbf{y}_j^*)\mathbf{e}_j \\ \in &- \partial_{\mathbf{y}} c(\mathbf{x}^*, \mathbf{y}^*) + \frac{\lambda_2}{\mu}\partial\left(\sum_{j\in[m]} \sum_{k\in\mathcal{M}_j} \varpi(h_k(\mathbf{y}_j^*))\right) + N_{\mathcal{Y}}(\mathbf{y}^*). \end{aligned} \tag{5.40}$$

Based on the above analysis, we introduce the following definitions to min-max problem (5.9).

Definition 5.1. *For $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$,*

- *if (5.39) and (5.40) hold for all $q_l^* \in \mathcal{D}(g_l(\mathbf{x}_i^*) - \mu)$ and $p_k^* \in \mathcal{D}(h_k(\mathbf{y}_j^*) - \mu)$ with $i \in [n]$, $l \in \mathcal{N}_i$ and $j \in [m]$, $k \in \mathcal{M}_j$, we call $(\mathbf{x}^*, \mathbf{y}^*)$ a **d-stationary point** of min-max problem (5.9);*
- *if there exist a couple of sequences $q_l^* \in \mathcal{D}(g_l(\mathbf{x}_i^*) - \mu)$ and $p_k^* \in \mathcal{D}(h_k(\mathbf{y}_j^*) - \mu)$ for $i \in [n]$, $l \in \mathcal{N}_i$ and $j \in [m]$, $k \in \mathcal{M}_j$, such that (5.39) and (5.40) hold, we call $(\mathbf{x}^*, \mathbf{y}^*)$ a **weak d-stationary point** of min-max problem (5.9).*

On one hand, if $(\mathbf{x}^*, \mathbf{y}^*)$ is a local saddle point of problem (5.9), then it is a (weak) d-stationary point of (5.9). On the other hand, if $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ is a weak d-stationary point of (5.9), then it satisfies (5.28) and (5.29).

Proposition 5.3. *Let density function ρ be defined as in Example 3.1 with $p = 1$ and $0 < \mu < \bar{\mu}_1$ with $\bar{\mu}_1$ defined in (5.30). If $(\mathbf{x}^*, \mathbf{y}^*)$ is a weak d-stationary point of (5.9), then the following statements hold.*

(i) *$g_l(\mathbf{x}_i^*) \notin (0, \mu)$, $\forall l \in \mathcal{N}_i$, $i \in [n]$ and $h_k(\mathbf{y}_j^*) \notin (0, \mu)$, $\forall k \in \mathcal{M}_j$, $j \in [m]$;*

(ii) *if $g_{\tilde{l}}(\mathbf{x}_{\tilde{i}}^*) = \mu$ for some $\tilde{i} \in [n]$ and $\tilde{l} \in \mathcal{N}_{\tilde{i}}$, then the $q_{\tilde{l}}^* \in \mathcal{D}(g_{\tilde{l}}(\mathbf{x}_{\tilde{i}}^*) - \mu)$ satisfying (5.39) is unique and $q_{\tilde{l}}^* = 1$;*

(iii) *if $h_{\tilde{k}}(\mathbf{y}_{\tilde{j}}^*) = \mu$ for some $\tilde{j} \in [m]$ and $\tilde{k} \in \mathcal{M}_{\tilde{j}}$, then the $p_{\tilde{k}}^* \in \mathcal{D}(h_{\tilde{k}}(\mathbf{y}_{\tilde{j}}^*) - \mu)$ satisfying (5.40) is unique and $p_{\tilde{k}}^* = 1$.*

Proof. From Proposition 5.1, (i) holds naturally. Next, we argue (ii) by contradiction and (iii) can be proved similarly. For item (ii), suppose there exist $\tilde{i} \in [n]$ and $\tilde{l} \in \mathcal{N}_{\tilde{i}}$ such that $g_{\tilde{l}}(\mathbf{x}_{\tilde{i}}^*) = \mu$ and (5.39) holds with $q_{\tilde{l}}^* = 2$. For any $l \in \mathcal{N}_{\tilde{i}}$ and $l \neq \tilde{l}$, by Assumption 5.1 and $\mu < \bar{\mu}_1 \leq \tau$, we have $\mathbf{x}_{\tilde{i}}^* \in \mathrm{int}(\tilde{\mathcal{X}}_{\tilde{i}})$ and $g_l(\mathbf{x}_{\tilde{i}}^*) \notin [0, \tau]$, which implies $q_l^* \in \mathcal{D}(g_l(\mathbf{x}_{\tilde{i}}^*) - \mu)$ is unique and $\varpi'_{q_l^*}(t)_{t=g_l(\mathbf{x}_{\tilde{i}}^*)-\mu} = \varpi'(t)_{t=g_l(\mathbf{x}_{\tilde{i}}^*)}$. Then, (5.39) gives

$$0 \in [\partial_{\mathbf{x}} c(\mathbf{x}^*, \mathbf{y}^*)]_{\tilde{i}} + \frac{\lambda_1}{\mu} g'_{\tilde{l}}(\mathbf{x}_{\tilde{i}}^*) + [N_{\mathcal{X}}(\mathbf{x}^*)]_{\tilde{i}}. \tag{5.41}$$

Using Assumption 5.1, (5.7) and (5.17), we confirm that $\frac{\lambda_1 \sigma}{\mu} \leq L_{c,1}$, which contradicts to the supposition on the value of μ and gives the result in (ii). □

For $t \in \mathbb{R}$, denote $\mathcal{A}_i^0(t) = \{l \in \mathcal{N}_i : g_l(t) = 0\}$ and $\mathcal{B}_j^0(t) = \{k \in \mathcal{M}_j : h_k(t) = 0\}$. For given $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ and $\mu \in \mathbb{R}_{++}$, consider the following two functions

$$W_{\mathbf{x}^*, \mathbf{y}^*, \mu}(\mathbf{x}) = c(\mathbf{x}, \mathbf{y}^*) + \frac{\lambda_1}{\mu} \sum\nolimits_{i=1}^{n} \sum\nolimits_{l \in \mathcal{A}_i^0(\mathbf{x}_i^*)} g_l(\mathbf{x}_i)_+, \tag{5.42}$$

$$V_{\mathbf{x}^*, \mathbf{y}^*, \mu}(\mathbf{y}) = -c(\mathbf{x}^*, \mathbf{y}) + \frac{\lambda_2}{\mu} \sum\nolimits_{j=1}^{m} \sum\nolimits_{k \in \mathcal{B}_j^0(\mathbf{y}_j^*)} h_k(\mathbf{y}_j)_+. \tag{5.43}$$

If g_l is convex for any $l \in [\hat{n}]$, by the convexity of $c(\cdot, \mathbf{y}^*)$, function $W_{\mathbf{x}^*, \mathbf{y}^*, \mu}$ is convex on $\mathcal{X}$, which gives that $\mathbf{x}^*$ is a minimizer of $W_{\mathbf{x}^*, \mathbf{y}^*, \mu}$ on $\mathcal{X}$ if and only if

$$\begin{aligned} 0 \in & \partial W_{\mathbf{x}^*, \mathbf{y}^*, \mu}(\mathbf{x}^*) + N_{\mathcal{X}}(\mathbf{x}^*) \\ = & \partial_{\mathbf{x}} c(\mathbf{x}^*, \mathbf{y}^*) + \frac{\lambda_1}{\mu} \sum\nolimits_{i=1}^{n} \left(\sum\nolimits_{l \in \mathcal{A}_i^0(\mathbf{x}_i^*)} [0, 1] g'_l(\mathbf{x}_i^*) \right) \mathbf{e}_i + N_{\mathcal{X}}(\mathbf{x}^*). \end{aligned} \tag{5.44}$$

Similarly, if h_k is convex for any $k \in [\hat{m}]$, $V_{\mathbf{x}^*, \mathbf{y}^*, \mu}$ is convex on $\mathcal{Y}$ and $\mathbf{y}^* \in \mathcal{Y}$ is a minimizer of $V_{\mathbf{x}^*, \mathbf{y}^*, \mu}$ on $\mathcal{Y}$ if and only if

$$0 \in \partial V_{\mathbf{x}^*, \mathbf{y}^*, \mu}(\mathbf{y}^*) + N_{\mathcal{Y}}(\mathbf{y}^*).$$

In what follows, we will verify that all weak d-stationary points of (5.9) are μ-strong local saddle points of (5.1).

