Mathematical Programming (2025) 212:479–518
https://doi.org/10.1007/s10107-024-02113-z

FULL LENGTH PAPER

Series A

Unified smoothing approach for best hyperparameter selection problem using a bilevel optimization strategy

Jan Harold Alcantara[1] · Chieu Thanh Nguyen[2] · Takayuki Okuno[1,3] · Akiko Takeda[1,4] · Jein-Shan Chen[5]

Received: 21 October 2021 / Accepted: 24 May 2024 / Published online: 8 August 2024

Abstract

Strongly motivated from applications in various fields including machine learning, the methodology of sparse optimization has been developed intensively so far. Especially, the advancement of algorithms for solving problems with nonsmooth regularizers has been remarkable. However, those algorithms suppose that weight parameters of regularizers, called hyperparameters hereafter, are pre-fixed, but it is a crucial matter how the best hyperparameter should be selected. In this paper, we focus on the hyperparameter selection of regularizers related to the ℓ_p function with $0 < p \leq 1$ and apply a bilevel programming strategy, wherein we need to solve a bilevel problem, whose lower-level problem is nonsmooth, possibly nonconvex and non-Lipschitz. Recently, for solving a bilevel problem for hyperparameter selection of the pure ℓ_p $(0 < p \leq 1)$ regularizer Okuno et al. discovered new necessary optimality conditions, called SB(scaled bilevel)-KKT conditions, and further proposed a smoothing-type algorithm using a specific smoothing function. However, this opti-

✉ Jein-Shan Chen
jschen@math.ntnu.edu.tw

Jan Harold Alcantara
janharold.alcantara@riken.jp

Chieu Thanh Nguyen
ntchieu@vnua.edu.vn

Takayuki Okuno
takayuki-okuno@st.seikei.ac.jp

Akiko Takeda
takeda@mist.i.u-tokyo.ac.jp

1 Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan

2 Department of Mathematics, Faculty of Information Technology, Vietnam National University of Agriculture, Hanoi 131000, Vietnam

3 Faculty of Science and Technology, Seikei University, Tokyo 180-8633, Japan

4 Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan

5 Department of Mathematics, National Taiwan Normal University, Taipei 11677, Taiwan

mality measure is loose in the sense that there could be many points that satisfy the SB-KKT conditions. In this work, we propose new bilevel KKT conditions, which are new necessary optimality conditions tighter than the ones proposed by Okuno et al. Moreover, we propose a unified smoothing approach using smoothing functions that belong to the Chen-Mangasarian class, and then prove that generated iteration points accumulate at bilevel KKT points under milder constraint qualifications. Another contribution is that our approach and analysis are applicable to a wider class of regularizers. Numerical comparisons demonstrate which smoothing functions work well for hyperparameter optimization via bilevel optimization approach.

Keywords Hyperparameter learning · Smoothing functions · Bilevel optimization

Mathematics Subject Classification 90C26 · 90C30

1 Introduction

A learning algorithm in machine learning usually involves solving the unconstrained optimization problem

$$\min_{\omega\in\Re^n} g(\omega) + \sum_{i=1}^{r} \lambda_i R_i(\omega), \tag{1}$$

where $\lambda = (\lambda_1, \ldots, \lambda_r)$ is called a hyperparameter, whose value is decided prior to implementation of the algorithm. Here, $R_i, g : \Re^n \to \Re$, $i = 2, \ldots, r$ are twice continuously differentiable functions, and

$$R_1(\omega) := \sum_{i=1}^{n} \psi\left(|\omega_i|^p\right) \quad (0 < p \leq 1), \tag{2}$$

with ψ satisfying the following assumption:

Assumption (A). $\psi : [0, \infty) \to \Re$ is twice continuously differentiable on $[0, \infty)$ and there exist two positive constants α, β such that $0 < \psi'(t) \leq \alpha$ and $-\beta \leq \psi''(t) \leq 0$ for all $t \in [0, \infty)$.

In this manuscript, we make Assumption (A) our blanket assumption on ψ. There are many penalty functions often used in statistics and signal reconstruction satisfying Assumption (A) (see Appendix A). It is well-known that the function $R_1(\omega)$ is nonsmooth, nonconvex, and even non-Lipschitz when $p \in (0, 1)$.

For notation purposes, we denote

$$G(\omega, \bar{\lambda}) := g(\omega) + \bar{\lambda}^T \bar{R}(w),$$

with $\bar{\lambda} := (\lambda_2, \ldots, \lambda_r)^T \in \Re^{r-1}$ and $\bar{R} : \Re^n \to \Re^{r-1}$ given by $\bar{R}(w) := (R_2(w), \ldots, R_r(w))^T$. Then problem (1) can be rewritten as

$$\min_{\omega\in\Re^n} G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega). \tag{3}$$

The problem of finding the optimal values of the hyperparameters for (3) can be accomplished using grid search and Bayesian optimization [2, 27]. This paper, on the other hand, is devoted to a bilevel optimization strategy to find the best hyperparameter. In particular, we focus on the bilevel nonsmooth programming problem

$$\begin{aligned} \min_{\omega_\lambda^*, \lambda} \quad & f(\omega_\lambda^*) \\ \text{s.t} \quad & \omega_\lambda^* \in \operatorname*{argmin}_{\omega \in \Re^n} G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega) \\ & (\lambda_1, \bar{\lambda}) \in \Omega_\epsilon \subset \Re^r, \end{aligned} \tag{4}$$

where $f : \Re^n \to \Re$ is continuously differentiable and

$$\Omega_\epsilon := \{(\lambda_1, \bar{\lambda}) \in \Re \times \Re^{r-1} : \lambda_1 \geq \epsilon,\ \bar{\lambda} \geq 0\}, \tag{5}$$

for some small parameter $\epsilon > 0$. Problem (3) that appears in the constraint set of (4) is called the *lower-level problem*, and the minimization of f is called the *upper-level problem*. Note that in the interest of obtaining sparse models, we impose a strict positive lower bound condition for the parameter λ_1 corresponding to the sparsity-promoting regularizer R_1.

Bilevel optimization problems were introduced by Bracken and McGill [5]. The reader is referred to [10, 11, 24] for a survey of methods for solving the bilevel optimization problem as well as their applications. Significant efforts have been put forth by many researchers in the past few decades to use bilevel optimization strategy to the problem of finding the best hyperparameter values. In particular, [3, 4] focused on a bilevel support-vector regression (SVR) problem where the lower-level optimization problem is cast as a convex quadratic program. The authors in [17, 18] proposed a bilevel cross-validation program for support-vector machine (SVM), where the upper-level problem is convex and nonsmooth, while the lower-level problem is differentiable. [20] used gradient-based methods for the bilevel optimization problem with nonsmooth convex lower-level problem (for example, sparse models based on the ℓ_1-norm). However, [3, 17, 20] only provided algorithms to solve the bilevel optimization, and theoretical guarantees are not established. [26] formulated the hyperparameter optimization problem through K-fold cross-validation as a bilevel optimization problem with LASSO regression and an ℓ_1-norm support-vector machine (SVM) in the lower-level problem. They used parametric programming theory to reformulate the bilevel optimization problem as a single level problem, which is called the bilevel and parametric optimization approach to hyperparameter optimization (HY-POP). Similarly, the authors only provided the numerical experiments to show the efficiency of HY-POP without any theoretical analysis. [16] considered bilevel optimization problems for variational image denoising models, where the upper-level problem is smooth while the lower-level problem is the ℓ_p regularizer with $p = \frac{1}{2}, 1, 2$. They proposed semismooth Newton method for solving the bilevel optimization problem including the ℓ_2-norm and the ℓ_1-norm. Especially, they only provided numerical experiments for the $\ell_{\frac{1}{2}}$-norm and leave the theoretical analysis for nonconvex $\ell_{\frac{1}{2}}$-norm to future work. Nevertheless, they showed that the $\ell_{\frac{1}{2}}$-norm has better denoising performance

than the ℓ_1-norm. Recently, [21] considered the bilevel program (4) with the function $R_1(\omega) := \|\omega\|_p^p = \sum_{i=1}^n |\omega_i|^p$ $(0 < p \leq 1)$ (i.e. the ℓ_p-regularizer) and $\epsilon = 0$ by employing a smoothing method via the twice continuously differentiable function

$$\varphi_\mu(\omega) = \sum_{i=1}^{n} \left(\omega_i^2 + \mu^2\right)^{\frac{p}{2}} \tag{6}$$

as a smooth approximation of R_1. Using such a smoothing function, problem (4) can be approximated by a smooth bilevel program, which then allows for use of several optimization techniques that normally require differentiability. The authors established the convergence analysis of their smoothing algorithm when ℓ_p-norm is used with $p \in (0, 1]$.

The following are the main theoretical contributions of our present work:

(I) First, we propose bilevel KKT conditions (BKKT conditions for short) for problem (4), which are new necessary optimality conditions for the relaxation of (4) obtained by replacing its lower-level optimization problem by the corresponding first order necessary conditions in terms of generalized subdifferentials (see Sect. 2.1), that is,

$$\begin{array}{ll} \min\limits_{\omega,\lambda} & f(\omega) \\ \text{s.t} & 0 \in \partial_\omega \left(G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega)\right) \\ & (\lambda_1, \bar{\lambda}) \in \Omega_\epsilon. \end{array} \tag{7}$$

Our proposed BKKT conditions are notably tighter than the scaled bilevel KKT conditions (SB-KKT conditions for short) discovered in [21]. As a special case, when $p = 1$ and the functions f, g and R_i $(i = 1, \ldots, r)$ are all convex functions, the proposed BKKT conditions are necessary optimality conditions for the original bilevel problem (4).

(II) Second, we consider a general framework for constructing smoothing functions for R_1 given by (2), where the associated ψ is any function that satisfies Assumption (A) and the absolute value mapping is smoothly approximated by a function generated via density functions, as inspired by the smoothing technique for plus functions by Chen and Mangasarian [7]. Based on this approach, we propose a smoothing algorithm and prove its convergence to BKKT points by utilizing only some information on the generating density function. That is, we do not rely on a specific formula of a smoothing function, and therefore our framework provides a unified theory for a class of smoothing algorithms for (4). Indeed, one novelty of this work is our unified convergence analysis that solely depends on density functions. Along with these, we only suppose weaker algorithmic assumptions and constraint qualifications, as opposed to the specific model and algorithm considered in [21]. Finally, in connection with contribution (I) described above, we obtain stronger results since we establish convergence to BKKT points, which are tighter necessary conditions than SB-KKT conditions.

The SB-KKT conditions proposed in [21] for problem (7) with $\epsilon = 0$ are more loose than our proposed BKKT conditions as mentioned in (I). Consequently, we

provide a better optimality measure for the relaxation (7) of the bilevel program (4). In fact, when $p = 1$, the SB-KKT conditions proposed in [21] are not even necessary conditions for the relaxed problem (7), but for another relaxation which has a larger feasible region (see model (12) and Proposition 3.1). Hence, our proposed BKKT conditions provide a significant improvement over the prior work.

Moreover, under an appropriate assumption on the algorithm iterates (see Remark 3), our convergence analysis significantly generalizes the existing technique of [21] that only holds for the case when $\epsilon = 0$, $\psi(t) \equiv t$, and the function φ_μ in (6) is used to smoothly approximate the ℓ_p norm in (4). In the said work, the formula of the smoothing function (6) was fully exploited to derive important inequalities that are specific to (6). The specific formula of (6) was also exhaustively utilized to obtain fundamental lemmas for establishing global subsequential convergence (see, for instance, [21, Lemma 7, Proposition 8, and the proof of Theorem 5]). Indeed, the lines of arguments used to establish the aforementioned results are only applicable to the chosen smoothing function (6). It should be noted that extension to a wider class of regularizers R_1 given by (2) with an arbitrary smooth approximation of the absolute value function is not trivial and requires more subtle arguments. To this end, the present work provides a unified analysis that derives alternative fundamental lemmas and properties. This is achieved using arguments that do not rely on the specific formula of a smoothing function but rather only on certain analytic properties of a density function generating the smoothing function. In turn, other important contributions of our work involve the flexibility of our algorithm concerning the smoothing functions used and its applicability to a considerably wider class of regularizers for the hyperparameter optimization problem. Compared to [21], our algorithm comes with convergence guarantees under less restrictive constraint qualifications and weaker algorithmic assumptions, all within the framework of the unified analysis.

From a practical point of view, the choice of smoothing functions is critical in achieving successful simulations with fast convergence. We compare the numerical performance of six smoothing functions generated via Chen and Mangasarian's method [7] to determine which function is more suitable for our smoothing approach. Our proposed algorithm involves the use of a semismooth Newton method to solve a sequence of bilevel KKT systems, thereby significantly improving upon the methodology proposed in [21]. As a result, one significant finding from our numerical experience indicates that some smoothing functions result to a faster algorithm that achieves sparse models with lower validation and test errors. Consequently, this gives insights on which smoothing function can work well with the proposed strategy.

This paper is organized as follows: In Sect. 2, we review some fundamental concepts in analysis and a brief review of the method for constructing smoothing functions of the plus function by means of density functions, which was proposed in [7]. This will serve as our basis to construct smoothing functions for $R_1(\omega)$, and our theoretical analysis will all be dependent on the density function. In Sect. 3, we recall the SB-KKT conditions introduced in [21], and then propose our BKKT conditions. In Sect. 4, we present our smoothing algorithm along with its convergence analysis. In Sect. 5, we compare the numerical performance of different smoothing functions generated from different density functions in solving (4).

Throughout this paper, we denote the vector $\omega \in \Re^n$ by $\omega = (\omega_1, \ldots, \omega_n)^T$. We let $|\omega| := (|\omega_1|, \ldots, |\omega_n|)^T$, and $|\omega|^p := (|\omega_1|^p, \ldots, |\omega_n|^p)^T$. We define $I(\omega) := \{j \in \{1, 2, \ldots, n\} \mid \omega_j = 0\}$ for any $\omega \in \Re^n$. The Hadamard product of two vectors $\omega \in \Re^n$ and $\breve{\omega} \in \Re^n$ is denoted by $\omega \odot \breve{\omega} := (\omega_1\breve{\omega}_1, \ldots, \omega_n\breve{\omega}_n)^T$. We define the sgn function as $\text{sgn}(t) = 1$ if $t > 0$, $\text{sgn}(t) = 0$ if $t = 0$, and $\text{sgn}(t) = -1$ if $t < 0$. For a differentiable function $f : \Re^n \to \Re$, we denote the gradient of f by ∇f with $\nabla f(\omega) := (\frac{\partial f(\omega)}{\partial \omega_1}, \ldots, \frac{\partial f(\omega)}{\partial \omega_n})^T \in \Re^n$. If f is twice differentiable, we denote the Hessian of f by $\nabla^2 f$ with $\nabla^2 f(\omega) := \left(\frac{\partial^2 f(\omega)}{\partial \omega_i \partial \omega_j}\right)_{1 \le i,j \le n} \in \Re^{n \times n}$.