Theorem 5.2. *Under conditions of Proposition 5.3, if $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ is a weak d-stationary point of (5.9), then it is a μ-strong local saddle point of problem (5.1).*

Proof. Since $(\mathbf{x}^*, \mathbf{y}^*)$ is a weak d-stationary point of (5.9), putting forward the results in Proposition 5.3 to (5.39), we have (5.44), which means that $\mathbf{x}^*$ is a global minimizer of $W_{\mathbf{x}^*,\mathbf{y}^*,\mu}$ on $\mathcal{X}$, i.e.

$$c(\mathbf{x}^*, \mathbf{y}^*) \leq c(\mathbf{x}, \mathbf{y}^*) + \frac{\lambda_1}{\mu} \sum\nolimits_{i=1}^{n} \sum\nolimits_{l \in \mathcal{A}_i^0(\mathbf{x}_i^*)} g_l(\mathbf{x}_i)_+, \quad \forall\, \mathbf{x} \in \mathcal{X}. \tag{5.45}$$

Then, (5.45) means that for any $\mathbf{x} \in \{\mathbf{x} \in \mathcal{X} : g_l(\mathbf{x}_i) \leq 0 \text{ if } g_l(\mathbf{x}_i^*) \leq 0 \text{ for } l \in \mathcal{N}_i \text{ and } i \in [n]\}$, it holds $c(\mathbf{x}^*, \mathbf{y}^*) \leq c(\mathbf{x}, \mathbf{y}^*)$. Similarly, Proposition 5.3 together with (5.43) implies that $\mathbf{y}^*$ is a global minimizer of $V_{\mathbf{x}^*,\mathbf{y}^*,\mu}$ on $\mathcal{Y}$ and we further have that $\mathbf{y}^*$ is a maximizer of $c(\mathbf{x}^*, \cdot)$ on $\{\mathbf{y} \in \mathcal{Y} : h_k(\mathbf{y}) \leq 0 \text{ if } h_k(\mathbf{y}_j^*) \leq 0 \text{ for } k \in \mathcal{M}_j \text{ and } j \in [m]\}$. Thus, from Theorem 2.1 and recalling Proposition 5.3-(i), $(\mathbf{x}^*, \mathbf{y}^*)$ is a μ-strong local saddle point of (5.1). □

Remark 5.2. *Following the proof of Proposition 5.3, when* $0 < \mu < \bar{\mu}_1$*, if* $\mathbf{x}^*$ *and* $\mathbf{y}^*$ *satisfy*

$$\mathbf{x}^* \in \arg\min_{\mathbf{x} \in \mathcal{X}} W_{\mathbf{x}^*,\mathbf{y}^*,\mu}(\mathbf{x}) \quad \textit{and} \quad \mathbf{y}^* \in \arg\min_{\mathbf{y} \in \mathcal{Y}} V_{\mathbf{x}^*,\mathbf{y}^*,\mu}(\mathbf{y}),$$

then $(\mathbf{x}^*, \mathbf{y}^*)$ *is a* μ*-strong local saddle point of* (5.1).

By [33, Proposition 17], any local saddle point is a local minimax point. Then, by Theorem 5.2, any weak d-stationary point of problem (5.9) is also a local minimax point of problem (5.1).

Since the continuous relaxation functions to the cardinality functions in (5.9) are DC functions and variable separated, the proximal operator of its subtracted convex function can be calculated directly in most cases. Moreover, to solve problem (5.9) with a nonsmooth function c efficiently, we can use a smoothing approximation of (5.9) as follows

$$\min_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} f^R(\mathbf{x}, \mathbf{y}, \mu, \varepsilon) := \tilde{c}(\mathbf{x}, \mathbf{y}, \varepsilon) + \lambda_1 \sum_{i=1}^{n} \phi_i^R(\mathbf{x}_i, \mu) - \lambda_2 \sum_{j=1}^{m} \psi_j^R(\mathbf{y}_j, \mu), \tag{5.46}$$

where $\tilde{c}$ is a smoothing function of c defined by (4.26). Similar to the expression in (5.36) and by Proposition 4.2-(ii), for fixed $\mu > 0$ and $\varepsilon > 0$, $\tilde{f}^R(\mathbf{x}, \mathbf{y}, \mu, \varepsilon)$ in (5.46) is a DC function with respect to $\mathbf{x}$ and $\mathbf{y}$, respectively. Thus, the d-stationary point and weak d-stationary point to (5.46) can be defined according to Definition 5.1. By using the gradient consistency (4.23)-(4.24), we have the following result.

Proposition 5.4. *Let* $\tilde{c} : \mathbb{R}^n \times \mathbb{R}^m \times (0, 1] \to \mathbb{R}$ *be defined as in* (4.26). *If* $\{(\mathbf{x}^k, \mathbf{y}^k)\}$ *is a sequence of weak d-stationary points of* (5.46) *with* $\varepsilon := \varepsilon_k \downarrow 0$*, then any accumulation point of* $\{(\mathbf{x}^k, \mathbf{y}^k)\}$ *is a weak d-stationary point of* (5.9).

5.2 Density function ρ under Assumption 3.2

Section 5.1 focuses on the study of (5.1) with a density function ρ satisfying Assumption 3.1. From Table 1, we find that the other three density functions satisfy Assumption 3.2 and the corresponding continuous relaxation function $r(\cdot, \mu)$ owns the continuous differentiability on $\mathbb{R}_{++}$, which may bring some convenience to its

algorithm research when c is smooth. Thus, in this subsection, we pay attention to the results of the continuous relaxation to (5.1) with density function ρ satisfying Assumption 3.2 and consider (5.1) under the following conditions:

(i) functions c, g_l, $l \in [\hat{n}]$ and h_k, $k \in [\hat{m}]$ are Lipschitz continuously differentiable;

(ii) the feasible regions are defined by the box constraints, i.e.

$$\mathcal{X} := \tilde{\mathcal{X}} = \{\mathbf{x} \in \mathbb{R}^n : \underline{\mathbf{u}} \leq \mathbf{x} \leq \overline{\mathbf{u}}\}, \quad \mathcal{Y} := \tilde{\mathcal{Y}} = \{\mathbf{y} \in \mathbb{R}^m : \underline{\mathbf{v}} \leq \mathbf{y} \leq \overline{\mathbf{v}}\},$$

where $\underline{\mathbf{u}}$, $\overline{\mathbf{u}}$, $\underline{\mathbf{v}}$ and $\overline{\mathbf{v}}$ are defined as in (5.4);

(iii) Assumption 5.1 holds, in which the conditions of (5.12) and (5.15) can be ignored;

(iv) under Assumption 3.2, there exist $\bar{\rho}_2 > 0$ and $\rho_0 > 0$ such that

$$\rho(s) \leq \bar{\rho}_2, \quad \forall s \in (0, \alpha) \quad \text{and} \quad \lim_{t \downarrow 0} \rho(t) = \rho_0. \tag{5.47}$$

In this case, (5.2) holds naturally and we will consider the second order necessary optimality condition of (5.1).

To proceed, we first introduce some notations on the existing parameters.

- By virtue of the Lipschitz continuous differentiability of c on $\mathcal{X} \times \mathcal{Y}$, there exists $L_{c,2}$ such that for any $\mathbf{x} \in \mathcal{X}$, $\mathbf{y} \in \mathcal{Y}$, it holds

$$\sup\{|H_{ii}|, |M_{jj}| : H \in \partial^2_{\mathbf{xx}} c(\mathbf{x}, \mathbf{y}), M \in \partial^2_{\mathbf{yy}} c(\mathbf{x}, \mathbf{y}), i \in [n], j \in [m]\} \leq L_{c,2}.$$

- Since $g_l : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous differentiable on $\mathcal{X}_i$ for $l \in \mathcal{N}_i$, and $\mathcal{X}_i$ is compact, there exists $L_{g,2}$ such that

$$\sup\{|\xi| : \xi \in \partial^2 g_l(\mathbf{x}_i), \mathbf{x}_i \in \mathcal{X}_i, i \in [n], l \in \mathcal{N}_i\} \leq L_{g,2}.$$

  Similarly, there exists $L_{h,2}$ such that

$$\sup\{|\eta| : \eta \in \partial^2 h_k(\mathbf{y}_j), \mathbf{y}_j \in \mathcal{Y}_j, j \in [m], k \in \mathcal{M}_j\} \leq L_{h,2}.$$

- For $t \in \mathbb{R}$, $i \in [n]$, $j \in [m]$ and $\delta > 0$, denote

$$\mathcal{A}^+_{\delta,i}(t) = \{l \in \mathcal{N}_i : 0 < g_l(t) < \delta\}, \quad \mathcal{B}^+_{\delta,j}(t) = \{k \in \mathcal{M}_j : 0 < h_k(t) < \delta\}.$$

Proceed to the next step, and let

$$\bar{\mu}_2 = \min\left\{\frac{\tau}{\alpha}, \frac{\lambda\sigma\underline{\rho}_2}{L_{c,1}}, \frac{\lambda_1\check{\rho}_2\sigma^2}{\tau L_{c,2}/\alpha + \lambda_1\bar{\rho}_2 L_{g,2}}, \frac{\lambda_2\check{\rho}_2\sigma^2}{\tau L_{c,2}/\alpha + \lambda_2\bar{\rho}_2 L_{h,2}}\right\} \tag{5.48}$$

with $\lambda = \min\{\lambda_1, \lambda_2\}$. In particular, when g_l, h_k are linear functions and $c(\cdot,\cdot)$ is also linear with respect to $\mathbf{x}$ and $\mathbf{y}$, respectively, then $\bar{\mu}_2 = \left\{\frac{\tau}{\alpha}, \frac{\lambda\sigma\underline{\rho}_2}{L_{c,1}}\right\}$. If we further choose ρ as in Example 3.3 or Example 3.4, then, $\bar{\mu}_2 = \frac{\tau}{\alpha}$.