2 Preliminaries

We review some important concepts in nonsmooth analysis. We also recall the method of Chen and Mangasarian to construct smoothing functions for the plus function, and discuss how to use this to obtain a smoothing function for the absolute value function.

2.1 Some concepts in analysis

The following facts can be found in Rockafellar and Wets [22].

Definition 2.1 [22, Definition 8.3] Let $f : \Re^n \to \Re \cup \{\infty\}$ be a proper function. For vectors $v \in \Re^n$ and $\bar{x} \in \Re^n$, one say that

1. v is a *regular subgradient* of f at $\bar{x}$, written $v \in \hat{\partial} f(\bar{x})$, if

$$f(x) \ge f(\bar{x}) + v^T(x - \bar{x}) + o(\|x - \bar{x}\|).$$

2. v is a *general subgradient* of f at $\bar{x}$, written $v \in \partial f(\bar{x})$, if there are sequences $\{x^\nu\} \subseteq \Re^n$ and $\{v^\nu\} \subseteq \Re^n$ such that

$$\lim_{\nu \to \infty} x^\nu = \bar{x} \text{ and } v^\nu \in \hat{\partial} f(x^\nu) \text{ with } \lim_{\nu \to \infty} v^\nu = v.$$

Note that a regular subgradient of f at $\bar{x}$ is also called a Fréchet subgradient of f at $\bar{x}$ (see in [15]). Moreover, if f is a proper and convex function, the regular subgradient of f coincides with the subgradient of f in the sense of convex analysis (see in [22, Proposition 8.12]).

Proposition 2.1 [22, Theorem 8.6] *For a function $f : \Re^n \to \Re \cup \{\infty\}$ and a point $\bar{x}$ where f is finite, the subgradient sets $\hat{\partial} f(\bar{x})$ and $\partial f(\bar{x})$ are closed, with $\hat{\partial} f(\bar{x})$ convex and $\hat{\partial} f(\bar{x}) \subset \partial f(\bar{x})$.*

Proposition 2.2 [22, Theorem 10.1] *If a proper function $f : \Re^n \to \Re \cup \{\infty\}$ has a local minimum at $\bar{x}$, then $0 \in \hat{\partial} f(\bar{x}) \subset \partial f(\bar{x})$.*

2.2 Smoothing functions of $|x|$ via density functions

We recall the general definition of a smoothing function.

Definition 2.2 [6, Definition 1] Let $h : \Re^n \to \Re$ be a continuous function. We say that $\phi : \Re_{++} \times \Re^n \to \Re$ is a *smoothing function* of h if it satisfies the following:

(i) $\phi(\mu, \cdot)$ is continuously differentiable for any $\mu > 0$;
(ii) $\lim_{w \to z, \mu \downarrow 0} \phi(\mu, w) = h(z)$ for any $z \in \Re^n$.

To construct a smoothing function for the absolute value function, we briefly recall from [6, 7] that the plus function $(x)_+ = \max\{x, 0\}$ for $x \in \Re$ can be smoothly approximated by

$$\hat{\phi}(\mu, x) = \int_{-\infty}^{+\infty} (x - t)_+ \hat{t}(\mu, t) dt = \int_{-\infty}^{x} (x - t) \hat{t}(\mu, t) dt, \tag{8}$$

where $\hat{t}(\mu, t) := \frac{1}{\mu} \rho\left(\frac{t}{\mu}\right)$, and $\rho : \Re \to \Re_+$ is a piecewise continuous density function[1] that satisfies

$$\rho(x) = \rho(-x) \text{ and } \kappa := \int_{-\infty}^{+\infty} |x| \rho(x) dx < +\infty. \tag{9}$$

Using the fact that $|x| = (x)_+ + (-x)_+$, we obtain a smoothing function for the absolute value function as follows:

$$\phi(\mu, x) := \hat{\phi}(\mu, x) + \hat{\phi}(\mu, -x) = \int_{-\infty}^{+\infty} |x - t| \hat{t}(\mu, t) dt. \tag{10}$$

Analogous to [7, Proposition 2.2], we have the following properties of $\phi(\mu, x)$.

Proposition 2.3 *Suppose that $\phi(\mu, x)$ is defined as in (10). Then, for a fixed $\mu > 0$, we have*

(a) $\phi(\mu, \cdot)$ is continuously differentiable.
(b) $0 \le \phi(\mu, x) - |x| \le \kappa \mu$ for all $x \in \Re$ and $\mu > 0$, where the constant $\kappa > 0$ is defined in (9).
(c) $\phi'(\mu, x)$ is bounded satisfying $-1 \le \phi'(\mu, x) \le 1$ for all $x \in \Re$, $\mu > 0$.

From Proposition 2.3, given any sequence $\{(x^k, \mu_k)\} \subset \Re \times \Re_{++}$ such that $x^k \to x \in \Re$ and $\mu_k \downarrow 0$, we have

$$\lim_{k \to \infty} \phi(\mu_k, x^k) = |x| \quad \forall x \in \Re,$$

and

$$\lim_{k \to \infty} \phi'(\mu_k, x^k) = \text{sgn}(x) \quad \forall x \neq 0. \tag{11}$$

[1] That is, ρ is a nonnegative function whose integral over $\Re$ is 1. Consequently, it is easy to see that $\hat{t}(\mu, x) \to \delta(x)$ as $\mu \to 0$ for all $x \in \Re$, where δ is the Dirac delta function provided that $\rho(0) > 0$.

We also have from Proposition 2.3(c) that subsequential limits of the sequence of $\{\phi'(\mu_k, x^k)\}$ exist and belong to $[-1, 1]$.

3 Necessary conditions

Using Proposition 2.2, the first-order optimality condition for the lower-level problem (3) is given by

$$0 \in \partial_\omega(G(\omega^*, \bar{\lambda}) + \lambda_1 R_1(\omega^*)),$$

where $\partial_\omega(G(\omega^*, \bar{\lambda}) + \lambda_1 R_1(\omega^*))$ is the general subgradient with respect to ω of $G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega)$ at ω^*. Then problem (4) can be transformed into the one-level problem given in (7).

3.1 Scaled bilevel KKT conditions

In [21], a smooth version of (7) with $R_1(\omega) = \|\omega\|_p^p$ was presented by replacing its lower-level problem by the scaled first-order necessary condition, which was originally introduced by Chen, Xu and Ye [8] for non-Lipschitz continuous functions. Since the function $G(\omega, \bar{\lambda}) + \lambda_1\|\omega\|_p^p$ may be non-Lipschitz, the scaled first-order necessary condition for the lower-level problem (3) proposed in [21], as adapted from [8], can be extended to our setting with R_1 given by (2), as in Definition 3.1. In particular, when $\psi(t) \equiv t$, the following definition reduces to the scaled first-order necessary condition given in [21].

Definition 3.1 We say that ω^* satisfies the *scaled first-order necessary condition* of (3) if

$$W_* \nabla_\omega G(\omega^*, \bar{\lambda}) + p\lambda_1 |W_*|^p \psi'(|\omega^*|^p) = 0,$$

where $W_* := \mathrm{diag}(\omega^*)$, $|W_*|^p := \mathrm{diag}(|\omega^*|^p)$, and

$$\psi'(|\omega^*|^p) := (\psi'(|\omega_1^*|^p), \psi'(|\omega_2^*|^p), \ldots, \psi'(|\omega_n^*|^p))^T.$$

Using this scaled first-order necessary condition, one can obtain the following one-level problem:

$$\begin{array}{ll} \min\limits_{\omega,\lambda} & f(\omega) \\ \text{s.t} & W\nabla_\omega G(\omega, \bar{\lambda}) + p\lambda_1 |W|^p \psi'(|\omega|^p) = 0 \\ & \lambda \in \Omega_\epsilon, \end{array} \tag{12}$$

where $W := \mathrm{diag}(\omega)$, $|W|^p := \mathrm{diag}(|\omega|^p)$, and

$$\psi'(|\omega|^p) := (\psi'(|\omega_1|^p), \psi'(|\omega_2|^p), \ldots, \psi'(|\omega_n|^p))^T.$$

Though this problem looks different from (7), the following proposition[2] indicates that the problems are indeed identical when $p \in (0, 1)$. However, when $p = 1$, the feasible region of problem (7) is contained in that of (12).

[2] The proof is analogous to [21, Lemma 3].

Proposition 3.1 *For $\omega \in \Re^n$ and $\lambda \in \Re^r_+$, if $0 \in \partial_\omega(G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega))$, then $W\nabla_\omega G(\omega, \bar{\lambda}) + p\lambda_1 |W|^p \psi'(|\omega|^p) = 0$. In particular, when $p < 1$, the converse is also true.*

Based on this scaling, one can extend the scaled bilevel KKT (SB-KKT) conditions proposed in [21] to our general setting of (4) as in the following definition.

Definition 3.2 We say that $(\omega^*, \lambda^*) \in \Re^n \times \Re^r$ is a scaled bilevel Karush-Kuhn-Tucker (SB-KKT) point for problem (4) if there exists a pair of vectors $(\zeta^*, \eta^*) \in \Re^n \times \Re^r$ such that

$$W_*^2 \nabla f(\omega^*) + H(\omega^*, \lambda^*)\zeta^* = 0, \tag{13}$$

$$W_* \nabla_\omega G(\omega^*, \bar{\lambda}^*) + p\lambda_1^* |W_*|^p \psi'(|\omega^*|^p) = 0, \tag{14}$$

$$p \sum_{j \notin I(\omega^*)} \operatorname{sgn}(\omega_j^*) |\omega_j^*|^{p-1} \psi'(|\omega_j^*|^p) \zeta_j^* = \eta_1^*, \tag{15}$$

$$\zeta_j^* = 0 \ (j \in I(\omega^*)), \tag{16}$$

$$\nabla R_j(\omega^*)^T \zeta^* - \eta_j^* = 0 \ (j = 2, 3, \ldots, r), \tag{17}$$

$$\lambda^* - \epsilon e_1 \geq 0, \quad \eta^* \geq 0, \quad (\lambda^* - \epsilon e_1)^T \eta^* = 0, \tag{18}$$

where $W_* := \operatorname{diag}(\omega^*)$, and $e_1 = (1, 0, \ldots, 0)^T \in \Re^r$. Here, we write

$$\begin{aligned} H(\omega, \lambda) = {} & W^2 \nabla^2_{\omega\omega} G(\omega, \bar{\lambda}) + \lambda_1 p(p-1) \operatorname{diag}(|W|^p \psi'(|\omega|^p)) \\ & + \lambda_1 p^2 \operatorname{diag}(|W|^{2p} \psi''(|\omega|^p)) \end{aligned}$$

with $W := \operatorname{diag}(\omega)$, $|W|^p := \operatorname{diag}(|\omega|^p)$, and $|W|^{2p} := \operatorname{diag}(|\omega|^{2p})$ for $\omega \in \Re^n$ and $\lambda \in \Re^r$.

The SB-KKT conditions are necessary optimality conditions for problem (12) as asserted in the following result, whose proof is essentially similar to [21, Theorem 2].

Proposition 3.2 *Let $(\omega^*, \lambda^*) \in \Re^n \times \Re^r$ be a local optimum of (12). Then, (ω^*, λ^*) together with some pair of vectors $(\zeta^*, \eta^*) \in \Re^n \times \Re^r$ satisfies the SB-KKT conditions (13)–(18) under an appropriate constraint qualification concerning the constraints $\frac{\partial G(\omega, \bar{\lambda})}{\partial \omega_j} + p \operatorname{sgn}(\omega_j) \lambda_1 |\omega_j|^{p-1} \psi'(|\omega_j|^p) = 0 \ (j \notin I(\omega^*))$, $\omega_j = 0 \ (j \in I(\omega^*))$, and $\lambda \in \Omega_\epsilon$.*

3.2 Bilevel KKT conditions

An immediate consequence of Propositions 3.1 and 3.2 is that when $p \in (0, 1)$, a local optimum of the one-level problem (7) satisfies the SB-KKT conditions under appropriate constraint qualifications. However, one main drawback of the SB-KKT conditions presented in the preceding section is that the process of "scaling" enlarges the feasible region of the relaxed one-level problem (7) when $p = 1$. In the following definition, we propose an alternative necessary condition which avoids the multiplication by W and W^2 as defined in Definition 3.2.

Definition 3.3 We say that $(\omega^*, \lambda^*) \in \Re^n \times \Re^r$ is a *bilevel Karush-Kuhn-Tucker point (BKKT point)* for problem (4) if there exists $(\zeta^*, \eta^*) \in \Re^n \times \Re^r$ such that

$$\nabla_{\tilde{\omega}} f(\omega^*) + \widetilde{H}(\omega^*, \lambda^*)\tilde{\zeta}^* = 0, \tag{19}$$
$$\nabla_{\tilde{\omega}} G(\omega^*, \bar{\lambda}^*) + p\lambda_1^* \psi'(|\tilde{\omega}^*|^p) \odot |\tilde{\omega}^*|^{p-1} \odot \operatorname{sgn}(\tilde{\omega}^*) = 0, \tag{20}$$
$$\breve{\zeta}^* = 0\,, \tag{21}$$
$$p\left(\psi'(|\tilde{\omega}^*|^p) \odot |\tilde{\omega}^*|^{p-1} \odot \operatorname{sgn}(\tilde{\omega}^*)\right)^T \tilde{\zeta}^* = \eta_1^*, \tag{22}$$
$$\nabla_{\tilde{\omega}} \bar{R}(\omega^*)^T \tilde{\zeta}^* - \bar{\eta}^* = 0, \tag{23}$$
$$\lambda^* - \epsilon e_1 \geq 0, \quad \eta^* \geq 0, \quad (\lambda^* - \epsilon e_1)^T \eta^* = 0, \tag{24}$$

where $\bar{\eta}^* = (\eta_2^*, \ldots, \eta_r^*)$,

$$\widetilde{H}(\omega, \lambda) := \nabla^2_{\tilde{\omega}\tilde{\omega}} G(\omega, \bar{\lambda}) + \lambda_1 p(p-1)|\tilde{W}|^{p-2}\psi'(|\tilde{\omega}|^p) + p^2\lambda_1 |\tilde{W}|^{2p-2}\psi''(|\tilde{\omega}|^p),$$
$$|\tilde{W}| := \operatorname{diag}(|\tilde{\omega}|),\ \tilde{\omega}^* := (\omega_i^*)_{i \notin I(\omega^*)},$$
$$\breve{\omega}^* := (\omega_i^*)_{i \in I(\omega^*)},\ \tilde{\zeta}^* := (\zeta_i^*)_{i \notin I(\omega^*)},\ \breve{\zeta}^* := (\zeta_i^*)_{i \in I(\omega^*)}.$$

We show in the following propositions that the proposed bilevel KKT conditions in Definition 3.3 are necessary conditions for the one-level relaxation (7) of the original bilevel problem (4). In other words, the scaling used in the preceding section, as extended from the prior work [21], is not needed.

Proposition 3.3 *Let $p \in (0, 1)$ and $(\omega^*, \lambda^*) \in \Re^n \times \Re^r$ be a local optimum of (7). Then, (ω^*, λ^*) is a BKKT point under an appropriate constraint qualification concerning the constraints $\frac{\partial G(\omega, \bar{\lambda})}{\partial \omega_j} + p\operatorname{sgn}(\omega_j)\lambda_1 |\omega_j|^{p-1}\psi'(|\omega_j|^p) = 0$ $(j \notin I(\omega^*))$, $\omega_j = 0$ $(j \in I(\omega^*))$, and $\lambda \in \Omega_\epsilon$.*

Proof Let $(\omega^*, \lambda^*) \in \Re^n \times \Re^r$ be a local minimum of (7), and consider the problem

$$\begin{aligned} \min_{\omega, \lambda} \quad & f(\omega) \\ \text{s.t} \quad & \nabla_{\tilde{\omega}} G(\omega, \bar{\lambda}) + p\lambda_1 \psi'(|\tilde{\omega}|^p) \odot |\tilde{\omega}|^{p-1} \odot \operatorname{sgn}(\tilde{\omega}) = 0, \\ & \breve{\omega} = 0, \\ & \lambda \in \Omega_\epsilon, \end{aligned} \tag{25}$$

where $\tilde{\omega} = (\omega_i)_{i \notin I(\omega^*)}$ and $\breve{\omega} = (\omega_i)_{i \in I(\omega^*)}$. We claim that (ω^*, λ^*) is a local minimum of (25). It is clear that (ω^*, λ^*) is feasible to (25). Let (ω, λ) be a feasible point of (25) in some sufficiently small neighborhood of (ω^*, λ^*) so that $\tilde{\omega} \neq 0$. Then with the first equality constraint in (25) together with the fact that $\partial_{\breve{\omega}}\left(G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega)\right) = \Re^{|I(\omega^*)|}$, it immediately follows that $0 \in \partial_\omega(G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega))$, and therefore (ω, λ) belongs to the feasible region of (7). Using the fact that (ω^*, λ^*) is a local minimum of (7), we indeed obtain that (ω^*, λ^*) is a local minimum of (25). Hence, there exist

Lagrange multipliers $(\hat{\zeta}^{(1)}, \hat{\zeta}^{(2)}, \hat{\eta}) \in \Re^{n-|I(\omega^*)|} \times \Re^{|I(\omega^*)|} \times \Re^r$ such that

$$\begin{bmatrix} \nabla_{\tilde{\omega}} f(\omega^*) \\ \nabla_{\breve{\omega}} f(\omega^*) \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} \widetilde{H}(\omega^*, \bar{\lambda}^*) & 0 \\ \nabla^2_{\breve{\omega}\tilde{\omega}} G(\omega^*, \bar{\lambda}^*) & I \\ p\psi'(|\tilde{\omega}^*|^p) \odot |\tilde{\omega}^*|^{p-1} \odot \operatorname{sgn}(\tilde{\omega}^*) & 0 \\ \nabla_{\tilde{\omega}} \bar{R}(\omega^*)^T & 0 \end{bmatrix} \begin{bmatrix} \hat{\zeta}^{(1)} \\ \hat{\zeta}^{(2)} \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ \hat{\eta}_1 \\ \bar{\hat{\eta}} \end{bmatrix} = 0,$$

$$\nabla_{\tilde{\omega}} G(\omega^*, \bar{\lambda})^* + p\lambda_1^* \psi'(|\tilde{\omega}^*|^p) \odot |\tilde{\omega}^*|^{p-1} \odot \operatorname{sgn}(\tilde{\omega}^*) = 0,$$

$$\lambda^* - \epsilon e_1 \geq 0, \quad \eta^* \geq 0, \quad (\lambda^* - \epsilon e_1)^T \eta^* = 0,$$

where $\bar{\hat{\eta}} = (\hat{\eta}_2, \hat{\eta}_3, \ldots, \hat{\eta}_r)^T$. Taking $\eta^* = \hat{\eta}$ and setting $\zeta^* \in \Re^n$ such that $\zeta_i^* = \hat{\zeta}_i^{(1)}$ for $i \notin I(\omega^*)$ and $\zeta_i^* = 0$ for $i \in I(\omega^*)$, one can easily check that all the conditions (19)–(24) are satisfied. □

For the case $p = 1$, we obtain a similar result provided that the feasibility condition

$$\|\nabla_{\breve{\omega}} G(\omega^*, \bar{\lambda}^*)\|_\infty < \lambda_1^* \psi'(0) \tag{26}$$

holds, where $\breve{\omega} = (\omega_i)_{i \in I(\omega^*)}$. We note that this is equivalent to saying that $-\frac{1}{\lambda_1^*} \frac{\partial G}{\partial \omega_i}(\omega^*, \bar{\lambda}^*)$ belongs to the interior of the subdifferential set $\partial \psi(|t|)|_{t=0}$ for all $i \in I(\omega^*)$. This condition is used in the convergence analysis of the smoothing algorithm in [21], but its connection with the necessary conditions for solutions of (7) was not explored. This precise connection is revealed in the following proposition.

Proposition 3.4 *Let $p = 1$ and $(\omega^*, \lambda^*) \in \Re^n \times \Re^r$ be a local optimum of (7) that satisfies (26). Then, (ω^*, λ^*) is a BKKT point under an appropriate constraint qualification concerning the constraints $\frac{\partial G(\omega, \bar{\lambda})}{\partial \omega_j} + p\operatorname{sgn}(\omega_j)\lambda_1 |\omega_j|^{p-1} \psi'(|\omega_j|^p) = 0$ $(j \notin I(\omega^*))$, $\omega_j = 0$ $(j \in I(\omega^*))$, and $\lambda \in \Omega_\epsilon$.*

Proof We consider a problem similar to (25) but with (26) as an added inequality constraint, that is,

$$\begin{aligned} \min_{\omega, \lambda} \quad & f(\omega) \\ \text{s.t} \quad & \nabla_{\tilde{\omega}} G(\omega, \bar{\lambda}) + \lambda_1 \psi'(|\tilde{\omega}|) \odot \operatorname{sgn}(\tilde{\omega}) = 0, \\ & \breve{\omega} = 0, \\ & \lambda \in \Omega_\epsilon \\ & \|\nabla_{\breve{\omega}} G(\omega, \bar{\lambda})\|_\infty < \lambda_1 \psi'(0), \end{aligned} \tag{27}$$

where $\tilde{\omega} = (\omega_i)_{i \notin I(\omega^*)}$ and $\breve{\omega} = (\omega_i)_{i \in I(\omega^*)}$. Note that $\|\nabla_{\breve{\omega}} G(\omega, \bar{\lambda})\|_\infty < \lambda_1 \psi'(0)$ is a non-binding inequality constraint of (27). Hence, following the proof of Proposition 3.3, it suffices to show that (ω^*, λ^*) is feasible to (27) and that the feasible region of (27) is contained in that of (7). The former is clear due to our hypothesis. To show the inclusion of the feasible regions, let (ω, λ) be a feasible point of (27). If $i \in I(\omega^*)$, then $\omega_i = 0$, which together with (26) implies that $0 \in \partial_{\omega_i}(G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega))$. If $i \notin I(\omega^*)$ and $\omega_i \neq 0$, it is clear that $0 \in \partial_{\omega_i}(G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega))$ from the first equality constraint in (27). On the other hand, if $i \notin I(\omega^*)$ but $\omega_i = 0$, we also have from

the first equality constraint in (27) that $\partial G(w, \bar{\lambda})/\partial \omega_i = 0$. Since $0 \in \partial \psi(|t|)|_{t=0}$, we again have $0 \in \partial_{\omega_i}(G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega))$. This completes the proof. □

Remark 1 We make some comments about the case $p = 1$.

(a) For this case, we always need to assume that a candidate bilevel KKT point (ω^*, λ^*) satisfies inequality (26). This will also be the standing assumption for our subsequent analysis when dealing with the case of $p = 1$, as we shall see in the next section.

(b) Note that if g and R_j, $j = 2, \ldots, r$ are convex functions, then we obtain a stronger result that *the bilevel KKT conditions are necessary conditions for the original bilevel problem* (4) under appropriate constraint qualifications, rather than just necessary conditions for the relaxed problem (7), which is the situation when $p \in (0, 1)$ (even if the functions g and R_j are all convex). Hence, in this case, bilevel KKT points are indeed candidate solutions to the bilevel problem (4).

4 Proposed algorithm and its convergence

In this section, we describe our smoothing algorithm for (4) and present our convergence results.

4.1 Smoothing approach and the algorithm

One main source of difficulty in solving the bilevel program (4) is the nonsmooth, nonconvex and possibly non-Lipschitz component $R_1(\omega) = \sum_{i=1}^{n} \psi(|\omega_i|^p)$, where $p \in (0, 1]$. To overcome this, we apply the smoothing technique to R_1 with the smoothing function ϕ defined in the previous section, yielding the following smooth approximation:

$$\varphi_\mu(\omega) := \sum_{j=1}^{n} \psi\left([\phi(\mu, \omega_j)]^p\right). \tag{28}$$

Then, as in [21], we consider problem (4) with φ_μ in place of R_1, and further replace the obtained smoothed lower-level problem with its first-order condition. Hence, the following problem is obtained:

$$\begin{aligned} \min_{\omega,\lambda} \quad & f(\omega) \\ \text{s.t} \quad & \nabla_\omega G(\omega, \bar{\lambda}) + \lambda_1 \nabla \varphi_\mu(\omega) = 0 \\ & \lambda \in \Omega_\epsilon. \end{aligned} \tag{29}$$

Next, we suppose that φ_μ is twice continuously differentiable from this moment, and we also recall Assumption (A) together with our differentiability assumptions on R_i $(i = 2, \ldots, r)$ and g. These properties enable us to consider the KKT conditions.[3]

[3] Without the $\mathcal{C}^2$ property of φ_μ, the KKT conditions cannot be considered because the constraint function $\nabla_\omega G(\omega, \bar{\lambda}) + \lambda_1 \nabla \varphi_\mu(\omega)$ may not necessarily be smooth.

By virtue of this fact, we can find candidate solutions to (29) by looking at its KKT points.

In fact, it is sufficient to obtain approximate KKT points: Given a parameter $\hat{\varepsilon} > 0$, we define an $\hat{\varepsilon}$-approximate KKT point for problem (29) as follows: We say that $\{(\omega, \lambda, \zeta, \eta)\} \subseteq \Re^n \times \Re^r \times \Re^n \times \Re^r$ is an $\hat{\varepsilon}$-approximate KKT point for (29) if there exists a vector $\hat{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5) \in \Re^n \times \Re \times \Re^{r-1} \times \Re^n \times \Re$ such that

$$\nabla f(\omega) + (\nabla^2_{\omega\omega} G(\omega, \bar{\lambda}) + \lambda_1 \nabla^2 \varphi_\mu(\omega))\zeta = \varepsilon_1, \tag{30}$$

$$\nabla \varphi_\mu(\omega)^T \zeta - \eta_1 = \varepsilon_2, \tag{31}$$

$$\nabla R_j(\omega)^T \zeta - \eta_j = (\varepsilon_3)_{j-1} \ (j = 2, 3, \ldots, r), \tag{32}$$

$$\nabla_\omega G(\omega, \bar{\lambda}) + \lambda_1 \nabla \varphi_\mu(\omega) = \varepsilon_4, \tag{33}$$

$$\lambda - \epsilon e_1 \geq 0, \quad \eta \geq 0, \quad (\lambda - \epsilon e_1)^T \eta = \varepsilon_5, \tag{34}$$

and

$$\|(\varepsilon_1^T, \varepsilon_2^T, \varepsilon_3^T, \varepsilon_4^T, \varepsilon_5^T)^T\| \leq \hat{\varepsilon},$$

where $\nabla^2_{\omega\omega} G(\omega, \bar{\lambda})$ is the Hessian of G with respect to ω. Note that when $\hat{\varepsilon} = 0$, an $\hat{\varepsilon}$-approximate KKT point is identical to a KKT point.

Now, by iteratively computing an $\hat{\varepsilon}$-approximate KKT point while decreasing $\hat{\varepsilon}$ and the smoothing parameter μ, we obtain the smoothing algorithm presented in Algorithm 1.

Algorithm 1 (A smoothing method for nonsmooth bilevel optimization)

Step 0 Choose $\mu_0 > 0$, $\beta_1, \beta_2 \in (0, 1)$ and $\hat{\varepsilon}_0 \geq 0$. Set $k := 0$.
Step 1 Find an $\hat{\varepsilon}_k$-approximate KKT point $\{(\omega^{k+1}, \lambda^{k+1}, \zeta^{k+1}, \eta^{k+1})\}$ for problem (29) with $\mu = \mu_k$.
Step 2 Set $\mu_{k+1} = \beta_1 \mu_k$, $\hat{\varepsilon}_{k+1} = \beta_2 \hat{\varepsilon}_k$ and $k := k + 1$.

Algorithm 1 is quite similar to the one proposed by Okuno et al [21]. However, whereas Okuno et al supposed to employ only $\sum_{i=1}^n (\omega_i^2 + \mu^2)^{\frac{p}{2}}$ as the smoothing function φ_μ, Algorithm 1 enjoys much more freedom in the choice of smoothing functions, as well as penalty functions ψ.

In our convergence analysis, we assume that at every iteration, an $\hat{\varepsilon}_k$-approximate KKT point can always be computed. In order to establish the global convergence of Algorithm 1, we will require some more properties of the density function ρ used for constructing the smoothing function ϕ, such as properties that will guarantee that φ_μ is $\mathcal{C}^2$ as supposed above.

4.2 Convergence analysis

4.2.1 Assumptions on density function

We have mentioned in the Introduction that the setting of all our analysis is based on density functions. That is, we wish to prove all our convergence results by solely looking at density functions used to induce the smoothing functions. To this end, we must be able to identify necessary properties of a given density function so that Algorithm 1 converges to a candidate solution of the main problem (4). Indeed, one novelty of our work is precisely the identification of these required properties and its application to the convergence analysis.