In what follows, suppose that ρ satisfies Assumption 3.2 and $0 < \mu < \bar{\mu}_2$ with $\bar{\mu}_2$ defined in (5.48). By the Lipschitz continuity of ρ on $\mathbb{R}_{++}$, $r(g_l(t), \mu)$ is Lipschitz

continuous differentiable on $\{t : g_l(t) > 0\}$, $\forall l \in [\hat{n}]$. When $g_l(t) > 0$, by (3.9), the second order generalized derivative of $r(g_l(t), \mu)$ with respect to t satisfies

$$\partial^2_{tt} r(g_l(t), \mu) \subseteq \tilde{\partial}^2_{tt} r(g_l(t), \mu) := \frac{\partial \rho(s)_{s=g_l(t)/\mu}}{\mu^2} (g_l'(t))^2 + \frac{\rho(g_l(t)/\mu)}{\mu} \partial^2 g_l(t), \quad (5.49)$$

where $\rho(g_l(t)/\mu) = 0$ if $g_l(t) \notin [0, \tau)$. Then, $\tilde{\partial}^2_{tt} r(g_l(t), \mu) = \{0\}$ when $g_l(t) \notin [0, \tau)$. Thus, if $\mathbf{x}^*$ is a local solution of $\min_{\mathbf{x}\in\mathcal{X}} f^R(\mathbf{x}, \mathbf{y}^*, \mu)$ and $\mathcal{A}^+_{\tau,i}(\mathbf{x}^*_i) \neq \emptyset$ for $i \in [n]$, by Assumption 5.1 and the second order necessary optimality condition [30], then

$$\begin{cases} \text{there exists a unique } \hat{l} \in \mathcal{N}_i \text{ such that } 0 < g_{\hat{l}}(\mathbf{x}^*_i) < \tau \text{ and} \\ \text{there exists a } \omega_i \in [\partial^2_{\mathbf{xx}} c(\mathbf{x}^*, \mathbf{y}^*)]_{ii} + \lambda_1 \tilde{\partial}^2_{tt} r(g_{\hat{l}}(t), \mu)_{t=\mathbf{x}^*_i} \text{ such that } \omega_i \geq 0, \end{cases} \quad (5.50)$$

which implies $\mathbf{x}^*_i \in \text{int}(\mathcal{X}_i)$. Similarly, if $\mathbf{y}^*$ is a local solution of $\max_{\mathbf{y}\in\mathcal{Y}} f^R(\mathbf{x}^*, \mathbf{y}, \mu)$ and $\mathcal{B}^+_{\tau,j}(\mathbf{y}^*_j) \neq \emptyset$ for $j \in [m]$, then

$$\begin{cases} \text{there exists a unique } \hat{k} \in \mathcal{M}_j \text{ such that } 0 < h_{\hat{k}}(\mathbf{y}^*_j) < \tau \text{ and} \\ \text{there exists a } \varpi_j \in -[\partial^2_{\mathbf{yy}} c(\mathbf{x}^*, \mathbf{y}^*)]_{jj} + \lambda_2 \tilde{\partial}^2_{tt} r(h_{\hat{k}}(t), \mu)_{t=\mathbf{y}^*_j} \text{ such that } \varpi_j \geq 0. \end{cases} \quad (5.51)$$

Thus, inspired by the first and second order necessary optimality conditions to

$$\mathbf{x}^* \in \arg\min_{\mathbf{x}\in\mathcal{X}} f^R(\mathbf{x}, \mathbf{y}^*, \mu) \quad \text{and} \quad \mathbf{y}^* \in \arg\max_{\mathbf{y}\in\mathcal{Y}} f^R(\mathbf{x}^*, \mathbf{y}, \mu), \quad (5.52)$$

we introduce the following definition.

Definition 5.2. *We call* $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ *a* **weak second order stationary point** *of problem* (5.9)*, if*

$$\mathbf{0} \in \tilde{\partial}_{\mathbf{x}} f^R(\mathbf{x}^*, \mathbf{y}^*, \mu) + N_{\mathcal{X}}(\mathbf{x}^*) \quad \textit{and} \quad \mathbf{0} \in -\tilde{\partial}_{\mathbf{y}} f^R(\mathbf{x}^*, \mathbf{y}^*, \mu) + N_{\mathcal{Y}}(\mathbf{y}^*), \quad (5.53)$$

where $\tilde{\partial}_{\mathbf{x}} f^R(\mathbf{x}^*, \mathbf{y}^*, \mu)$ *and* $\tilde{\partial}_{\mathbf{y}} f^R(\mathbf{x}^*, \mathbf{y}^*, \mu)$ *are defined in* (5.26) *and* (5.27)*, and for any* $i \in [n]$ *with* $\mathcal{A}^+_{\tau,i}(\mathbf{x}^*_i) \neq \emptyset$ *and* $j \in [m]$ *with* $\mathcal{B}^+_{\tau,j}(\mathbf{y}^*_j) \neq \emptyset$*,* (5.50) *and* (5.51) *hold, respectively.*

It is clear that (5.53) and (5.50)-(5.51) are weaker than the general first and second order necessary optimality conditions to (5.9), respectively, so we call it "weak" stationary point.

Theorem 5.3. *Suppose problem* (5.1) *satisfies Assumption 5.1, density function* ρ *satisfies Assumption 3.2 and* $0 < \mu < \bar{\mu}_2$ *with* $\bar{\mu}_2$ *defined in* (5.48)*. Then, the following statements hold.*

(i) *If* $(\mathbf{x}^*, \mathbf{y}^*)$ *is a weak second order stationary point of* (5.9)*, then*

$$g_l(\mathbf{x}^*_i) \notin (0, \alpha\mu), \forall l \in \mathcal{N}_i,\ i \in [n]; \quad h_k(\mathbf{y}^*_j) \notin (0, \alpha\mu), \forall k \in \mathcal{M}_j,\ j \in [m]. \quad (5.54)$$

(ii) $(\mathbf{x}^*, \mathbf{y}^*)$ *is a saddle point of problem* (5.1) *if and only if it is a saddle point of* (5.9)*.*

(iii) $(\mathbf{x}^*, \mathbf{y}^*)$ *is an* $\alpha\mu$*-strong local saddle point of* (5.1) *if it is a local saddle point of* (5.9).

(iv) *When functions* g_l, h_k *are convex for all* $l \in [\hat{n}]$ *and* $k \in [\hat{m}]$, $(\mathbf{x}^*, \mathbf{y}^*)$ *is an* $\alpha\mu$*-strong local saddle point of* (5.1) *if it is a weak second order stationary point of* (5.9).

Proof. To prove (i), we argue the results in (5.54) by contradiction. Suppose there exist $\tilde{i} \in [n]$ and $\tilde{l} \in \mathcal{N}_{\tilde{i}}$ such that $0 < g_{\tilde{l}}(\mathbf{x}^*_{\tilde{i}}) < \alpha\mu$.

By Assumption 5.1, since $\alpha\mu < \tau$, then $\mathbf{x}^*_{\tilde{i}} \in \mathrm{int}(\mathcal{X}_{\tilde{i}})$, and for any $l \in \mathcal{N}_{\tilde{i}}$, $l \neq \tilde{l}$, $g_l(\mathbf{x}^*_{\tilde{i}}) \notin [0, \tau]$, which together with (3.4) implies that

$$\nabla_t r(t,\mu)_{t=g_l(\mathbf{x}^*_{\tilde{i}})} = 0 \quad \text{and} \quad \nabla^2_t r(t,\mu)_{t=g_l(\mathbf{x}^*_{\tilde{i}})} = 0.$$

Next, we obtain the contradiction to $0 < g_{\tilde{l}}(\mathbf{x}^*_{\tilde{i}}) < \alpha\mu$ from two cases.

Case 1: $\sup\{a : a \in \partial\rho(g_{\tilde{l}}(\mathbf{x}^*_{\tilde{i}})/\mu)\} > -\check{\rho}_2$. By Assumption 3.2, it means that $\rho(g_{\tilde{l}}(\mathbf{x}^*_{\tilde{i}})/\mu) \geq \underline{\rho}_2$. Similar to the discussion in Proposition 5.1, by $\mu < \bar{\mu}_2 \leq \lambda_1\sigma\underline{\rho}_2/L_{c,1}$, it brings a contradiction.

Case 2: $\sup\{a : a \in \partial\rho(g_{\tilde{l}}(\mathbf{x}^*_{\tilde{i}})/\mu)\} \leq -\check{\rho}_2$. By (5.49) and (5.50), there exist $\xi_{\tilde{i}} \in [\partial^2_{\mathbf{xx}} c(\mathbf{x}^*, \mathbf{y}^*)]_{\tilde{i}\tilde{i}}$, $\eta_{\tilde{i}} \in \partial\rho(t)_{t=g_{\tilde{l}}(\mathbf{x}^*_{\tilde{i}})/\mu}$ and $\zeta_{\tilde{i}} \in \partial^2 g_{\tilde{l}}(t)_{t=\mathbf{x}_{\tilde{i}^*}}$ such that

$$\xi_{\tilde{i}} + \lambda_1 \frac{\eta_{\tilde{i}}}{\mu^2}(g'_{\tilde{l}}(\mathbf{x}^*_{\tilde{i}}))^2 + \lambda_1 \frac{\rho(g_{\tilde{l}}(\mathbf{x}^*_{\tilde{i}})/\mu)}{\mu}\zeta_{\tilde{i}} \geq 0. \tag{5.55}$$

Recalling Assumption 3.2 and by $\mu < \frac{\tau}{\alpha}$, an estimation on the left side of (5.55) gives

$$0 \leq \frac{L_{c,2}\tau}{\alpha} - \lambda_1\check{\rho}_2\sigma^2/\mu + \lambda_1\bar{\rho}_2 L_{g,2}, \tag{5.56}$$

which contradicts to $\mu < \bar{\mu}_2 \leq \frac{\lambda_1\check{\rho}_2\sigma^2}{\tau L_{c,2}/\alpha + \lambda_1\bar{\rho}_2 L_{g,2}}$ given in (5.48). Therefore, $g_l(\mathbf{x}^*_i) \notin (0, \alpha\mu)$, $\forall l \in \mathcal{N}_i$, $i \in [n]$. Similarly, $h_k(\mathbf{y}^*_j) \notin (0, \alpha\mu)$, $\forall k \in \mathcal{M}_j$, $j \in [m]$. Thus, (i) holds. Moreover, (i) implies Assumption 4.1. By Theorem 4.1, we can obtain (ii) and (iii).