We summarize necessary assumptions on the density function ρ that we will use in the next subsection.

Assumption (B). Let $\rho : \Re \to \Re_+$ be a density function. Then, the following properties hold:

(B1) ρ is symmetric, i.e. $\rho(x) = \rho(-x)$ for all $x \in \Re$.
(B2) ρ is continuous and nonincreasing on $[0, \infty)$.
(B3) There exist positive constants $c, r > 0$ such that

$$2\int_0^S \rho(x)\,dx \geq 1 - \frac{c}{S^r + c} \quad \text{for all } S \geq 0.$$

(B4) If $p = 1$, we have $\rho(x) > 0$ for all $x \in \Re$.

Some remarks are in order: First, although Assumption (B1) was already supposed in Sect. 2.2, we have restated it for later use. Under this assumption, note that the smoothing function $\phi(\mu, x)$ is strictly positive for all $\mu > 0$ and $x \in \Re$. Indeed, we already have from Proposition 2.3(b) that $\phi(\mu, x) > 0$ for $x \neq 0$. On the other hand, since ρ is a symmetric density function by Assumption (B1), then ρ is not identical to the zero function on the interval $[0, \infty)$. Consequently, $\int_0^{+\infty} t\rho(t)dt > 0$, which together with (8) and (10) yields $\phi(\mu, 0) > 0$. Moreover, we can easily calculate the first and second derivatives of the induced $\phi(\mu, x)$ as

$$\phi'(\mu, x) = 2\mathrm{sgn}(x)\int_0^{|x|} \frac{1}{\mu}\rho\left(\frac{t}{\mu}\right) dt = 2\mathrm{sgn}(x)\int_0^{\frac{|x|}{\mu}} \rho(t)\,dt, \tag{35}$$

and

$$\phi''(\mu, x) = \frac{2}{\mu}\rho\left(\frac{x}{\mu}\right), \tag{36}$$

respectively. From Eq. (36) and strict positivity of $\phi(\mu, x)$, we see that $\phi(\mu, \cdot)$ is twice continuously differentiable approximation of $|x|$ by the continuity assumption in (B2).

Assumptions (B1) and (B2) will also have other important roles in the proofs of our main result. Among them are formulas for the limits of $\{(\nabla\varphi_{\mu_{k-1}}(\omega^k))_j\}$ and $\{(\nabla^2\varphi_{\mu_{k-1}}(\omega^k))_{jj}\}$ when $j \notin I(\omega^*)$, where φ_μ is the smooth function given by (28).

First, with simple calculations, the components of $\nabla\varphi_\mu(\omega) \in \Re^n$ are given by

$$(\nabla\varphi_\mu(\omega))_j = p\psi'\left([\phi(\mu,\omega_j)]^p\right)\phi'(\mu,\omega_j)[\phi(\mu,\omega_j)]^{p-1}, \tag{37}$$

while $\nabla^2\varphi_\mu(\omega) \in \Re^{n\times n}$ is a diagonal matrix whose diagonal entries are given by

$$\begin{aligned}(\nabla^2\varphi_\mu(\omega))_{jj} = {} & p^2\psi''\left([\phi(\mu,\omega_j)]^p\right)[\phi'(\mu,\omega_j)[\phi(\mu,\omega_j)]^{p-1}]^2 \\ & + p(p-1)\psi'\left([\phi(\mu,\omega_j)]^p\right)[\phi'(\mu,\omega_j)]^2[\phi(\mu,\omega_j)]^{p-2} \\ & + p\psi'\left([\phi(\mu,\omega_j)]^p\right)[\phi(\mu,\omega_j)]^{p-1}\phi''(\mu,\omega_j),\end{aligned} \tag{38}$$

for $j = 1,\ldots,n$. Since $\phi(\mu,x)$ is strictly positive under Assumption (B1), the factor $[\phi(\mu,\omega_j)]^{p-1}$ that appears in (37) and (38) is real-valued for any $p \in (0,1]$. In addition, it is clear from (38) that the components of the Hessian of φ_μ are continuous by using Assumption (A), Eq. (36), and Assumptions (B1)-(B2).

We now list some important formulas that will be useful in our subsequent analysis.

Lemma 4.1 *Suppose that Assumptions (B1)-(B2) hold. Let* $\{(\omega^k,\mu_{k-1})\} \subseteq \Re^n \times \Re_{++}$ *be an arbitrary sequence converging to* $(\omega^*,0)$. *Then for* $j \notin I(\omega^*)$, *we have*

$$\lim_{k\to\infty}(\nabla\varphi_{\mu_{k-1}}(\omega^k))_j = p\,\mathrm{sgn}(\omega_j^*)|\omega_j^*|^{p-1}\psi'\left(|\omega_j^*|^p\right), \tag{39}$$

$$\lim_{k\to\infty}(\nabla^2\varphi_{\mu_{k-1}}(\omega^k))_{jj} = p^2\psi''\left(|\omega_j^*|^p\right)|\omega_j^*|^{2p-2} + p(p-1)\psi'\left(|\omega_j^*|^p\right)|\omega_j^*|^{p-2}. \tag{40}$$

Proof Let $j \notin I(\omega^*)$. We have from (11) that

$$\lim_{k\to\infty}\phi'(\mu_{k-1},\omega_j^k) = \mathrm{sgn}(\omega_j^*).$$

Since $\phi(\mu,x)$ is a smoothing function of $|x|$ that is strictly positive by Assumption (B1), we have $[\phi(\mu_{k-1},\omega_j^k)]^{p-1} \to |\omega_j^*|^{p-1}$ as $k\to\infty$. Moreover, since ψ is $\mathcal{C}^2$ by Assumption (A), then we easily obtain (39) by letting $k\to\infty$ in Eq. (37) with $\omega = \omega^k$ and $\mu = \mu_{k-1}$.

For Eq. (40), we first show that $\lim_{k\to\infty}\phi''(\mu_{k-1},\omega_j^k) = 0$, which by (36) is equivalent to showing that

$$\lim_{k\to\infty}\frac{1}{\mu_{k-1}}\rho\left(\frac{\omega_j^k}{\mu_{k-1}}\right) = \lim_{k\to\infty}\frac{1}{\mu_{k-1}}\rho\left(\frac{|\omega_j^k|}{\mu_{k-1}}\right) = 0. \tag{41}$$

Since $j \notin I(\omega^*)$, then $|\omega_j^*|/2 < |\omega_j^k|$ for all k sufficiently large. Thus,

$$0 \le \frac{1}{\mu_{k-1}}\rho\left(\frac{\omega_j^k}{\mu_{k-1}}\right) \le \frac{1}{\mu_{k-1}}\rho\left(\frac{\omega_j^*}{2\mu_{k-1}}\right) \to \delta(\omega_j^*/2) = 0 \qquad \text{as } k\to\infty,$$

by invoking Assumptions (B1)-(B2) and the definition of the Dirac delta function. This proves (41). Finally, taking the limit in (38) when $k \to \infty$, we obtain

$$\begin{aligned}
&\lim_{k\to\infty} (\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj} \\
&\quad = p^2\psi''\left(|\omega_j^*|^p\right)[\operatorname{sgn}(\omega_j^*)|\omega_j^*|^{p-1}]^2 + p(p-1)\psi'\left(|\omega_j^*|^p\right)[\operatorname{sgn}(\omega_j^*)]^2|\omega_j^*|^{p-2} \\
&\quad = p^2\psi''\left(|\omega_j^*|^p\right)|\omega_j^*|^{2p-2} + p(p-1)\psi'\left(|\omega_j^*|^p\right)|\omega_j^*|^{p-2}.
\end{aligned}$$

This completes the proof of the lemma. □

On the other hand, the other technical assumption on ρ, namely (B3), is important in our subsequent analysis. Without knowing definitively the formula for $\phi(\mu, x)$, the analysis becomes extremely difficult. In particular, if $\omega^k = (\omega_1^k, \ldots, \omega_n^k)$ and μ_{k-1} are generated from Algorithm 1, it is challenging to understand the behavior of the sequence $\left\{S_j^k\right\}$, where

$$S_j^k := \frac{|\omega_j^k|}{\mu_{k-1}}, \tag{42}$$

when $\omega_j^k \to 0$ as $k \to \infty$ for some $j \in \{1, \ldots, n\}$. Nonetheless, this problem can be addressed thanks to the simple assumption (B3). Finally, Assumption (B4) will later be important in proving the unboundedness of the sequence $\left\{|\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj}|\right\}$ when $p = 1$ and $\omega_j^k \to 0$ as $k \to \infty$. Interestingly, we shall see shortly that (B4) is not needed for the case $p \in (0, 1)$.

In the forthcoming discussions, we will see in great detail how these assumptions on ρ will play a central role in establishing the main convergence result. In Appendix B, we provide some specific examples of density functions satisfying Assumption (B).

4.2.2 Subsequential convergence

We now prove our main result that accumulation points of the sequence generated by Algorithm 1 are in fact bilevel KKT points (see Definition 3.3), which in turn are candidate solutions for (7) when $p < 1$, and candidate solutions for the original bilevel problem (4). As mentioned in Remark 1, we will assume that for any given accumulation point of such sequence generated by Algorithm 1, the inequality (26) holds when $p = 1$.

Theorem 4.1 *Let $p \in (0, 1]$ and assume that $(\omega^*, \lambda^*, \zeta^*, \eta^*)$ is an accumulation point of a sequence $\{(\omega^k, \lambda^k, \zeta^k, \eta^k)\}$ generated by Algorithm 1. Then (ω^*, λ^*) is a bilevel KKT point for the original problem (4) provided that Assumptions (A) and (B) hold.*

To prove Theorem 4.1, we show that $(\omega^*, \lambda^*, \zeta^*, \eta^*)$ satisfies Eqs. (19)–(24).

To this end, we prove a series of lemmas, and in particular, we do the following:

(i) Prove that $\{S_j^k\}_{k\in K}$ is bounded, where S_j^k is given by (42), $j \in I(\omega^*)$, $K \subset \{1, 2, \ldots, \}$ such that $(\omega^k, \lambda^k) \to (\omega^*, \lambda^*)$ as $k \in K \to \infty$ and $\{(\omega^k, \lambda^k)\}$ is generated by Algorithm 1;

(ii) Using the boundedness of $\{S_j^k\}_{k\in K}$, we compute the limit of the sequence $\left\{(\nabla^2\varphi_{\mu_{k-1}}(\omega^k))_{jj}\right\}_{k\in K}$, where $j \in I(\omega^*)$ and the index set K is as described in (i);
(iii) Using (ii) and Lemma 4.1, we show that $\breve{\zeta}^* = 0$ and equation (22) holds.

The above objectives are formally stated and proved, respectively, in Lemma 4.2 to Lemma 4.4. We will prove these results without knowledge of the exact formula for the smoothing function $\phi(\mu, x)$ used to construct φ_μ given by (28), that is, only using Assumption (B) on the density function.

To facilitate our subsequent analysis, we note here that a sequence $\{(\omega^k, \lambda^k, \zeta^k, \eta^k)\}$ generated by Algorithm 1 satisfies

$$\nabla f(\omega^k) + (\nabla^2_{\omega\omega} G(\omega^k, \bar{\lambda}^k) + \lambda_1^k \nabla^2 \varphi_{\mu_{k-1}}(\omega^k))\zeta^k = \varepsilon_1^{k-1}, \tag{43}$$

$$\nabla \varphi_{\mu_{k-1}}(\omega^k)^T \zeta^k - \eta_1^k = \varepsilon_2^{k-1}, \tag{44}$$

$$\nabla R_j(\omega^k)^T \zeta^k - \eta_j^k = (\varepsilon_3^{k-1})_{j-1} \ (j = 2, 3, \ldots, r), \tag{45}$$

$$\nabla_\omega G(\omega^k, \bar{\lambda}^k) + \lambda_1^k \nabla \varphi_{\mu_{k-1}}(\omega^k) = \varepsilon_4^{k-1}, \tag{46}$$

$$\lambda^k - \epsilon e_1 \geq 0, \quad \eta^k \geq 0, \quad (\lambda^k - \epsilon e_1)^T \eta^k = \varepsilon_5^{k-1}, \tag{47}$$

for all k.

In [21], the authors proved that when $\psi(t) = t$ and $\phi(\mu, x) = \sqrt{x^2 + \mu^2}$ and under Assumptions (B1)-(B3), there exists some $\gamma > 0$ such that

$$\mu_{k-1}^2 \geq \gamma |\omega_j^k|^{\frac{2}{2-p}} \quad (j \in I(\omega^*)) \tag{48}$$

for all sufficiently large $k \in K$, where K is the subsequence described in (i). This result is especially important in proving results related to (ii)-(iii) above. However, in order to derive inequality (48), [21] takes advantage of the specific function $\phi(\mu, x)$ chosen, which is not the case in the present work. Nevertheless, we have found out that such a strong result is not necessarily required to prove (ii)-(iii). In particular, it suffices to establish (i), which is indeed a weaker property. To this end, Assumption (B3) will play a very significant role without which the analysis becomes extremely difficult.

We now prove our first lemma which establishes property (i).