(iv) Suppose $(\mathbf{x}^*, \mathbf{y}^*)$ is a weak second order stationary point of (5.9). To proceed the proof, we use a slight modification of functions in (5.42) and (5.43) as follows

$$W_{\mathbf{x}^*,\mathbf{y}^*,\mu}(\mathbf{x}) = c(\mathbf{x}, \mathbf{y}^*) + \frac{\lambda_1\rho_0}{\mu}\sum\nolimits_{i=1}^{n}\sum\nolimits_{l\in\mathcal{A}^0_i(\mathbf{x}^*_i)} g_l(\mathbf{x}_i)_+, \tag{5.57}$$

$$V_{\mathbf{x}^*,\mathbf{y}^*,\mu}(\mathbf{y}) = -c(\mathbf{x}^*, \mathbf{y}) + \frac{\lambda_2\rho_0}{\mu}\sum\nolimits_{j=1}^{m}\sum\nolimits_{k\in\mathcal{B}^0_j(\mathbf{y}^*_j)} h_k(\mathbf{y}_j)_+, \tag{5.58}$$

which are convex on $\mathcal{X}$ and $\mathcal{Y}$, respectively. By Assumption 3.2 and (3.3), if $g_l(\mathbf{x}^*_i) \geq \alpha\mu$ or $g_l(\mathbf{x}^*_i) < 0$ for some $i \in [n]$ and $l \in \mathcal{N}_i$, then $\partial_t r(t,\mu)_{t=g_l(\mathbf{x}^*_i)} g'_l(\mathbf{x}^*_i) = 0$. From (i), $(\mathbf{x}^*, \mathbf{y}^*)$ satisfies (5.54). Thus, using (3.3) again and by (5.53), we have

$$\mathbf{0} \in \nabla_{\mathbf{x}} c(\mathbf{x}^*, \mathbf{y}^*) + \frac{\lambda_1\rho_0}{\mu}\sum\nolimits_{i=1}^{n}\left(\sum\nolimits_{l\in\mathcal{A}^0_i(\mathbf{x}^*_i)}[0,1]g'_l(\mathbf{x}^*_i)\right)\mathbf{e}_i + N_{\mathcal{X}}(\mathbf{x}^*), \tag{5.59}$$

which implies $0 \in \partial_{\mathbf{x}} W_{\mathbf{x}^*,\mathbf{y}^*,\mu}(\mathbf{x}^*) + N_{\mathcal{X}}(\mathbf{x}^*)$. Thus, $\mathbf{x}^*$ is a global minimizer of $W_{\mathbf{x}^*,\mathbf{y}^*,\mu}(\mathbf{x})$ on $\mathcal{X}$. In what follows, similar to the analysis in Theorem 5.2, we get that $(\mathbf{x}^*, \mathbf{y}^*)$ is an $\alpha\mu$-strong local saddle point of (5.1). □

5.3 Continuous relaxations defined by different density functions

In this subsection, we use three examples to explain the different properties of the continuous relaxation problems constructed by the density functions that satisfy Assumption 3.1 or 3.2. In particular, we use the density functions in Examples 3.1 and 3.3 to construct two different continuous relaxation problems, which have different relations with min-max problem (1.1) regarding local saddle points and strong local saddle points.

- In Example 5.1, we show that we can provide a possible larger lower bound to the saddle points of (5.1) by the analysis on the continuous relaxation models with different density functions.
- It is interesting to see in Example 5.2 that the bounds in (4.1) and (4.2) with $0 < \mu < \bar{\mu}_1$, $\bar{\mu}_1$ in (5.30) and $\alpha = 1$ (given in subsection 5.1 by the continuous relation model with a density function in Example 3.1 and $p = 1$) is satisfied by the global minimax points of this example, but these bounds with $0 < \mu < \bar{\mu}_2$, $\bar{\mu}_2$ in (5.48), and α in Assumption 3.2 (given in subsection 5.2 by the continuous relation model with a density function satisfying Assumption 3.2) may not hold to the global minimax points.
- Note that all the functions $r(\cdot, \mu)$ in Examples 3.1-3.4 can be expressed by DC functions and continuously differentiable on $(0, \alpha\mu)$, where $p = 1$ in Example 3.1. Then, when c is continuously differentiable, both the weak d-stationary point and weak second order stationary point to these continuous relaxation models are well-defined. In Example 5.3, we will show that a weak second order stationary point is not necessary to be a weak d-stationary point of the continuous relaxation problem with a density function in Example 3.1 and $p = 1$. Moreover, a weak d-stationary points is also not necessary to be a weak second order stationary point of the continuous relaxation problem with a density function in Example 3.3.

Example 5.1. *Consider*

$$\min_{\mathbf{x}\in[-2,2]} \max_{\mathbf{y}\in[-2,2]} f(\mathbf{x}, \mathbf{y}) := (\mathbf{x} - 1)(\mathbf{y} - 1) + 3\|\mathbf{x}\|_0 - 3\|\mathbf{y}\|_0. \tag{5.60}$$

In Example 2.1, we have verified that $(0, 0)$ *is the unique saddle point of* (5.60). *Assumption 5.1 holds with* $\sigma = 1$, $\tau = 2$, $L_{c,1} = 3$ *and* $L_{c,2} = L_{g,2} = L_{h,2} = 0$.

Case 1: Choose the density function ρ *in Example 3.1 with* $p = 1$ *to build up its continuous relaxation. Then,* $\alpha = 1$, $\underline{\rho} = 1$ *and then* $\bar{\mu}_1 = 1$ *in* (5.30). *Since we can choose any* μ *in* $(0, \bar{\mu}_1)$, *by Theorem 5.1, it gives that the saddle points and global minimax points of* (5.60) *satisfy the lower bounds that*

$$\textit{either } \mathbf{x} = 0 \textit{ or } |\mathbf{x}| \geq \nu \textit{ and either } \mathbf{y} = 0 \textit{ or } |\mathbf{y}| \geq \nu \tag{5.61}$$

with $\nu = 1$.

Case 2: Choose the density function ρ in Example 3.3 with $\alpha = 2$ to build up its continuous relaxation. Then, the analysis in subsection 5.2 gives that $\bar{\mu}_2 = 1$. By Theorem 5.3, we have that any saddle point of (5.60) satisfies the lower bound in (5.61) with $\nu = 2$.

Example 5.2. *Consider*

$$\min_{\mathbf{x}\in[-2,2]} \max_{\mathbf{y}\in[-2,2]} f(\mathbf{x},\mathbf{y}) := (\mathbf{x}-1)(\mathbf{y}-1) + \|\mathbf{x}\|_0 - \|\mathbf{y}\|_0.$$

Example 2.1 shows that $(3/2, 0)$ and $(3/2, 2)$ are global minimax points of this problem.

By basic calculation, $\bar{\mu}_1 = 1/3$ when we define ρ by Example 3.1 with $p = 1$. Then, by Theorem 5.1, any global minimax point of this example satisfies (5.61) with any $\nu = \mu < \bar{\mu}_1$.

However, when we define ρ by Example 3.3 with $\alpha = 2$, then $\bar{\mu}_2 = 1$. It is obvious that neither of the two global minimax points satisfies (5.61) with $\nu = \alpha\mu$ when $3/4 < \mu < \bar{\mu}_2$.

Example 5.3. *Consider*

$$\min_{\mathbf{x}\in[-2,2]} \max_{\mathbf{y}\in[-2,2]} f(\mathbf{x},\mathbf{y}) := (\mathbf{x}-1)(1-\mathbf{y}) + \|\mathbf{x}\|_0 - \|\mathbf{y}\|_0. \tag{5.62}$$

On one hand, choose the density function ρ in Example 3.1 with $p = 1$ and $\mu = 1/4$ to build up its continuous relaxation, where $0 < \mu < \bar{\mu}_1 = 1/3$. For this case, we can verify that $(-1/4, 1/4)$ is a weak second order stationary point of its continuous relaxation model, but it is not a weak d-stationary point of it and is also not a local saddle point of (5.62).

On the other hand, choose the density function ρ in Example 3.3 with $\alpha = 1$ and $\mu = 1$ to build up its continuous relaxation, where $0 < \mu < \bar{\mu}_2 = 2$. For this case, we can easily check that $(1, 1)$ is a weak d-stationary point but not a weak second order stationary point of this continuous relaxation model, and it is not a local saddle point of (5.62).

At the end of this subsection, we summarize the relations between min-max problem (5.1) and its continuous relaxation problem (5.9) in Fig. 1. From Fig. 1, we find that both weak d-stationary point and weak second order stationary point of (5.9) are necessary conditions to the saddle points of (5.1) and (5.9), but sufficient conditions to the ν-strong local saddle points of (5.1). Getting a bound on ν in (2.8) satisfied by all saddle points of (5.1) would allow us to discard a certain number of local saddle points which are not saddle points. When problem (5.1) satisfies Assumption 5.1, by Theorem 5.1 and Theorem 5.2, we can conclude that any saddle point of (5.1) satisfies the lower bounds in (2.8) with $\nu := \mu < \bar{\mu}_1$, which is obtained by the continuous relaxation model with ρ in Example 3.1 and $p = 1$. For a more special case, we also obtain from Theorem 5.3 that any saddle point of (5.1) satisfies the lower bounds in (2.8) with $\nu := \alpha\mu < \alpha\bar{\mu}_2$, which is obtained by the continuous relaxation model with ρ satisfying Assumption 3.2.

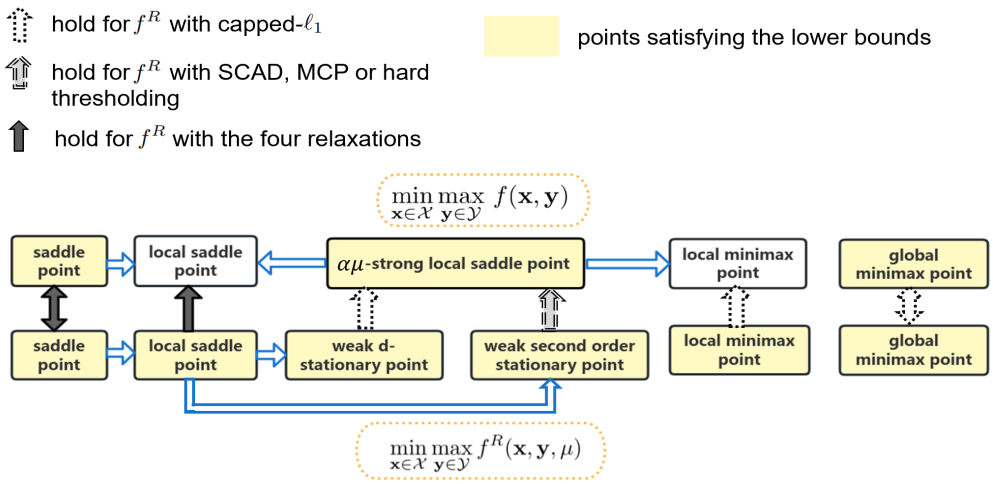


Fig. 1: Relations between problems (5.1) and (5.9) with different relaxations

6 Applications

In this section, we use three examples to explain the motivation and theoretical results of this paper. Moreover, we present numerical results for the third example.