Lemma 4.2 *Suppose that Assumptions (B1)-(B3) hold. Let (ω^*, λ^*) be an arbitrary accumulation point of the sequence $\{(\omega^k, \lambda^k)\}$ generated by Algorithm 1, and let $\{(\omega^k, \lambda^k)\}_{k\in K}$ be an arbitrary subsequence converging to (ω^*, λ^*). For any $j \in I(\omega^*)$, $\{S_j^k\}_{k\in K}$ is bounded. Moreover, $S_j^k \to 0$ as $k \in K \to \infty$ if $p \in (0, 1)$.*

Proof Denote

$$F_j(\omega^k, \bar{\lambda}^k) := \frac{\partial G(\omega^k, \bar{\lambda}^k)}{\partial \omega_j}. \tag{49}$$

From Eqs. (46) and (37), we have

$$F_j(\omega^k, \bar{\lambda}^k) + p\lambda_1^k \psi'\left([\phi(\mu_{k-1}, \omega_j^k)]^p\right) \phi'(\mu_{k-1}, \omega_j^k)[\phi(\mu_{k-1}, \omega_j^k)]^{p-1} = (\varepsilon_4^{k-1})_j. \tag{50}$$

Case 1. Suppose that $p = 1$. Then

$$F_j(\omega^k, \bar{\lambda}^k) + \lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k))\phi'(\mu_{k-1}, \omega_j^k) = (\varepsilon_4^{k-1})_j. \tag{51}$$

Rearranging the terms and using Assumptions (A) and (B3), there are constants $c, r > 0$ such that for sufficiently large k,

$$\frac{|(\varepsilon_4^{k-1})_j - F_j(\omega^k, \bar{\lambda}^k)|}{\lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k))} = |\phi'(\mu_{k-1}, \omega_j^k)| = 2\int_0^{S_j^k} \rho(s)\, ds \geq 1 - \frac{c}{(S_j^k)^r + c},$$

where the second equality holds by (35), and $\lambda_1^k > 0$ for sufficiently large k since $\lambda_1^* \geq \epsilon > 0$. Consequently, we get

$$\begin{aligned} \frac{c}{(S_j^k)^r + c} &\geq 1 - \frac{|(\varepsilon_4^{k-1})_j - F_j(\omega^k, \bar{\lambda}^k)|}{\lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k))} \\ &= \frac{\lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k)) - |(\varepsilon_4^{k-1})_j - F_j(\omega^k, \bar{\lambda}^k)|}{\lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k))}. \end{aligned}$$

Note that $\lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k)) - |(\varepsilon_4^{k-1})_j - F_j(\omega^k, \bar{\lambda}^k)| > 0$ for all large k by using the fact that (26) holds at (ω^*, λ^*) when $p = 1$ and by invoking Proposition 2.3(c) and (51). Then

$$\begin{aligned} 0 \leq \frac{(S_j^k)^r}{c} \leq \frac{(S_j^k)^r + c}{c} &\leq \frac{\lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k))}{\lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k)) - |(\varepsilon_4^{k-1})_j - F_j(\omega^k, \bar{\lambda}^k)|} \\ &\to \frac{\lambda_1^* \psi'(0)}{\lambda_1^* \psi'(0) - |F_j(\omega^*, \bar{\lambda}^*)|} \text{ as } k \in K \to \infty, \end{aligned}$$

where the finiteness of the limit is guaranteed by (26). Hence, it easily follows that $\{S_j^k\}_{k \in K}$ is bounded.

Case 2. Now, suppose $p \in (0, 1)$. From Eq. (50), we have

$$\frac{|(\varepsilon_4^{k-1})_j - F_j(\omega^k, \bar{\lambda}^k)|}{p\lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k))} = |\phi'(\mu_{k-1}, \omega_j^k)| \cdot [\phi(\mu_{k-1}, \omega_j^k)]^{p-1}.$$

Using Assumptions (B1) and (B3) and by (35), we get

$$[\phi(\mu_{k-1}, \omega_j^k)]^{1-p} \cdot \frac{|(\varepsilon_4^{k-1})_j - F_j(\omega^k, \bar{\lambda}^k)|}{p\lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k))} = |\phi'(\mu_{k-1}, \omega_j^k)| \geq 1 - \frac{c}{(S_j^k)^r + c} \geq 0.$$

Meanwhile, note that since $1 - p > 0$ and $\omega_j^k \to 0$ as $k \in K \to \infty$, we have

$$\lim_{k \in K \to \infty} [\phi(\mu_{k-1}, \omega_j^k)]^{1-p} = 0.$$

Since $\lambda_1^* > 0$
and $\psi'(0) > 0$ by Assumption (A), then

$$\lim_{k \in K \to \infty} \frac{|(\varepsilon_4^{k-1})_j - F_j(\omega^k, \bar{\lambda}^k)|}{p\lambda_1^k \psi'(\phi(\mu_{k-1}, \omega_j^k))} = \frac{|F_j(\omega^*, \bar{\lambda}^*)|}{p\lambda_1^* \psi'(0)}.$$

Thus,

$$\lim_{k \in K \to \infty} \frac{c}{(S_j^k)^r + c} = 1.$$

It follows that $\lim_{k \in K \to \infty} S_j^k = 0$, as desired. This completes the proof. □

We now focus on the sequence $\left\{\nabla^2 \varphi_{\mu_{k-1}}(\omega^k)\right\}_{k \in K}$. Using Lemma 4.2, we prove the following important result.

Lemma 4.3 *Suppose that Assumptions (B1)-(B4) hold. Let (ω^*, λ^*) be an arbitrary accumulation point of the sequence $\{(\omega^k, \lambda^k)\}$ generated by Algorithm 1 and let $\{(\omega^k, \lambda^k)\}_{k \in K}$ be an arbitrary subsequence converging to (ω^*, λ^*). Then*

$$\lim_{k \in K \to \infty} |(\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj}| = \infty \text{ for } j \in I(\omega^*).$$

Proof We first consider the case when $p = 1$. In this instance, we have from Eq. (38) that

$$(\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj} = \psi''\left(\phi(\mu_{k-1}, \omega_j^k)\right) \phi'(\mu_{k-1}, \omega_j^k)^2 + \psi'\left(\phi(\mu_{k-1}, \omega_j^k)\right) \phi''(\mu_{k-1}, \omega_j^k).$$

It follows from Assumption (A) and Proposition 2.3(c) that

$$\psi''\left(\phi(\mu_{k-1}, \omega_j^k)\right) \phi'(\mu_{k-1}, \omega_j^k)^2 \geq -\beta,$$

and there exists $\gamma > 0$ such that for sufficient large k,

$$\psi'\left(\phi(\mu_{k-1}, \omega_j^k)\right) \geq \gamma > 0.$$

Hence, for sufficiently large k, we obtain

$$(\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj} \geq -\beta + \gamma \phi''(\mu_{k-1}, \omega_j^k) = -\beta + \gamma \frac{2}{\mu_{k-1}} \rho\left(\frac{\omega_j^k}{\mu_{k-1}}\right),$$

where $\phi''(\mu_{k-1}, \omega_j^k)$ is defined in (36).

By Lemma 4.2, there exists $M > 0$ such that $\frac{|\omega_j^k|}{\mu_{k-1}} \leq M$ for all $k \in K$. Since ρ is nonincreasing on $[0, \infty)$ by Assumption (B2), we have

$$\begin{aligned}\lim_{k\in K\to\infty} (\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj} &\geq -\beta + \lim_{k\in K\to\infty} \frac{2\gamma}{\mu_{k-1}} \rho\left(\frac{\omega_j^k}{\mu_{k-1}}\right) \\ &\geq -\beta + \lim_{k\in K\to\infty} \frac{2\gamma}{\mu_{k-1}} \rho(M) = \infty,\end{aligned}$$

where the rightmost equality holds since $\rho(t) > 0$ on $\Re$ by Assumption (B4). This proves the claim for $p = 1$.

We now consider the case when $0 < p < 1$. We look at two disjoint subsets of K:

$$U_1^j := \{k \in K \mid \omega_j^k = 0\}, \quad \text{and} \quad U_2^j := \{k \in K \mid \omega_j^k \neq 0\},$$

and the corresponding subsequences. For $k \in U_1^j$, we get from (35) that $\phi'(\mu_{k-1}, 0) = 0$. From (38), we have

$$(\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj} = p\psi'\left([\phi(\mu_{k-1}, \omega_j^k)]^p\right) [\phi(\mu_{k-1}, \omega_j^k)]^{p-1} \phi''(\mu_{k-1}, \omega_j^k). \quad (52)$$

Meanwhile, from (36),

$$\lim_{k\in U_1^j\to\infty} \phi''(\mu_{k-1}, \omega_j^k) = \lim_{k\in U_1^j\to\infty} \frac{2}{\mu_{k-1}} \rho\left(\frac{\omega_j^k}{\mu_{k-1}}\right) = \lim_{k\in U_1^j\to\infty} \frac{2}{\mu_{k-1}} \rho(0) = \infty, \quad (53)$$

where we note that $\rho(0) > 0$ by Assumption (B1) and the definition of density function. Moreover, it is clear that

$$\lim_{k\in U_1^j\to\infty} [\phi(\mu_{k-1}, \omega_j^k)]^{p-1} = \infty. \quad (54)$$

It follows from (52), (53), (54) and Assumption (A) that

$$\lim_{k\in U_1^j\to\infty} |(\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj}| = \infty.$$

For $k \in U_2^j$, we obtain

$$\lim_{k\in U_2^j\to\infty} \phi''(\mu_{k-1}, \omega_j^k) = \lim_{k\in U_2^j\to\infty} \frac{2}{\mu_{k-1}} \rho\left(\frac{\omega_j^k}{\mu_{k-1}}\right) = \infty, \quad (55)$$

by using Eq. (36) and the facts that $\lim_{k\in U_2^j\to\infty}\frac{|\omega_j^k|}{\mu_{k-1}}=0$ from Lemma 4.2 and ρ is continuous by Assumption (B2).

Meanwhile, we obtain $\phi''(\mu_{k-1},\omega_j^k)\neq 0$ for sufficiently large k using Lemma 4.2, Eq. (36) and Assumption (B1). Thus, invoking Assumption (A), we have for large $k\in U_2^j$ that

$$\begin{aligned}
&\psi''\left([\phi(\mu_{k-1},\omega_j^k)]^p\right)\frac{[\phi'(\mu_{k-1},\omega_j^k)]^2[\phi(\mu_{k-1},\omega_j^k)]^{p-1}}{\phi''(\mu_{k-1},\omega_j^k)}\\
&=\psi''\left([\phi(\mu_{k-1},\omega_j^k)]^p\right)\frac{[\phi'(\mu_{k-1},\omega_j^k)]^2[\phi(\mu_{k-1},\omega_j^k)]^{p}}{\phi''(\mu_{k-1},\omega_j^k)\phi(\mu_{k-1},\omega_j^k)}\\
&\geq -\beta\frac{[\phi'(\mu_{k-1},\omega_j^k)]^2[\phi(\mu_{k-1},\omega_j^k)]^{p}}{\phi''(\mu_{k-1},\omega_j^k)\phi(\mu_{k-1},\omega_j^k)}.
\end{aligned}\tag{56}$$

Using (36) again, the symmetry of ρ, and Proposition 2.3(b), we have

$$\phi''(\mu_{k-1},\omega_j^k)\phi(\mu_{k-1},\omega_j^k)\geq\frac{2}{\mu_{k-1}}\rho\left(\frac{\omega_j^k}{\mu_{k-1}}\right)|\omega_j^k|=2S_j^k\rho(S_j^k).$$

Using this fact together with (56) and Eq. (35), we obtain

$$\begin{aligned}
&\psi''\left([\phi(\mu_{k-1},\omega_j^k)]^p\right)\frac{[\phi'(\mu_{k-1},\omega_j^k)]^2[\phi(\mu_{k-1},\omega_j^k)]^{p-1}}{\phi''(\mu_{k-1},\omega_j^k)}\\
&\geq -\beta\frac{2\left[\int_0^{S_j^k}\rho(t)dt\right]^2[\phi(\mu_{k-1},\omega_j^k)]^{p}}{S_j^k\rho(S_j^k)}\\
&\geq -\beta\frac{2S_j^k(\rho(0))^2[\phi(\mu_{k-1},\omega_j^k)]^{p}}{\rho(S_j^k)}\ \to\ 0 \text{ as } k\in U_2^j\to\infty,
\end{aligned}\tag{57}$$

where the last inequality holds since $\left[\int_0^{S_j^k}\rho(t)dt\right]^2\leq(S_j^k)^2(\rho(0))^2$. Similarly, we also have

$$\begin{aligned}
(p-1)\frac{[\phi'(\mu_{k-1},\omega_j^k)]^2}{\phi''(\mu_{k-1},\omega_j^k)\phi(\mu_{k-1},\omega_j^k)}&\geq(p-1)\frac{2\left[\int_0^{S_j^k}\rho(t)dt\right]^2}{S_j^k\rho(S_j^k)}\\
&\geq(p-1)\frac{2S_j^k(\rho(0))^2}{\rho(S_j^k)}\ \to\ 0 \text{ as } k\in U_2^j\to\infty.
\end{aligned}\tag{58}$$

For brevity, denote $\phi_k \coloneqq \phi(\mu_{k-1}, \omega_j^k)$, $\phi_k' \coloneqq \phi'(\mu_{k-1}, \omega_j^k)$, $\phi_k'' \coloneqq \phi''(\mu_{k-1}, \omega_j^k)$. Then (38) can be written as

$$(\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj} = p\phi_k'' \phi_k^{p-1} \left[p\psi''\left(\phi_k^p\right) \frac{[\phi_k']^2 \phi_k^{p-1}}{\phi_k''} + (p-1)\psi'\left(\phi_k^p\right) \frac{[\phi_k']^2}{\phi_k'' \phi_k} + \psi'\left(\phi_k^p\right) \right]. \tag{59}$$

Since $0 < \psi'(t) \leq \alpha$ by Assumption (A), we have

$$\lim_{k \in U_2^j \to \infty} \left[p\psi''\left(\phi_k^p\right) \frac{[\phi_k']^2 \phi_k^{p-1}}{\phi_k''} + (p-1)\psi'\left(\phi_k^p\right) \frac{[\phi_k']^2}{\phi_k'' \phi_k} + \psi'\left(\phi_k^p\right) \right] > 0$$

by using the obtained limits (57) and (58). On the other hand, it is clear from (55) that $\phi_k'' \phi_k^{p-1} \to \infty$ as $k \in U_2^j \to \infty$. Hence, taking the limit in (59),

$$\lim_{k \in U_2^j \to \infty} (\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj} = \infty.$$

This completes the proof. □

Remark 2 Note that the main result presented in Theorem 4.1 considers an arbitrary accumulation point of a sequence $\{(\omega^k, \lambda^k, \zeta^k, \eta^k)\}$ generated by Algorithm 1. The existence of such accumulation points is guaranteed if the generated sequence is bounded. We defer our discussion on the boundedness of the sequence to Sect. 4.3 in order to focus solely on the main ideas for proving Theorem 4.1. We highlight that Lemma 4.3, which relies only on Assumption (B), serves as one of our primary tools for establishing boundedness (see Proposition 4.2). However, an additional assumption and lemma are required to prove boundedness, and thus, we postpone the discussion of the details to Sect. 4.3.

Lemma 4.4 *Suppose that Assumptions (B1)-(B4) hold, and $(\omega^*, \lambda^*, \zeta^*, \eta^*)$ is an accumulation point of the sequence $\{(\omega^k, \lambda^k, \zeta^k, \eta^k)\}$ generated by Algorithm 1. Then*

(i) $\zeta_j^* = 0$ for all $j \in I(\omega^*)$, that is, $\breve{\zeta}^* = 0$; and
(ii) $p \sum_{j \notin I(\omega^*)} \operatorname{sgn}(\omega_j^*) |\omega_j^*|^{p-1} \psi'(|\omega_j^*|^p) \zeta_j^* = \eta_1^*$.

Proof Let $\{(\omega^k, \lambda^k, \zeta^k, \eta^k)\}_{k \in K}$ be a subsequence converging to $(\omega^*, \lambda^*, \zeta^*, \eta^*)$. From (43), we have for all $k \in K$ that

$$(\nabla f(\omega^k))_j + \left(\nabla^2_{\omega\omega} G(\omega^k, \bar{\lambda}^k)\zeta^k\right)_j + \lambda_1^k (\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj} \zeta_j^k = (\varepsilon_1^{k-1})_j, \ (j = 1, 2, \ldots, n).$$

Since G is twice continuously differentiable and f is continuously differentiable, then $\{\lambda_1^k (\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj} \zeta_j^k\}_{k \in K}$ is a bounded sequence for each j. Consequently,

$\zeta_j^* = 0$ for all $j \in I(\omega^*)$, since $\lambda_1^* > 0$ and $\lim_{k\in K\to\infty}(\nabla^2\varphi_{\mu_{k-1}}(\omega^k))_{jj} = +\infty$ for each $j \in I(\omega^*)$ by Lemma 4.3. This proves part (i).

To prove part (ii), we note from (46) that for all $k \in K$,

$$F_j(\omega^k, \bar{\lambda}^k) + \lambda_1^k(\nabla\varphi_{\mu_{k-1}}(\omega^k))_j = (\varepsilon_4^{k-1})_j,\ (j = 1, 2, \ldots, n),$$

so that $\{(\nabla\varphi_{\mu_{k-1}}(\omega^k))_j\}_{k\in K}$ is convergent since F_j given by (49) is continuous and $\lambda_1^* > 0$. Hence, we obtain from item (i) that $\lim_{k\in K\to\infty}(\nabla\varphi_{\mu_{k-1}}(\omega^k))_j\zeta_j^k = 0$, $(j \in I(\omega^*))$ and therefore,

$$\lim_{k\in K\to\infty}\sum_{j\in I(\omega^*)}(\nabla\varphi_{\mu_{k-1}}(\omega^k))_j\zeta_j^k = 0.$$

Together with (39) and (44), it follows that

$$\begin{aligned}\eta_1^* = \lim_{k\in K\to\infty}\eta_1^k &= \lim_{k\in K\to\infty}\nabla\varphi_{\mu_{k-1}}(\omega^k)^T\zeta^k\\ &= \lim_{k\in K\to\infty}\sum_{j\notin I(\omega^*)}(\nabla\varphi_{\mu_{k-1}}(\omega^k))_j\zeta_j^k\\ &= \sum_{j\notin I(\omega^*)} p\,\mathrm{sgn}(\omega_j^*)|\omega_j^*|^{p-1}\psi'(|\omega_j^*|^p)\zeta_j^*,\end{aligned}$$

proving part (ii). □

Remark 3 Crucial in the proofs of Lemmas 4.2–4.4 is the strict positivity of the regularization parameter λ_1^* that corresponds to an accumulation point of $\{\lambda_1^k\}$. This is automatically guaranteed by the constraint set (5), as opposed to the work of [21] where the parameter ϵ is set to 0. In turn, [21] requires the assumption that $\liminf_{k\to\infty}\lambda_1^k > 0$ to ensure that $\lambda_1^* > 0$, but such an assumption is difficult to guarantee for the iterates generated by Algorithm 1.

Having derived all the necessary lemmas, we can now prove our main result.

Proof of Theorem 4.1 Let $\{(\omega^k, \lambda^k, \zeta^k, \eta^k)\}_{k\in K}$ be a subsequence converging to an accumulation point $(\omega^*, \lambda^*, \zeta^*, \eta^*)$. It is clear from (45) and (47) that Eqs. (23) and (24) hold. Meanwhile, we obtain from (43) and (46), respectively, that

$$\nabla_{\tilde{\omega}} f(\omega^k) + (\nabla^2_{\tilde{\omega}\tilde{\omega}} G(\omega^k, \bar{\lambda}^k)\tilde{\zeta}^k + (\nabla^2_{\tilde{\omega}\check{\omega}} G(\omega^k, \bar{\lambda}^k)\check{\zeta}^k + \lambda_1^k\nabla^2_{\tilde{\omega}}\varphi_{\mu_{k-1}}(\omega^k))\tilde{\zeta}^k = \tilde{\varepsilon}_1^{k-1}, \tag{60}$$

$$\nabla_{\tilde{\omega}} G(\omega^k, \bar{\lambda}^k) + \lambda_1^k\nabla_{\tilde{\omega}}\varphi_{\mu_{k-1}}(\omega^k) = \tilde{\varepsilon}_4^{k-1}, \tag{61}$$

where $\tilde{\varepsilon}_1^{k-1} = \{(\varepsilon_1^{k-1})_j\}_{j\notin I(\omega^*)}$ and $\tilde{\varepsilon}_4^{k-1} = \{(\varepsilon_4^{k-1})_j\}_{j\notin I(\omega^*)}$. Using Lemma 4.1 and Lemma 4.4(i), and letting $k \in K \to \infty$ in (60) and (61), we obtain the bilevel KKT conditions (19) and (20). Finally, (21) and (22) hold by Lemma 4.4. This completes the proof of Theorem 4.1. □

4.3 Boundedness

In the preceding discussion, we have shown that accumulation points of $\{(\omega^k, \lambda^k, \zeta^k, \eta^k)\}$ correspond to bilevel KKT points. The existence of these accumulation points is guaranteed by boundedness of the full sequence $\{(\omega^k, \lambda^k, \zeta^k, \eta^k)\}$. In this section, we show that the boundedness of $\{(\omega^k, \lambda^k)\}$ is sufficient to conclude that the $\{(\zeta^k, \eta^k)\}$ is likewise bounded.

4.3.1 Weaker constraint qualification

In [21], linearly independent constraint qualification (LICQ) was one of the assumptions used to obtain the boundedness of the sequence $\{(\omega^k, \lambda^k, \zeta^k, \eta^k)\}$. In this present work, we only assume that the Mangasarian-Fromovitz constraint qualification holds at accumulation points of a sequence generated by the smoothing algorithm.

Assumption (C). Let $(\omega^*, \lambda^*) \in \Re^n \times \Re^r$ be an accumulation point of $\{(\omega^k, \lambda^k)\}$ generated by Algorithm 1. Denote

$$I(\lambda^*) := \{i \in \{1, 2, \ldots, r\} \mid \lambda_i^* = \epsilon(e_1)_i, \}$$

where $e_1 = (1, 0, \ldots, 0) \in \Re^r$, and

$$\Phi_j(\omega, \lambda) := \frac{\partial G(\omega, \bar{\lambda})}{\partial \omega_j} + p\text{sgn}(\omega_j)\lambda_1|\omega_j|^{p-1}\psi'(|\omega_j|^p) \ (j \notin I(\omega^*)).$$

Then, the Mangasarian-Fromovitz constraint qualification (MFCQ) holds at $(\omega, \lambda) = (\omega^*, \lambda^*)$ for the constraints $\Psi_j(\omega, \lambda) = 0$ for all $j = 1, \ldots, n$ and $\lambda \geq 0$, where

$$\Psi_j(\omega, \lambda) := \begin{cases} \Phi_j(\omega, \lambda) & \text{if } j \notin I(\omega^*), \\ \omega_j & \text{if } j \in I(\omega^*). \end{cases}$$

That is, $\{\nabla_{(\omega,\lambda)}\Psi_j(\omega^*, \lambda^*)\}_{j=1}^n$ is linearly independent and there exists $\bar{d} \in \Re^{n+r}$ such that

$$\nabla_{(\omega,\lambda)}\Psi_j(\omega^*, \lambda^*)^T\bar{d} = 0 \quad \forall j = 1, \ldots, n, \tag{62}$$

$$(\nabla_{(\omega,\lambda)}\lambda_i|_{(\omega,\lambda)=(\omega^*,\lambda^*)})^T\bar{d} > 0 \qquad \forall i \in I(\lambda^*). \tag{63}$$

The following lemma is needed for subsequent analysis.

Lemma 4.5 *Suppose that (ω^*, λ^*) is an arbitrary accumulation point of the sequence $\{(\omega^k, \lambda^k)\}$ such that Assumption (C) holds. Then, $\{\nabla_{(\tilde{\omega},\lambda)}\Phi_j(\omega^*, \lambda^*)\}_{j\notin I(\omega^*)}$ is linearly independent and there exists a vector $d \in \Re^{n-|I(\omega^*)|+r}$ such that*

$$\nabla_{(\tilde{\omega},\lambda)}\Phi_j(\omega^*, \lambda^*)^T d = 0 \quad \forall j \notin I(\omega^*), \tag{64}$$

$$(\nabla_{(\tilde{\omega},\lambda)}\lambda_i|_{(\omega,\lambda)=(\omega^*,\lambda^*)})^T d > 0 \quad \forall i \in I(\lambda^*), \tag{65}$$

where $\tilde{\omega} := (\omega_j)_{j \notin I(\omega^*)}$.

Proof See Appendix C. □

4.3.2 Boundedness of algorithm iterates

We will now show the boundedness of the sequence of Lagrange multiplier vectors $\{(\zeta^k, \eta^k)\}$ in the following lemma.

Proposition 4.2 *Suppose that Assumptions (B) and (C) hold. Let* $\{(\zeta^k, \eta^k)\} \subseteq \Re^n \times \Re^r$ *be a sequence of the accompanying Lagrange multiplier vectors generated by Algorithm 1. If* $\{(\omega^k, \lambda^k)\}$ *is bounded, then* $\{(\zeta^k, \eta^k)\}$ *is bounded.*

Proof For convenience, denote

$$\xi^k := ((\zeta^k)^T, (\eta^k)^T)^T, \quad \hat{\zeta}^k := \frac{\zeta^k}{\|\xi^k\|}, \quad \hat{\eta}^k := \frac{\eta^k}{\|\xi^k\|}$$

for each k. We prove by contradiction that the sequence $\{(\zeta^k, \eta^k)\}$ is bounded. Without loss of generality, we may assume that

$$\|\xi^k\| \to \infty, \ \lim_{k\to\infty} \frac{\xi^k}{\|\xi^k\|} = \hat{\xi}^*,$$

where $\hat{\xi}^* := ((\hat{\zeta}^*)^T, (\hat{\eta}^*)^T)^T$ with $\hat{\zeta}^*$ and $\hat{\eta}^*$ are accumulation points of $\{\hat{\zeta}^k\}$ and $\{\hat{\eta}^k\}$, respectively. We may suppose without loss of generality that $\lim_{k\to\infty}(\omega^k, \lambda^k) = (\omega^*, \lambda^*)$. Dividing by $\|\xi^k\|$ both sides of (30), (31), (32) and (34) evaluated at $(\omega, \lambda, \zeta, \eta) = (\omega^k, \lambda^k, \zeta^k, \eta^k)$ and $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4, \varepsilon_5) = (\varepsilon_1^{k-1}, \varepsilon_2^{k-1}, \varepsilon_3^{k-1}, \varepsilon_4^{k-1}, \varepsilon_5^{k-1})$, we have for each k the following equations:

$$\frac{(\nabla f(\omega^k))_j}{\|\xi^k\|} + \left(\nabla^2_{\omega\omega} G(\omega^k, \bar{\lambda}^k)\hat{\zeta}^k\right)_j + \lambda_1^k (\nabla^2 \varphi_{\mu_{k-1}}(\omega^k))_{jj}\hat{\zeta}_j^k = \frac{(\varepsilon_1^{k-1})_j}{\|\xi^k\|} \ (j = 1, 2, \ldots, n), \tag{66}$$

$$\nabla \varphi_{\mu_{k-1}}(\omega^k)^T \hat{\zeta}^k - \hat{\eta}_1^k = \frac{\varepsilon_2^{k-1}}{\|\xi^k\|}, \tag{67}$$

$$\nabla R_j(\omega^k)^T \hat{\zeta}^k - \hat{\eta}_j^k = \frac{(\varepsilon_3^{k-1})_{j-1}}{\|\xi^k\|} \ (j = 2, 3, \ldots, r), \tag{68}$$

$$\lambda_j^k - \epsilon(e_1)_j \geq 0,$$

$$\hat{\eta}_j^k \geq 0, \quad (\lambda_j^k - \epsilon(e_1)_j)\hat{\eta}_j^k \leq \sum_{j=1}^{r} (\lambda_j^k - \epsilon(e_1)_j)\hat{\eta}_j^k = \frac{\varepsilon_5^{k-1}}{\|\xi^k\|} \quad (j = 1, \ldots, r). \tag{69}$$

Since $\lim_{k\to\infty} \frac{\varepsilon_l^{k-1}}{\|\xi^k\|} = 0, l = 1, 2, 3, 5$, letting $k \to \infty$ in inequality (69) gives

$$\hat{\eta}_j^* = 0 \ (j \notin I(\lambda^*)) \ \text{ and } \ \hat{\eta}_j^* \geq 0 \ (j \in I(\lambda^*)). \tag{70}$$

Since $\|\hat{\xi}^*\| = 1$, we get from (70) that

$$1 = \|\hat{\zeta}^*\|^2 + \|\hat{\eta}^*\|^2 = \|\hat{\zeta}^*\|^2 + \sum_{j\in I(\lambda^*)} |\hat{\eta}_j^*|^2. \tag{71}$$

Meanwhile, since G is twice continuously differentiable and f is continuously differentiable, then both $\{\nabla^2_{\omega\omega} G(\omega^k, \bar{\lambda}^k)\hat{\zeta}^k\}$ and $\left\{\frac{\nabla f(\omega^k)}{\|\xi^k\|}\right\}$ are bounded. This implies the boundedness of $\{\lambda_1^k(\nabla^2\varphi_{\mu_{k-1}}(\omega^k))_{jj}\hat{\zeta}_j^k\}$ for each j by (66). Consequently, we obtain $\lim_{k\to\infty} \hat{\zeta}_j^k = 0$ for $j \in I(\omega^*)$ by Lemma 4.3 and noting that $\lambda_1^* > 0$. That is,

$$\hat{\zeta}_j^* = 0 \ (j \in I(\omega^*)). \tag{72}$$

Letting $k \to \infty$ in (33), it is clear that

$$\lim_{k\to\infty} |\nabla(\varphi_{\mu_{k-1}}(\omega^k))_j| = \frac{|F_j(\omega^*, \bar{\lambda}^*)|}{\lambda_1^*},$$

where $F_j(\omega^*, \bar{\lambda}^*)$ is given by (49). This together with (72) gives us

$$\lim_{k\to\infty} \sum_{j\in I(\omega^*)} (\nabla\varphi_{\mu_{k-1}}(\omega^k))_j \hat{\zeta}_j^k = 0. \tag{73}$$