6.1 Distributionally robust sparse convex regression

The sparse convex regression problem

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}[\varphi(\mathbf{x};\mathbf{c}_\xi,d_\xi)] + \lambda_1\|\mathbf{x}\|_0$$

has wide applications in data science, where $\mathcal{X}=\{\mathbf{x}\in\mathbb{R}^n : \underline{\mathbf{u}}\le\mathbf{x}\le\overline{\mathbf{u}}\}$ with $\underline{\mathbf{u}}<\overline{\mathbf{u}}$, $(\mathbf{c}_\xi,d_\xi)\in\mathbb{R}^n\times\mathbb{R}$ represents a random data set of interest, $\varphi(\cdot;\mathbf{c}_\xi,d_\xi):\mathbb{R}^n\to\mathbb{R}$ is a convex loss function and $\mathbb{E}$ is the expectation. Widely used convex loss functions include the censored function $(\max(\mathbf{c}_\xi^\top\mathbf{x},0)-d_\xi)^2$ and the ℓ_1 function $|\mathbf{c}_\xi^\top\mathbf{x}-d_\xi|$, which are nonsmooth functions. Then, the distributionally robust sparse convex regression problem can be expressed by

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\hat{\mathcal{Y}}} \sum_{i=1}^m \mathbf{y}_i\varphi_i(\mathbf{x}) + \lambda_1\|\mathbf{x}\|_0 \tag{6.1}$$

with $\varphi_i(\mathbf{x}) := \varphi(\mathbf{x};\mathbf{c}_i,d_i)$, a set of m samples $\{\mathbf{c}_i,d_i\}_{i=1}^m$ and the approximation of the ambiguity set $\hat{\mathcal{Y}}=\{\mathbf{y}\in\mathbb{R}^m : \mathbf{y}\ge 0,\ \mathbf{e}^\top\mathbf{y}=1,\ \|\mathbf{A}\mathbf{y}-\mathbf{b}\|\le\delta\}$. Here $(\mathbf{A},\mathbf{b},\delta)\in\mathbb{R}^{k\times m}\times\mathbb{R}^k\times\mathbb{R}_+$ describes the approximated ambiguity set in a general moment form.

Taking account of the constraint on $\mathbf{y}$, the following penalty form

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} \underbrace{\sum_{i=1}^{m}\mathbf{y}_i\varphi_i(\mathbf{x}) - \beta\max\{\|\mathbf{A}\mathbf{y}-\mathbf{b}\|^2-\delta^2,0\}}_{c(\mathbf{x},\mathbf{y})} + \lambda_1\|\mathbf{x}\|_0 \tag{6.2}$$

for (6.1) is promising, where $\beta > 0$ is a penalty parameter and $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} \geq 0,\ \mathbf{e}^\top\mathbf{y} = 1\}$. In (6.2), $c(\mathbf{x},\mathbf{y})$ is nonsmooth with respect to both $\mathbf{x}$ and $\mathbf{y}$. However, thanks to the method in subsection 4.3, a smoothing function to c can be easily constructed with the properties in (4.23) and (4.24). For example, if $\varphi_i(\mathbf{x}) = |\mathbf{c}_i^\top\mathbf{x} - d_i|$, then we can set

$$\tilde{c}(\mathbf{x},\mathbf{y},\varepsilon) = \sum_{i=1}^{m}\mathbf{y}_i\theta(\mathbf{c}_i^\top\mathbf{x}-d_i,\varepsilon) - \beta\phi(\|\mathbf{A}\mathbf{y}-\mathbf{b}\|^2-\delta^2,\varepsilon),$$

where $\phi(s,\varepsilon)$ is a smoothing function of the plus function s_+ and $\theta(s,\varepsilon)$ is a smoothing function of the absolute value function $|s|$. Note that $\phi(s,\varepsilon)$ can be defined by any one of the following formulations:

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

$$\phi(s,\varepsilon) = \begin{cases} s_+ & \text{if } |s| > \varepsilon \\ \dfrac{(s+\varepsilon)^2}{4\varepsilon} & \text{if } |s| \leq \varepsilon, \end{cases} \qquad \phi(s,\varepsilon) = \begin{cases} s + \dfrac{\varepsilon}{2}e^{-\frac{s}{\varepsilon}} & \text{if } s > 0 \\ \dfrac{\varepsilon}{2}e^{\frac{s}{\varepsilon}} & \text{if } s \leq 0, \end{cases}$$

and $\theta(s,\varepsilon)$ can be given by $\theta(s,\varepsilon) = \phi(s,\varepsilon) + \phi(-s,\varepsilon)$. From Definition 4.1, it is clear that $\tilde{c}$ is a smoothing convex-concave function of c. Moreover, by Proposition 4.1, it satisfies (4.23) and (4.24).

6.2 Robust bond portfolio construction

We consider a portfolio of n bonds with quantities $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^n_+$ and time periods $t = 1,\ldots,T$, where the set $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \underline{\mathbf{u}} \leq \mathbf{x} \leq \overline{\mathbf{u}}\}$ gives a range of possible quantities for each bond. Let $\alpha_{i,t}$ denote the cash flow from bond i in period t, which includes the coupon payments and the payment of the face value at maturity.

Let $p \in \mathbb{R}^n_+$ denote the price of the bonds with

$$p_i = \sum_{t=1}^{T}\alpha_{i,t}\exp(-t(u_t+s_i)), \quad i = 1,\ldots,n,$$

where $s_i \geq 0$ is the spread for bond i and u_t is the yield curve at time t. The portfolio value is given by $p^\top\mathbf{x}$. Let ϕ be a smooth convex nominal function that may include tracking error against a benchmark, a risk term and a transaction cost term.

Let $\mathbf{y} = (u^\top, s^\top)^\top \in \mathbb{R}^{n+T}$. The set $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^{n+T} : \underline{\mathbf{v}} \leq \mathbf{y} \leq \overline{\mathbf{v}}\}$ gives a range of possible values for each point in the yield curve and for each spread.

A version of the robust bond portfolio construction model in [35] is the following convex-concave saddle point problem

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} c(\mathbf{x},\mathbf{y}) := \phi(\mathbf{x}) - \lambda\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{x}_i\alpha_{i,t}\exp(-t(\mathbf{y}_t+\mathbf{y}_{T+i})) - \beta\|\mathbf{A}\mathbf{y}-\mathbf{b}\|_1,$$

where $\|\mathbf{A}\mathbf{y}-\mathbf{b}\|_1$ describes the uncertainties in yield curves and spreads, and $c(\mathbf{x},\mathbf{y})$ is a nonsmooth function with respect to $\mathbf{y}$. A robust bond portfolio construction with sparse selection of bonds is as follows

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} c(\mathbf{x},\mathbf{y}) + \lambda_1\|\mathbf{x}\|_0. \tag{6.3}$$

Problem (6.3) is a nonsmooth convex-concave saddle point problem with cardinality penalty $\|\mathbf{x}\|_0$, where $\mathcal{X}$ is a convex set with $\mathrm{int}(\mathcal{X})\neq\emptyset$, and $\mathcal{Y}$ is a convex set. Note that the assumption $\mathrm{int}(\mathcal{Y})\neq\emptyset$ in (5.2) and Assumption 5.1-(ii) can be removed, since (6.3) does not have a cardinality function of $\mathbf{y}$. A smoothing function of $\|\mathbf{A}\mathbf{y}-\mathbf{b}\|_1$ in the function c can be constructed by $\theta(s,\varepsilon)$ in subsection 6.1.

6.3 Sparse convex-concave logistic regression saddle point problems

Motivated by the unconstrained convex-concave logistic regression saddle point problem in [5], we consider the following saddle point problem

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} c(\mathbf{x},\mathbf{y}) := \sum_{k=1}^{N}\log(1+e^{-\alpha_k\mathbf{a}_k^\top\mathbf{x}}) + \mathbf{x}^\top\mathbf{A}\mathbf{y} - \sum_{k=1}^{N}\log(1+e^{-\beta_k\mathbf{b}_k^\top\mathbf{y}}), \tag{6.4}$$

where $\mathcal{X}=\{\mathbf{x}:\|\mathbf{x}\|_\infty\le 1\}$, $\mathcal{Y}=\{\mathbf{y}:\|\mathbf{y}\|_\infty\le 1\}$, $\mathbf{a}_k\in\{0,1\}^n, \mathbf{b}_k\in\{0,1\}^m$, $\mathbf{A}\in\{0,1\}^{n\times m}$ and $\alpha_k,\beta_k\in\{-1,1\}$, for all $k\in[N]$. To find a sparse solution, we consider the following min-max model

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}} c(\mathbf{x},\mathbf{y}) + \lambda_1\|\mathbf{x}\|_0 - \lambda_2\|\mathbf{y}\|_0 \tag{6.5}$$

with $\lambda_1>0$ and $\lambda_2>0$. It is clear that c is a smooth convex-concave function and Assumption 5.1 holds for (6.5) with $\tau=\sigma=1$. It has

$$\nabla_{\mathbf{x}}c(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{N}\frac{-\alpha_k e^{-\alpha_k\mathbf{a}_k^\top\mathbf{x}}}{1+e^{-\alpha_k\mathbf{a}_k^\top\mathbf{x}}}\mathbf{a}_k + \mathbf{A}\mathbf{y},\ \nabla_{\mathbf{y}}c(\mathbf{x},\mathbf{y}) = -\sum_{k=1}^{N}\frac{-\beta_k e^{-\beta_k\mathbf{b}_k^\top\mathbf{y}}}{1+e^{-\beta_k\mathbf{b}_k^\top\mathbf{y}}}\mathbf{b}_k + \mathbf{A}^\top\mathbf{x}.$$

By simple calculation, we can set $L_{c,1}$ in (5.22) by

$$L_{c,1} = \max\left\{\|\mathbf{a}\|_\infty + \|\mathbf{A}\|_\infty, \|\mathbf{b}\|_\infty + \|\mathbf{A}^\top\|_\infty, 1\right\}, \tag{6.6}$$

where $\mathbf{a}=(\mathbf{a}_1,\ldots,\mathbf{a}_N)$ and $\mathbf{b}=(\mathbf{b}_1,\ldots,\mathbf{b}_N)$.

If we choose the density function in Example 3.1 with $p = 1$ to construct continuous relaxation function $f^R(\mathbf{x}, \mathbf{y}, \mu)$, then $\alpha = \underline{\rho} = 1$. From the weak d-stationary point defined in Definition 5.1 and by Theorem 5.2, if $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{X} \times \mathcal{Y}$ is a weak d-stationary point of $\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} f^R(\mathbf{x}, \mathbf{y}, \mu)$, then $(\mathbf{x}^*, \mathbf{y}^*)$ is a μ-strong local saddle point of (6.5), that is

$$\begin{cases} |\mathbf{x}_i^*| \notin (0, \mu),\ |\mathbf{y}_j^*| \notin (0, \mu), & \forall i \in [n],\ j \in [m], \\ 0 \in [\nabla_{\mathbf{x}} c(\mathbf{x}^*, \mathbf{y}^*)]_i + N_{\mathcal{X}_i}(\mathbf{x}_i^*), & \text{for } i \in [n] \text{ satisfying } |\mathbf{x}_i^*| \geq \mu, \\ 0 \in -[\nabla_{\mathbf{y}} c(\mathbf{x}^*, \mathbf{y}^*)]_j + N_{\mathcal{Y}_j}(\mathbf{y}_j^*), & \text{for } j \in [m] \text{ satisfying } |\mathbf{y}_j^*| \geq \mu. \end{cases} \tag{6.7}$$

There are many interesting algorithms for min-max problems [1, 4, 15, 25, 36, 42, 48, 49]. To illustrate our theoretical results, we solve convex-concave min-max problem (6.4) by the Proximal Gradient Descent Ascent (PGDA) algorithm proposed in [15] as follows

$$\mathbf{x}^{k+1} = \arg\min_{\mathbf{x}\in\mathcal{X}} Q(\mathbf{x}, \mathbf{x}^k; \mathbf{y}^k, \mathbf{y}^k), \qquad \mathbf{y}^{k+1} = \arg\max_{\mathbf{y}\in\mathcal{Y}} Q(\mathbf{x}^{k+1}, \mathbf{x}^k; \mathbf{y}, \mathbf{y}^k),$$

where

$$Q(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{y}, \tilde{\mathbf{y}}) := \langle \nabla_{\mathbf{x}} c(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{x} - \tilde{\mathbf{x}} \rangle + \langle \nabla_{\mathbf{y}} c(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}), \mathbf{y} - \tilde{\mathbf{y}} \rangle + \frac{1}{2}\gamma \|\mathbf{x} - \tilde{\mathbf{x}}\|^2 - \frac{1}{2}\gamma \|\mathbf{y} - \tilde{\mathbf{y}}\|^2,$$

and $\gamma \geq \max\{\|\mathbf{a}\|_\infty, \|\mathbf{b}\|_\infty\} \geq \max_{\mathbf{x}\in\mathcal{X}, \mathbf{y}\in\mathcal{Y}, i\in[n], j\in[m]} \{|[\nabla^2_{\mathbf{xx}} c(\mathbf{x}, \mathbf{y})]_{ii}|, |[\nabla^2_{\mathbf{yy}} c(\mathbf{x}, \mathbf{y})]_{jj}|\}$. If $(\mathbf{x}^k, \mathbf{y}^k)$ generated by PGDA converges to $(\bar{\mathbf{x}}, \bar{y})$, then

$$\mathbf{0} \in \nabla_{\mathbf{x}} c(\bar{\mathbf{x}}, \bar{\mathbf{y}}) + N_{\mathcal{X}}(\bar{\mathbf{x}}), \quad \mathbf{0} \in -\nabla_{\mathbf{y}} c(\bar{\mathbf{x}}, \bar{\mathbf{y}}) + N_{\mathcal{Y}}(\bar{\mathbf{y}}),$$

which implies that $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a saddle point of (6.4) by the convexity-concavity of c.

To find a sparse local saddle point of (6.5), we define the continuous relaxation

$$Q_{d_{\tilde{\mathbf{x}}}, d_{\tilde{\mathbf{y}}}}(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{y}, \tilde{\mathbf{y}}; \mu) := Q(\mathbf{x}, \tilde{\mathbf{x}}; \mathbf{y}, \tilde{\mathbf{y}}) + \lambda_1 \sum_{i=1}^{n} \Phi^{d_{\tilde{\mathbf{x}}_i}}(\mathbf{x}_i, \mu) - \lambda_2 \sum_{j=1}^{m} \Phi^{d_{\tilde{\mathbf{y}}_j}}(\mathbf{y}_j, \mu), \tag{6.8}$$

where $\Phi^{d_{\tilde{s}}}(s, \mu) = \begin{cases} \frac{1}{\mu}|s| & \text{if } |\tilde{s}| < \mu \\ 1 & \text{if } |\tilde{s}| \geq \mu. \end{cases}$ Notice that for fixed $\tilde{\mathbf{x}} \in \mathcal{X}$, $\tilde{\mathbf{y}} \in \mathcal{Y}$ and $\mu > 0$, $Q_{d_{\tilde{\mathbf{x}}}, d_{\tilde{\mathbf{y}}}}(\cdot, \tilde{\mathbf{x}}; \cdot, \tilde{\mathbf{y}}; \mu)$ is convex-concave.

Combining the PGDA with the alternating index at $\mathbf{x}^k$ and $\mathbf{y}^k$ in [3], we propose the following Alternating Proximal Gradient Descent Ascent (APGDA) algorithm

$$\begin{aligned} \mathbf{x}^{k+1} &= \arg\min_{\mathbf{x}\in\mathcal{X}} Q_{d_{\mathbf{x}^k}, d_{\mathbf{y}^k}}(\mathbf{x}, \mathbf{x}^k; \mathbf{y}^k, \mathbf{y}^k; \mu), \\ \mathbf{y}^{k+1} &= \arg\max_{\mathbf{y}\in\mathcal{Y}} Q_{d_{\mathbf{x}^{k+1}}, d_{\mathbf{y}^k}}(\mathbf{x}^{k+1}, \mathbf{x}^k; \mathbf{y}, \mathbf{y}^k; \mu). \end{aligned}$$

Although the two steps in APGDA can be considered as a generalization Algorithm 3.1 in [3] for solving two minimization problems: $\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x},\mathbf{y}^k)$ and $\min_{\mathbf{y}\in\mathcal{Y}} -f(\mathbf{x}^{k+1},\mathbf{y})$, the convergence analysis of APGDA is not trivial. In what follows, we give a preliminary result on the convergence of APGDA.

Proposition 6.1. *For any initial point $(\mathbf{x}^0,\mathbf{y}^0)\in\mathcal{X}\times\mathcal{Y}$, if $(\mathbf{x}^k,\mathbf{y}^k)$ generated by APGDA with $\mu < \frac{\min\{\lambda_1,\lambda_2\}}{L_{c,1}}$ converges to a point $(\bar{\mathbf{x}},\bar{\mathbf{y}})$, then it is a μ-strong local saddle point of (6.5).*

Proof. Let $\Phi_1(s,\mu) = |s|/\mu$ and $\Phi_2(s,\mu) = 1$. There is a subsequence of $\{(\mathbf{x}^k,\mathbf{y}^k)\}$ (also denoted by $\{(\mathbf{x}^k,\mathbf{y}^k)\}$) and vectors $t\in\mathbb{R}^n$, $w\in\mathbb{R}^m$ with $t_i, w_j\in\{1,2\}$ such that $\Phi^{d_{\mathbf{x}_i^k}}(s,\mu) = \Phi_{t_i}(s,\mu)$ and $\Phi^{d_{\mathbf{y}_j^k}}(s,\mu) = \Phi_{w_j}(s,\mu)$ for any $i\in[n]$, $j\in[m]$ and $k\in\mathbb{N}$. From APGDA, we have

$$
\begin{aligned}
&0\in[\nabla_{\mathbf{x}}c(\mathbf{x}^k,\mathbf{y}^k)]_i + \gamma(\mathbf{x}_i^{k+1}-\mathbf{x}_i^k) + \lambda_1\nabla_s\Phi_{t_i}(\mathbf{x}_i^{k+1},\mu) + N_{\mathcal{X}_i}(\mathbf{x}_i^{k+1}), \quad \forall i\in[n],\\
&0\in-[\nabla_{\mathbf{y}}c(\mathbf{x}^k,\mathbf{y}^k)]_j + \gamma(\mathbf{y}_j^{k+1}-\mathbf{y}_j^k) + \lambda_2\nabla_s\Phi_{w_j}(\mathbf{y}_j^{k+1},\mu) + N_{\mathcal{Y}_j}(\mathbf{y}_j^{k+1}), \quad \forall j\in[m].
\end{aligned}
$$

Letting $k\to\infty$, for any $i\in[n]$ and $j\in[m]$, we obtain

$$
\begin{aligned}
&0\in[\nabla_{\mathbf{x}}c(\bar{\mathbf{x}},\bar{\mathbf{y}})]_i + \lambda_1\nabla_s\Phi_{t_i}(\bar{\mathbf{x}}_i,\mu) + N_{\mathcal{X}_i}(\bar{\mathbf{x}}_i),\\
&0\in-[\nabla_{\mathbf{y}}c(\bar{\mathbf{x}},\bar{\mathbf{y}})]_j + \lambda_2\nabla_s\Phi_{w_j}(\bar{\mathbf{y}}_j,\mu) + N_{\mathcal{Y}_j}(\bar{\mathbf{y}}_j).
\end{aligned}
\tag{6.9}
$$

If there exists $i\in[n]$ such that $|\bar{\mathbf{x}}_i|\in(0,\mu)$, then $\bar{\mathbf{x}}_i\in\mathrm{int}(\mathcal{X}_i)$ and the first inclusion in (6.9) gives $L_{c,1}\le\frac{\lambda_1}{\mu}$, which leads a contradiction. Thus, $|\bar{\mathbf{x}}_i|\notin(0,\mu)$, $\forall i\in[n]$. Similarly, $|\bar{\mathbf{y}}_j|\notin(0,\mu)$, $\forall j\in[m]$.