Thus,

$$\begin{aligned}
\hat{\eta}_1^* &\overset{(67)}{=} \lim_{k\to\infty} \nabla\varphi_{\mu_{k-1}}(\omega^k)^T \hat{\zeta}^k \\
&= \lim_{k\to\infty} \left(\sum_{j\in I(\omega^*)} (\nabla\varphi_{\mu_{k-1}}(\omega^k))_j \hat{\zeta}_j^k + \sum_{j\notin I(\omega^*)} (\nabla\varphi_{\mu_{k-1}}(\omega^k))_j \hat{\zeta}_j^k \right) \\
&\overset{(73)}{=} \lim_{k\to\infty} \sum_{j\notin I(\omega^*)} (\nabla\varphi_{\mu_{k-1}}(\omega^k))_j \hat{\zeta}_j^k \\
&\overset{(39)}{=} \sum_{j\notin I(\omega^*)} p\,\mathrm{sgn}(\omega_j^*)|\omega_j^*|^{p-1}\psi'(|\omega_j^*|^p)\hat{\zeta}_j^*.
\end{aligned} \tag{74}$$

On the other hand, we have from (68) and (72) that

$$\hat{\eta}_j^* = \sum_{i\notin I(\omega^*)} \frac{\partial R_j(\omega^*)}{\partial \omega_i} \hat{\zeta}_i^*, \quad (j = 2, \ldots, r). \tag{75}$$

Meanwhile, letting $k \to \infty$ in (66) and using Eqs. (72) and (40), we have

$$\left(\nabla^2_{\tilde{\omega}\tilde{\omega}} G(\omega^*, \bar{\lambda}^*) + \lambda_1^* p^2 \psi''\left(|\tilde{\omega}_j^*|^p\right)|\tilde{\omega}_j^*|^{2p-2} + \lambda_1^* p(p-1)\psi'\left(|\tilde{\omega}_j^*|^p\right)|\tilde{\omega}_j^*|^{p-2}\right)\tilde{\hat{\zeta}}^* = 0, \tag{76}$$

where $\tilde{\hat{\zeta}}^* = (\hat{\zeta}_i^*)_{i \notin I(\omega^*)}$. Combining Eqs. (74), (75) and (76), we obtain

$$\sum_{j \notin I(\omega^*)} \hat{\zeta}_j^* \nabla_{(\tilde{\omega},\lambda)} \Phi_j(\omega^*, \lambda^*) - \sum_{j \in I(\lambda^*)} \hat{\eta}_j^* \nabla_{(\tilde{\omega},\lambda)} \lambda_j(\omega^*, \lambda^*) = 0, \tag{77}$$

where $\tilde{\omega} := (\omega_j)_{j \notin I(\omega^*)}$ and $\Phi_j(\omega, \lambda)$ ($j \notin I(\omega^*)$) are as defined in Assumption (C). On the other hand, by Lemma 4.5, we can find a vector $d \in \Re^{n-|I(\omega^*)|+r}$ such that (64) and (65) hold. From (77), we have

$$\sum_{j \notin I(\omega^*)} \hat{\zeta}_j^* \nabla_{(\tilde{\omega},\lambda)} \Phi_j(\omega^*, \lambda^*)^T d - \sum_{j \in I(\lambda^*)} \hat{\eta}_j^* \nabla_{(\tilde{\omega},\lambda)} \lambda_j(\omega^*, \lambda^*)^T d = 0.$$

Together with equation (64), we obtain

$$\sum_{j \in I(\lambda^*)} \hat{\eta}_j^* \nabla_{(\tilde{\omega},\lambda)} \lambda_j(\omega^*, \lambda^*)^T d = 0.$$

Consequently, we have from (70) and (65) that $\hat{\eta}_j^* = 0$ for all $j \in I(\lambda^*)$. In turn, (77) implies that

$$\sum_{j \notin I(\omega^*)} \hat{\zeta}_j^* \nabla_{(\tilde{\omega},\lambda)} \Phi_j(\omega^*, \lambda^*) = 0.$$

Since $\{\nabla_{(\tilde{\omega},\lambda)} \Phi_j(\omega^*, \lambda^*)\}_{j \notin I(\omega^*)}$ is linearly independent by Lemma 4.5, then $\hat{\zeta}_j^* = 0$ for all $j \notin I(\omega^*)$. Together with (72), we have $\hat{\zeta}^* = 0$ which in turn implies that $\|\hat{\zeta}^*\|^2 + \sum_{j \in I(\lambda^*)} |\hat{\eta}_j^*|^2 = 0$. This contradicts (71). Therefore, the sequence $\{(\zeta^k, \eta^k)\}$ is bounded. □

5 Numerical results

We compare the efficiency of different smoothing functions, namely the functions ϕ_i ($i = 1, 2, \ldots, 6$) presented in Appendix B by means of numerical simulation. The program is coded in MATLAB R2022b and run on a machine with Intel(R) Core(TM) i7-7500U CPU@2.70GHz and 8.0 GB RAM.

5.1 Problem with an elastic-net-type regularizer

We solve the following bilevel problem arising from squared linear regression using an Elastic-Net-type regularizer:

$$\begin{aligned} \min_{\omega,\lambda} \quad & \frac{1}{2}\|A_1\omega - b_1\|_2^2 \\ \text{s.t.} \quad & \omega \in \operatorname*{argmin}_{\hat{\omega}\in\Re^n} \left\{ \frac{1}{2}\|A_2\hat{\omega} - b_2\|_2^2 + \lambda_1\|\hat{\omega}\|_p^p + \lambda_2\|\hat{\omega}\|_2^2 \right\} \\ & \lambda_1 \geq \epsilon, \quad \lambda_2 \geq 0, \end{aligned} \tag{78}$$

where $A_i \in \Re^{m_i \times n}$, $b_i \in \Re^{m_i}$ for $i \in \{1, 2\}$ and $\epsilon = 10^{-6}$. We produce 20 synthetic problems for $(n, m_1, m_2) = (500, 1000, 1000)$ and for $(n, m_1, m_2) = (500, 300, 300)$ generated in Matlab as follows:

$$\begin{aligned} & A_i := \texttt{rand}(m_i, n), \quad \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} := \begin{bmatrix} A_1 * \theta \\ A_2 * \theta + 0.01 * (2 * \texttt{rand}\,(m_2, 1) - \texttt{ones}\,(m_2, 1)) \end{bmatrix}, \\ & \theta := \texttt{zeros}(n, 1), \quad \theta(\texttt{randsample}(n, 0.15 * n)) = -5 + 10 * \texttt{rand}(0.15 * n, 1), \end{aligned}$$

with `rand`, `randn`, `randsample`, `ones`, and `zeros` being MATLAB commands, and apply Algorithms 1 with the smoothing functions ϕ_i $(i = 1, 2, \ldots, 6)$ to the problems (78) with the generated data. The random number generator is initialized at `default`. The test data $A_3 \in \Re^{m_3 \times n}$ and $b_3 \in \Re^{m_3}$ are generated in the same manner as A_i and b_i for $i = 1, 2$, with $m_3 := m_1$.

In order to compute a KKT point of the smoothed subproblem for (78) in Step 1 of Algorithm 1, we utilize the MATLAB solver `fmincon` with "MaxIterations= 10^4" and opt for the interior-point method as an algorithm that runs within `fmincon`. We initialize `fmincon` for (29) at some initial point (ω^0, λ^0) in the first iteration $k = 0$, and then use the previous iteration point $(\omega^{k-1}, \lambda^{k-1})$ as the initial point for the succeeding iterations, i.e., for $k \geq 1$. The smoothing parameter is initialized at $\mu_0 = 0.1$, and the factor of decrease is set to $\beta_1 = 0.8$. To obtain a reasonable initial point (ω^0, λ^0), we employ a semismooth Newton (SSN) method for solving the KKT system (30)–(34). We first use a complementarity function to reformulate the conditions (34) with $\varepsilon_5 = 0$ as a system of equations [1]. In particular, we use the Fischer-Burmeister function $\phi_{\text{FB}} : \Re^r \to \Re^r$ given by

$$\phi_{\text{FB}}(x, y) = x + y - \sqrt{x^2 + y^2},$$

where the operations are understood to be taken component-wise, so that the conditions (34) with $\varepsilon_5 = 0$ are equivalent to solving

$$\phi_{\text{FB}}(\lambda - \epsilon e_1, \eta) = 0. \tag{79}$$

With this, a KKT point satisfying (30)–(34) can be obtained by solving approximately the equation

$$\Phi^{\mu}_{\mathrm{FB}}(\omega, \lambda, \zeta, \eta) := \begin{pmatrix} \nabla f(\omega) + (\nabla^2_{\omega\omega} G(\omega, \bar{\lambda}) + \lambda_1 \nabla^2 \varphi_\mu(\omega))\zeta \\ \nabla \varphi_\mu(\omega)^T \zeta - \eta_1 \\ \nabla \bar{R}(\omega)^T \zeta - \bar{\eta} \\ \nabla_\omega G(\omega, \bar{\lambda}) + \lambda_1 \nabla \varphi_\mu(\omega) \\ \phi_{\mathrm{FB}}(\lambda - \epsilon e_1, \eta) \end{pmatrix} = 0. \tag{80}$$

By our differentiability assumptions on f, g, R_j $(j = 2, \ldots, r)$ and the smoothness of φ_μ, Eqs. (31)–(33) are all smooth. On the other hand, from Eqs. (36) and (38) and invoking Assumptions (A) and (B), Eq. (30) is semismooth provided that ρ is semismooth, which is the case for piecewise smooth functions [13, Proposition 7.4.6], such as the density functions that we considered in Appendix B. Finally, since ϕ_{FB} is strongly semismooth [13, Proposition 7.4.8], then Eq. (79) is likewise a semismooth equation. Hence, for solving (80), we may employ the semismooth Newton method, the convergence of which has already been established (see, for instance, [13, Theorem 7.5.2]). Our warmstarting algorithm to obtain an initial point (ω^0, λ^0) is described in Algorithm 2.[4] Similar to Algorithm 1, we consider a sequence of Eq. (80) for decreasing values of μ. In our experiments, we set $\tau_1 = 0.8$, $\tau_2 = 0.1$, $\gamma_{\min} = 10^{-6}$, $\mu_0 = 10$, and $\gamma_0 = 0.1$. We initialize Algorithm 2 with $\omega^0 = 100 * \texttt{ones}(n, 1)$ and $\lambda^0 = (\epsilon, 0)$, $\zeta^0 = 0$, and $\eta^0 = 0$.

Algorithm 2 (A semismooth Newton method for warmstarting Algorithm 1)

Step 0 Choose $\mu_0, \gamma_0 > 0$, $\tau_1, \tau_2, \gamma_{\min} \in (0, 1)$, and $z^0 := (\omega^0, \lambda^0, \zeta^0, \eta^0)$. Set $k := 0$.

Step 1 Select an element $V^k \in \partial_C \Phi^{\mu_k}_{\mathrm{FB}}(z^k)$, where ∂_C denotes the Clarke subdifferential (see [13, Definition 7.1.1]), and solve the linear system

$$\Phi^{\mu_k}_{\mathrm{FB}}(z^k) + V^k \Delta z^k = 0.$$

Step 2 Set $z^{k+1} := z^k + \Delta z^k$.

Step 3 Set

$$(\mu_{k+1}, \gamma_{k+1}) := \begin{cases} (\mu_k, \gamma_k) & \text{if } \|\Phi^{\mu_k}_{\mathrm{FB}}(z^k)\| > \gamma_k \\ (\tau_1 \mu_k, \min\{\tau_2 \gamma_k, \gamma_{\min}\}) & \text{if } \|\Phi^{\mu_k}_{\mathrm{FB}}(z^k)\| \le \gamma_k \end{cases}$$

If $\|\Phi^{\mu_k}_{\mathrm{FB}}(z^k)\| < \gamma_{\min}$, terminate the algorithm. Otherwise, go to Step 1 and set $k := k + 1$.

[4] One may consider employing the semismooth Newton method, i.e., Algorithm 2, as a standalone algorithm for obtaining BKKT points. However, relying on this algorithm alone does not yield accurate solution for the BKKT system, given as well that its convergence guarantee is only local. Alternatively, a globally subsequentially convergent semismooth Newton method, incorporating an Armijo-type linesearch algorithm, was proposed in [12]. However, due to the linesearch scheme, this approach is more computationally expensive, making it less suitable especially as a warmstarting algorithm. Moreover, another challenge in semismooth Newton-type algorithms is that apart from providing an initial guess for primal variables (ω^0, λ^0), we also need to provide an initial guess for the Lagrange multipliers (ζ^0, η^0) when using Algorithm 2, which could influence the quality of the solution obtained by the algorithm.

In light of the SB-KKT conditions (13)–(18) and the value of the smoothing parameter μ, we terminate the algorithm when either one of the following criteria is met:

1. The norms of the residuals of the equations in (19)–(24) are smaller than 10^{-2}. To estimate the index set $I(\omega^*)$ in conditions (15) and (16), we regard ω_i^k as zero if $|\omega_i^k| \leq 10^{-5}$.
2. $\mu_{k+1} \leq 10^{-8}$.

The obtained results are summarized in Tables 1, 2, 3, 4, in which each column is described as follows. Here, the averages are taken over the set of problems that are counted in success(%).

i:	the smoothing function ϕ_i
val:	average validation error at the resulting solution; the validation error is the least squares error for the validation data A_1 at the obtained solution, i.e., the value of the objective function at the resulting solution
test:	average test error at the resulting solution; the test error is the the least squares error for the validation data A_3 at the obtained solution, i.e., the value of the objective function at the resulting solution
bkkt:	average residual of the BKKT conditions
sparsity(%):	average ratio of zero elements of the resulting solution ω^*, in which each element w_i is counted as zero if $\|\omega_i\| \leq 10^{-5}$
time(s):	average time spent by the algorithm in seconds; in parenthesis, we include the average time spent in the initialization phase via Algorithm 2
ssn.iter:	average number of iterations for the initialization phase
iter:	average number of iterations of Algorithm 1 executed by employing Matlab's fmincon built-in function.
success(%):	ratio of problems for which BKKT points are computed successfully in the sense that the first termination condition in the above is satisfied

The best values are displayed in bold in the tables, with the results for the smoothing function ϕ_5 excluded from the tables due to the overflow that often occurred when computing its gradient as μ gets smaller. Now, the following insights are obtained from the numerical results.

Comparison with $p = 0.5, 1$ ***and*** $m_2 = 300, 1000$

In terms of the sparsity of solutions obtained, we see that $\ell_{0.5}$ tends to attain sparser solutions than ℓ_1. Indeed, it is evident from Table 2 (resp., Table 4) that the solutions obtained for $p = 0.5$ are sparser than those obtained by $p = 1$ shown in Table 1 (resp., Table3). This is by virtue of the nonconvexity of ℓ_p with $p < 1$. Moreover, $\ell_{0.5}$ tends to attain solutions with better validation errors than ℓ_1.