If there exists $i\in[n]$ such that $|\bar{\mathbf{x}}_i|=\mu$ and $t_i=1$, then the first inclusion in (6.9) also brings a contradiction to the value of μ. Putting forward these results into (6.9) gives

$$
\begin{aligned}
&|\bar{\mathbf{x}}_i|\notin(0,\mu), \quad |\bar{\mathbf{y}}_j|\notin(0,\mu), \quad \forall i\in[n],\, j\in[m],\\
&0\in[\nabla_{\mathbf{x}}c(\bar{\mathbf{x}},\bar{\mathbf{y}})]_i + N_{\mathcal{X}_i}(\bar{\mathbf{x}}_i), \quad \forall i\in[n],\,\bar{\mathbf{x}}_i\neq 0,\\
&0\in[-\nabla_{\mathbf{y}}c(\bar{\mathbf{x}},\bar{\mathbf{y}})]_j + N_{\mathcal{Y}_j}(\bar{\mathbf{y}}_j), \quad \forall j\in[m],\,\bar{\mathbf{y}}_j\neq 0.
\end{aligned}
$$

Recalling the results in Theorem 2.1, we confirm that $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ is a μ-strong local saddle point of (6.5). □

For a given point, to determine whether it is a saddle point of (6.4) or a μ-strong local saddle point of (6.5), by the normal cones of $\mathcal{X}=\{\mathbf{x}:\|\mathbf{x}\|_\infty\le 1\}$ and $\mathcal{Y}=\{\mathbf{y}:\|\mathbf{y}\|_\infty\le 1\}$, we define the following evaluation functions

$$
R_i(\mathbf{x}) = \begin{cases} ([\nabla_{\mathbf{x}}c(\mathbf{x},\mathbf{y})]_i)_+ & \text{if } \mathbf{x}_i = 1\\ (-[\nabla_{\mathbf{x}}c(\mathbf{x},\mathbf{y})]_i)_+ & \text{if } \mathbf{x}_i = -1\\ |[\nabla_{\mathbf{x}}c(\mathbf{x},\mathbf{y})]_i| & \text{otherwise,}\end{cases}
\quad
S_j(\mathbf{y}) = \begin{cases} (-[\nabla_{\mathbf{y}}c(\mathbf{x},\mathbf{y})]_j)_+ & \text{if } \mathbf{y}_j = 1\\ ([\nabla_{\mathbf{y}}c(\mathbf{x},\mathbf{y})]_j)_+ & \text{if } \mathbf{y}_j = -1\\ |[\nabla_{\mathbf{y}}c(\mathbf{x},\mathbf{y})]_j| & \text{otherwise.}\end{cases}
$$

It is clear that $R_i(\mathbf{x})\ge 0$, $\forall i\in[n]$ and $S_j(\mathbf{y})\ge 0$, $\forall j\in[m]$. For $\bar{\mathbf{x}}\in\mathcal{X}$ and $\bar{\mathbf{y}}\in\mathcal{Y}$, it holds

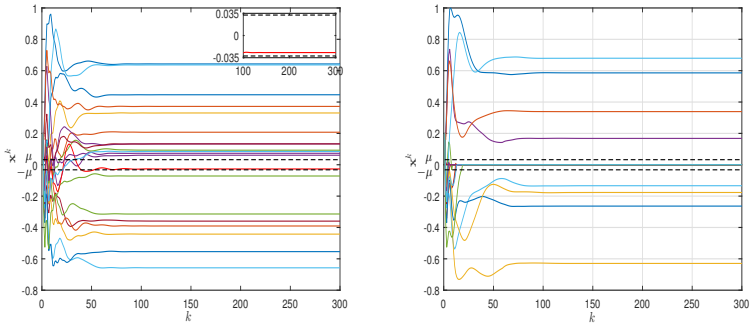
- $p(\bar{\mathbf{x}}) := \sum_{i=1}^{n} R_i(\bar{\mathbf{x}}) = 0$ and $q(\bar{\mathbf{y}}) := \sum_{j=1}^{m} S_j(\bar{\mathbf{y}}) = 0$ if and only if $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a saddle point of (6.4).
- $\tilde{p}(\bar{\mathbf{x}}) := \sum_{i:\bar{\mathbf{x}}_i \neq 0}(R_i(\bar{\mathbf{x}}) + \max\{\mu - |\bar{\mathbf{x}}_i|, 0\}) = 0$ and $\tilde{q}(\bar{\mathbf{y}}) := \sum_{j:\bar{\mathbf{y}}_j \neq 0}(S_j(\bar{\mathbf{y}}) + \max\{\mu - |\bar{\mathbf{y}}_j|, 0\}) = 0$ if and only if $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a μ-strong local saddle point of (6.5).

Although the sequence convergence of PGDA and APGDA cannot be guaranteed, we can compare the behaviour of the sequences generated by PGDA and APGDA from the same initial points. We conduct a simple test experiment with $n = 20$, $m = 30$, $N = 50$, $\lambda_1 = \lambda_2 = 1$ in Matlab. We randomly generate a binary matrix $\mathbf{A} \in \{0, 1\}^{n \times m}$, and for $k \in [N]$ randomly generate $\mathbf{a}_k \in \{0, 1\}^n$, $\mathbf{b}_k \in \{0, 1\}^m$ with 2 nonzero elements, $\alpha_k, \beta_k \in \{-1, 1\}$. We compute the constant $L_{c,1}$ as in (6.6) and obtain $\mu = 0.0323 < \bar{\mu}_1$, where $\bar{\mu}_1 = \min\{1, 1/L_{c,1}\}$ is defined as in (5.30). We choose an initial point $(\mathbf{x}^0, \mathbf{y}^0) = 0.2\mathbf{e}$ for running both PGDA and APGDA.

In Fig. 2, (a), (c), (e) plot convergent sequences of $\mathbf{x}^k$, $\mathbf{y}^k$, $p(\mathbf{x}^k)$, $q(\mathbf{y}^k)$ generated by PGDA and (b), (d), (f) plot convergent sequences of $\mathbf{x}^k$, $\mathbf{y}^k$, $\tilde{p}(\mathbf{x}^k)$, $\tilde{q}(\mathbf{y}^k)$ generated by APGDA, where each curve in (a)-(d) represents one component of the corresponding vectors. From Fig. 2, we find that the limit point of the sequence $(\mathbf{x}^k, \mathbf{y}^k)$ generated by PGDA does not have a zero element and some elements of it do not satisfy the lower bounds in (6.7). However, more than half elements of the limit point of $(\mathbf{x}^k, \mathbf{y}^k)$ generated by APGDA are zero, and all elements of it satisfy the lower bounds in (6.7). This is consistent with the theoretical results and shows the superiority of (6.5) in finding a sparse solution. Moreover, in Fig. 2-(e), from the convergence of $p(\mathbf{x}^k)$ and $q(\mathbf{y}^k)$ on $(\mathbf{x}^k, \mathbf{y}^k)$ generated by PGDA, we confirm that the limit point of $(\mathbf{x}^k, \mathbf{y}^k)$ is a saddle point of (6.4), while Fig. 2-(f) shows the convergence of $\tilde{p}(\mathbf{x}^k)$ and $\tilde{q}(\mathbf{y}^k)$ on $(\mathbf{x}^k, \mathbf{y}^k)$ generated by APGDA, which confirms that the limit point of this sequence is a μ-strong local saddle point of (6.5).

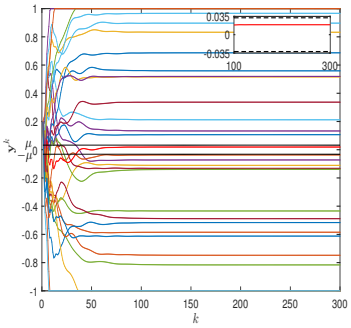
7 Conclusion

In this paper, we prove the existence of local saddle points and global minimax points of problem (1.1) and define a class of strong local saddle points of it. To construct interesting continuous relaxations to (1.1) based on convolution, we introduce two classes of density functions which satisfy Assumptions 3.1 and 3.2, respectively. The induced continuous relaxations include the capped-ℓ_p with $0 < p \leq 1$, scaled SCAD, scaled MCP, hard thresholding functions as special cases. Moreover, we establish the relations between problem (1.1) and its continuous relaxation (3.12) regarding their saddle points, local saddle points and global minimax points by using the lower bound properties of $g(\mathbf{x})$ and $h(\mathbf{y})$ in (4.1)-(4.2) at the local saddle points and global minimax points of the continuous relaxation problem. Moreover, we define the weak d-stationary points and weak second order stationary points of problem (5.1), which are necessary conditions for the local saddle points of its continuous relaxation problem (5.9), while sufficient conditions for the strong local saddle points of (5.1). In addition, we study the smoothing approximation of (5.9) by using a smoothing convex-concave function of nonsmooth c and prove that any accumulation point of weak d-stationary points of the smoothing approximation problem is a weak d-stationary point of (5.9) as the smoothing parameter goes to zero.

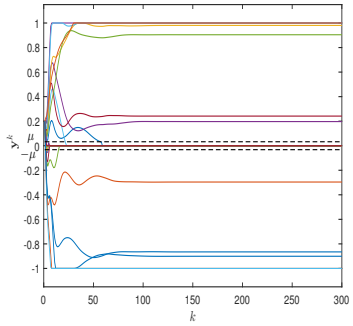


(a) $\mathbf{x}^k$ by PGDA

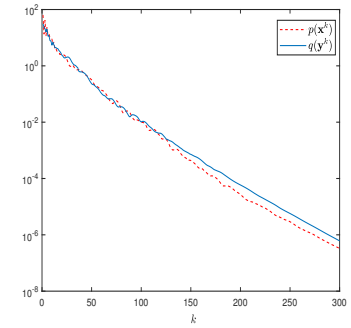
(b) $\mathbf{x}^k$ by APGDA



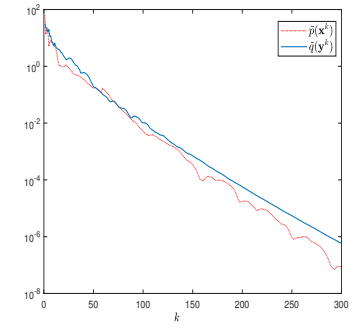
(c) $\mathbf{y}^k$ by PGDA



(d) $\mathbf{y}^k$ by APGDA



(e) $p(\mathbf{x}^k)$ and $q(\mathbf{y}^k)$ by PGDA



(f) $\tilde{p}(\mathbf{x}^k)$ and $\tilde{q}(\mathbf{y}^k)$ by APGDA

Fig. 2: Convergence of $\mathbf{x}^k$, $\mathbf{y}^k$, $p(\mathbf{x}^k)$ and $q(\mathbf{y}^k)$ generated by PGDA and convergence of $\mathbf{x}^k$, $\mathbf{y}^k$, $\tilde{p}(\mathbf{x}^k)$ and $\tilde{q}(\mathbf{y}^k)$ generated by APGDA

References

[1] Attouch, H., Wets, R.J.B.: A convergence theory for saddle functions. Trans. Am. Math. Soc. **280**, 1–44 (1983)

[2] Beck, A., Hallak, N.: Optimization problems involving group sparsity terms. Math. Program. **178**, 39–67 (2019)

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

[4] Bot, R.I., Bohm, A.: Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. SIAM J. Optim. **33**, 1884–1913 (2023)

[5] Bullins, B., Lai, K.A.: Higher-order methods for convex-concave min-max optimization and monotone variational inequalities. SIAM J. Optim. **32**, 2208–2229 (2022)

[6] Burke, J.V., Chen, X., Sun, H.: The subdifferential of measurable composite max integrands and smoothing approximation. Math. Program. **181**, 229–264 (2020)

[7] Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**, 34–81 (2009)

[8] Campbell, S.L., Gear, C.W.: The index of general nonlinear DAES. Numer. Math. **72**, 173–196 (1995)

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

[10] Chen, C., Mangasarian, O.L.: A class of smoothing functions for nonlinear and mixed complementarity problems. Comput. Optim. Appl. **5**, 97–138 (1996)

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

[12] Chen, X., Kelley, C.T.: Min-max optimization for robust nonlinear least squares problems. http://arxiv.org/abs/2402.12679 (2024)

[13] Chen, X., Xu, F., Ye, Y.: Lower bound theory of nonzero entries in solutions of l_2-l_p minimization. SIAM J. Sci. Comput. **32**, 2832–2852 (2010)

[14] Clarke, F.H.: Optimization and Nonsmooth Analysis. SIAM, New York (1990)

[15] Cohen, E., Teboulle, M.: Alternating and parallel proximal gradient methods for nonsmooth, nonconvex minimax: a unified convergence analysis. Math. Oper. Res.

(2024) https://doi.org/10.1287/moor.2022.0294

[16] Cui, Y., Liu, J., Pang, J.-S.: The minimization of piecewise functions: pseudo stationarity. Convex Anal. **30**, 793–834 (2023)

[17] Cui, Y., Pang, J.-S.: Modern Nonconvex Nondifferentiable Optimization. SIAM, New York (2021)

[18] Dai, Y., Zhang, L.: Optimality conditions for constrained minimax optimization. CSIAM Trans. Appl. Math. **1**, 296–315 (2020)

[19] Debreu, G.: Saddle point existence theorems. In: RCowles Commission Discussion Paper: Mathematicas No.412 (1952)

[20] Facchinei, F., Pang, J.-S.: Finite-Dimensional Variational Inequalities and Complementarity Problems, Volume I. Springer, New York (2002)

[21] Fan, K.: Minimax theorems. Proc. Nat. Acad. Sci. **39**, 42–47 (1953)

[22] Fan, J.: Comments on "Wavelets in statistics: A review" by A. Antoniadis. J. Italian Statist. Soc. **6**, 131–138 (1997)

[23] Goodfellow, I.: NIPS2016 totorial: Generative adversarial networks. http://arxiv.org/abs/1701.00160 (2016)

[24] Goodfellow, I., Pouget-Abadie, J., Mirza, M., etc.: Generative adversarial nets. in Advances in Neural Information Processing Systems **27**, 2672–2680 (2014)

[25] Grimmer, B., Lu, H., Worah, P., Mirrokni, V.: The landscape of the proximal point method for nonconvex-nonconcave minimax optimization. Math. Program. **201**, 373–407 (2023)

[26] Guo, L., Ye, J.J., Zhang, J.: Sensitivity analysis of the maximal value function with applications in nonconvex minimax programs. Math. Oper. Res. **49**, 536–556, (2024)

[27] Han, S., Cui, Y., Pang, J.-S: Analysis of a class of minimization problems lacking lower semicontinuity. https://optimization-online.org/wp-content/uploads/2023/10/non-lsc.pdf (2023)

[28] He, Z., Han, S., Gómez, A., Cui, Y., Pang, J.-S.: Comparing solution paths of sparse quadratic minimization with a Stieltjes matrix. Math. Program. **204**, 517–566 (2024)

[29] Hiriart-Urruty, J.B., Lemaréchal, C.: Fundamentals of Convex Analysis. Springer, Berlin (2001)

[30] Hiriart-Urruty, J.B., Strodiot, J.-J., Nguyen, V.H.: Generalized Hessian matrix

and second-order optimality conditions for problems with $C^{1,1}$ data. Appl. Math. Optim. **11**, 43–56 (1984)

[31] Jiang, J., Chen, X.: Optimality conditions for nonsmooth nonconvex-nonconcave min-max problems and generative adversarial networks. SIAM J. Math. Data Sci. **5**, 693–722 (2023)

[32] Jiao, Y., Jin, B., Lu, X.: A primal dual active set with continuation algorithm for the ℓ^0-regularized optimization problem. Appl. Comput. Harmon. Anal. **39**, 400–426 (2015)

[33] Jin, C., Netrapalli, P., Jordan, M.I.: What is local optimality in nonconvex-nonconcave minimax optimization? In: Proceedings of the 37th International Conference on Machine Learning, pp. 4880–4889 (2020)

[34] Kanzow, C., Schwartz, A., Weiß, F.: The sparse(st) optimization problem: reformulations, optimality, stationarity, and numerical results. http://arxiv.org/abs/2210.09589 (2022)

[35] Luxenberg, E., Schiele, P., Boyd, S.: Robust bond portfolio construction via convex-concave saddle point optimization. To appear in J. Optim. Theory Appl. (2024)

[36] Nemirovski, A.: Prox-method with rate of convergence $o(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. SIAM J. Optim. **15**, 229–251 (2004)

[37] Neumann, J.: Zur Theorie der Gesellschaftsspiele. Math. Ann. **100**, 295–320 (1928)

[38] Nikaido, H.: On von Neumann's minimax theorem. Pacific J. Math. **4**, 65–72 (1954)

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

[40] Peleg, D., Meir, R.: A bilinear formulation for vector sparsity optimization. Signal Process. **88**, 375–389 (2008)

[41] Qi, L., Chen, X.: A globally convergent successive approximation method for severely nonsmooth equations. SIAM J. Control Optim. **33**, 402–418 (1995)

[42] Rockafellar, R.T.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**, 877–898 (1976)

[43] Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton, NJ (1997)

[44] Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin, Germany (1998)

[45] Shiffiman, M.: On the equality $\min\max = \max\min$, and the theory of games. In: RAND Report, RM-243 (1949)

[46] Sion, M.: On general minimax theorems. Pacific J. Math. **8**, 171–176 (1958)

[47] Soubies, E., Blanc-Féraud, L., Aubert, G.: A unified view of exact continuous penalties for ℓ_2-ℓ_0 minimization. SIAM J. Optim. **27**, 2034–2060 (2017)

[48] Xu, Z., Zhang, H., Xu, Y., Lan, G.: A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. Math. Program. **201**, 635–706 (2022)

[49] Yang, J., Orvieto, A., Lucchi, A., He, N.: Faster single-loop algorithms for minimax optimization without strong concavity. In: Proceedings of Machine Learning Research, 7377–7389 (2022)

[50] Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. **38**, 894–942 (2010)