On the other hand, the problems with $m_2 < n$ is related to the problem of finding sparse solutions of underdetermined linear systems. Such kind of problems are often regarded more intractable than those with $m_2 \geq n$, as illustrated by the obtained numerical results. When $p = 1$, the success rate of the smoothing algorithm is largely diminished when $m_2 < n$. In addition, it is clear from Tables 1 and 3 that for this

Table 1 Averaged results for $(n, m_2, m_1, p) = (500, 1000, 1000, 1)$

i	val	Test	bkkt	Sparsity(%)	Time(s)	ssn.iter	iter	Success(%)
1	**7.32e-03**	**7.31e-03**	4.98e-03	**45.8**	187.5 (7.2)	**182.1**	34.9	**100**
2	7.90e-03	7.83e-03	5.34e-03	43.4	175.5 (7.2)	183.7	32.5	**100**
3	9.17e-03	9.07e-03	**4.23e-03**	36.3	157.4 (7.2)	183.7	27.7	95
4	1.08e-02	1.06e-02	4.59e-03	29.1	**148.9** (7.1)	185.2	**26.4**	**100**
6	1.01e-02	9.95e-03	4.80e-03	31.4	165.7 (7.2)	182.9	30.9	**100**

Table 2 Averaged results for $(n, m_2, m_1, p) = (500, 1000, 1000, 0.5)$

i	val	Test	bkkt	Sparsity(%)	Time(s)	ssn.iter	iter	Success(%)
1	1.39e-03	1.39e-03	**5.76e-03**	**84.7**	**198.9** (7.5)	180.9	**31.2**	**100**
2	1.36e-03	1.37e-03	6.66e-03	**84.7**	206.8 (7.4)	181.7	32.8	**100**
3	**1.35e-03**	**1.36e-03**	6.38e-03	83.5	213.3 (8.0)	180.8	33.0	**100**
4	3.01e-03	2.95e-03	6.96e-03	75.8	210.9 (7.6)	183.3	32.5	**100**
6	3.02e-03	2.96e-03	6.05e-03	76.4	202.2 (7.4)	**180.4**	32.4	**100**

Table 3 Averaged results for $(n, m_2, m_1, p) = (500, 300, 300, 1)$

i	val	Test	bkkt	Sparsity(%)	time(s)	ssn.iter	iter	Success(%)
1	1.07e-02	1.10e-02	4.43e-03	58.6	**244.5** (4.9)	124.8	**44.4**	**50**
2	**9.68e-03**	**1.02e-02**	4.56e-03	**59.9**	268.6 (4.7)	116.7	45.7	30
3	1.23e-02	1.28e-02	**3.85e-03**	56.2	262.0 (5.2)	132.9	45.1	35
4	1.17e-02	1.23e-02	5.32e-03	58.4	314.1 (6.4)	147.2	52.7	**50**
6	2.24e-02	2.27e-02	7.66e-03	58.5	299.5 (4.4)	**114.5**	56.0	20

instance, the algorithm required more time to solve the problems as compared when $m_2 \geq n$. When $p < 1$, while the average times spent by the algorithm are apparently not very distinct for both $m_2 = 1000$ and $m_2 = 300$, it is evident from Tables 2 and 4 that the success rate is also diminished for the latter case. Meanwhile, for the case $m_2 < n$, we note that the success rate when $p = 0.5$ is significantly better than when $p = 1$.

Comparison of the five smoothing functions

In view of the validation and test errors, bilevel KKT residuals, sparsity, average time and success rates, the qualities of the resulting solutions as well as the efficiency of the algorithm with different smoothing functions are comparable. From Table1, we see that Algorithm 1 with ϕ_4 is the fastest method obtaining a 100% success rate in solving the problems, but the solutions obtained are neither the sparsest ones, nor do they correspond to the lowest validation and test errors. As these factors are quite important in evaluating the performance of the model, we observe that Algorithm 1 equipped with smoothings functions ϕ_1 and ϕ_2 provide higher quality of solutions

Table 4 Averaged results for $(n, m_2, m_1, p) = (500, 300, 300, 0.5)$

i	val	Test	bkkt	Sparsity(%)	Time(s)	ssn.iter	iter	Success(%)
1	2.11e-03	2.25e-03	**5.62e-03**	**80.8**	185.3 (4.9)	**114.5**	33.5	**100**
2	**2.06e-03**	2.19e-03	6.00e-03	73.0	**180.6** (4.7)	129.2	**33.2**	85
3	2.11e-03	2.20e-03	5.88e-03	75.2	189.5 (5.2)	119.5	35.2	90
4	**2.06e-03**	**2.15e-03**	5.68e-03	79.2	200.9 (6.4)	146.1	37.5	95
6	2.58e-03	2.58e-03	5.90e-03	75.2	196.6 (4.4)	120.5	36.4	90

attained at a running time not significantly longer than that required by ϕ_4. Considering these important criteria along with the success rates of the algorithms, we also observe from Tables 2, 3, 4 that the algorithm equipped with ϕ_1 consistently obtains the best success rates with low validation error, as well as sparser solutions.

5.2 Problems with other regularizers

In this section, we solve problem (78) with the regularizers $\psi_2(\|\omega\|_p^p)$ and $\psi_3(\|\omega\|_p^p)$ in place of $\|\omega\|_p^p$, with ψ_2 and ψ_3 defined in Appendix A, where we set $a = 1$ and $p = 0.5$. Both the experimental settings and the 20 synthetic problem-data of A_i, b_i $(i = 1, 2, 3)$ are identical to the ones used in the preceding section. The obtained results are summarized in Tables 5, 6, 7, 8.

Similar to the remarks in the preceding sections, we observe that taking into account the quality of the solutions obtained as reflected by the validation errors and sparsity, together with the running times and success rates of the algorithm, we observe that Algorithm 1 with the smoothing function ϕ_1 has a consistent good performance among all the functions considered.

5.3 Comparisons with Bayesian optimization

As mentioned in the introduction, two popular methods for dealing with the hyperparameter learning problem include the grid search method and Bayesian optimization. For practical purposes, however, grid search algorithm is not a viable approach due to the necessity of solving the lower level problem (1) for many values of the hyperpa-

Table 5 Averaged results for $(n, m_2, m_1, p) = (500, 1000, 1000, 0.5)$ using ψ_2

i	val	Test	bkkt	Sparsity(%)	Time(s)	ssn.iter	iter	Success(%)
1	1.39e-03	1.39e-03	**5.76e-03**	84.7	276.7 (4.9)	180.9	**31.2**	**100**
2	1.36e-03	1.37e-03	6.66e-03	**84.7**	284.4 (4.7)	181.7	32.8	**100**
3	**1.35e-03**	**1.36e-03**	6.22e-03	84.1	259.4 (5.2)	180.7	32.9	95
4	3.01e-03	2.95e-03	6.96e-03	75.8	262.6 (6.4)	183.3	32.5	**100**
6	3.10e-03	3.05e-03	5.88e-03	76.0	**244.1 (4.4)**	**180.3**	32.1	95

Table 6 Averaged results for $(n, m_2, m_1, p) = (500, 300, 300, 0.5)$ using ψ_2

i	val	Test	bkkt	Sparsity(%)	Time(s)	ssn.iter	iter	Success(%)
1	**2.06e-03**	2.17e-03	5.78e-03	**80.7**	200.4 (4.9)	**112.5**	**32.8**	**95**
2	**2.06e-03**	2.19e-03	6.00e-03	73.0	211.8 (4.7)	129.2	33.2	85
3	2.11e-03	2.17e-03	5.90e-03	74.9	**196.3** (5.2)	119.6	33.2	85
4	**2.06e-03**	**2.15e-03**	**5.68e-03**	79.2	237.3 (6.4)	146.1	37.5	**95**
6	2.58e-03	2.58e-03	5.90e-03	75.2	217.2 (4.4)	120.5	36.4	90

Table 7 Averaged results for $(n, m_2, m_1, p) = (500, 1000, 1000, 0.5)$ using ψ_3

i	val	Test	bkkt	Sparsity(%)	Time(s)	ssn.iter	iter	Success(%)
1	1.39e-03	1.39e-03	**5.76e-03**	84.7	193.1 (4.9)	180.9	**31.2**	**100**
2	1.37e-03	1.38e-03	6.68e-03	**84.8**	193.9 (4.7)	181.6	31.7	95
3	**1.35e-03**	**1.36e-03**	6.38e-03	83.5	195.0 (5.2)	180.8	33.0	**100**
4	3.01e-03	2.95e-03	6.96e-03	75.8	**192.4** (6.4)	183.3	32.5	**100**
6	3.02e-03	2.96e-03	6.05e-03	76.4	214.5 (5.4)	**180.4**	32.4	**100**

Table 8 Averaged results for $(n, m_2, m_1, p) = (500, 300, 300, 0.5)$ using ψ_3

i	val	Test	bkkt	Sparsity(%)	time(s)	ssn.iter	iter	Success(%)
1	2.11e-03	2.25e-03	**5.62e-03**	**80.8**	227.6 (4.9)	**114.5**	33.5	**100**
2	**2.06e-03**	2.19e-03	6.00e-03	73.0	227.8 (4.7)	129.2	**33.2**	85
3	2.11e-03	2.20e-03	5.88e-03	75.2	231.4 (5.2)	119.5	35.2	90
4	**2.06e-03**	**2.15e-03**	5.68e-03	79.2	**220.9** (6.4)	146.1	37.5	95
6	2.58e-03	2.58e-03	5.90e-03	75.2	231.5 (4.4)	120.5	36.4	90

Table 9 Averaged results for $(n, m_2, m_1) = (250, 500, 500)$ using ψ_1

p	Method	val	Test	Sparsity(%)	Time(s)
1	Algorithm 1 w/ ϕ_1	**4.15e-03**	**4.12e-03**	41.7	**39.4**
	Bayesian optimization	7.10e+00	6.79e+00	**50.4**	314.8
0.5	Algorithm 1 w/ ϕ_1	**6.60e-04**	**6.67e-04**	**72.7**	**39.9**
	Bayesian optimization	4.47e+01	4.47e+01	19.1	160.8

rameters $(\lambda_1, \ldots, \lambda_r)$, as was also demonstrated in [21]. Hence, we only compare our approach with Bayesian optimization. As shown in Table 9, our approach needed only roughly 25% of the time required by Bayesian optimization for $p = 0.5$, while still achieving low validation and test errors, as well as sparse models. For $p = 1$, while the Bayesian optimization strategy attained sparser solutions, it required almost eight times more computing time, and the validation and test errors are significantly larger than the one obtained by our approach.

6 Conclusion

This paper considers a class of nonsmooth, possibly nonconvex and non-Lipschitz regularizers for the best hyperparameter selection problem using a bilevel programming strategy. The class of regularizers we consider subsumes the traditional ℓ_p regularizer, which is the focus of the earlier work [21]. We propose new bilevel KKT conditions which are tighter than the SBKKT conditions proposed in [21]. These are necessary conditions for the original bilevel problem (4) when $p = 1$, and are necessary conditions for the relaxed problem (7) when $p < 1$. The convergence analysis of the smoothing algorithm presented in this paper is unified, in the sense that it is not limited to the chosen smoothing function, unlike the previous work [21] where the analysis is centered on the selected smoothing function. Finally, we proved our main convergence result under a milder constraint qualification. More precisely, we only assumed the Mangasarian-Fromovitz constraint qualification (MFCQ) for our convergence analysis, which is weaker than the linearly independent constraint qualification (LICQ) used in [21]. For our numerical simulations, we compared the performance of six smoothing functions in solving the bilevel programming problem using different regularizers. Theoretically, we can use these smoothing functions for all the regularizers considered as their corresponding density functions satisfy Assumption (B). On the other hand, our practical experience revealed that the smoothing function ϕ_1 provides the best performance when taking into account the validation and test errors of the resulting model, as well as the sparsity of the solution and running time of the algorithm. Interestingly, the function ϕ_1 is the closest approximation to the regularizer R_1 among all the smoothing functions, as proved in Appendix B.

Appendix A Penalty functions that satisfy Assumption (A)

We consider four penalty functions as follows:

$$\psi_1(t) = t, \quad \psi_2(t) = \log(1 + at), \quad \psi_3(t) = \frac{at}{1 + at}, \quad \psi_4(t) = \frac{-1}{1 + at},$$

where a is positive number. In particular,

(1) ψ_1 is a soft-thresholding penalty function [14, 25]. We have $\psi_1'(t) = 1$ and $\psi_1''(t) = 0$. Hence, it satisfies Assumption (A).

(2) ψ_2 is a logistic penalty function [19]. We have

$$\psi_2'(t) = \frac{a}{1 + at}, \quad \psi_2''(t) = -\frac{a^2}{(1 + at)^2},$$

which implies that $0 < \lim_{t\to 0} \psi_2'(t) = a$ and $|\psi_2''(t)| \leq a^2$. Hence, it satisfies Assumption (A).

(3) ψ_3 is fraction penalty function [9, 19]. We have

$$\psi_3'(t) = \frac{a}{(1+at)^2}, \quad \psi_3''(t) = -\frac{2a^2}{(1+at)^3},$$

which implies that $0 < \lim_{t\to 0} \psi_3'(t) = a$ and $|\psi_3''(t)| \leq 2a^2$. Hence, it satisfies Assumption (A).

(4) For function ψ_4, we have

$$\psi_4'(t) = \frac{a}{(1+at)^2}, \quad \psi_4''(t) = -\frac{2a^2}{(1+at)^3},$$

which implies that $0 < \lim_{t\to 0} \psi_4'(t) = a$ and $|\psi_4''(t)| \leq 2a^2$. Hence, it satisfies Assumption (A).

Appendix B Examples of smoothing functions

A key aspect in successful numerical implementations of a smoothing algorithm is the choice of the approximating functions. Here, we enumerate six smoothing functions that we will use in our numerical simulations.

There are many density functions commonly used and called kernel functions in statistics (see also [7]). Some density functions satisfying (9) are given as follows.

$$\rho_1(x) \coloneqq \begin{cases} \frac{35}{32}(1-x^2)^3 & \text{if } |x| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$\rho_2(x) \coloneqq \begin{cases} \frac{15}{16}(1-x^2)^2 & \text{if } |x| \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$\rho_3(x) \coloneqq \begin{cases} \frac{3}{4}(1-x^2) & \text{if } |x| \leq 1, \\ 0 & \text{otherwise,} \end{cases}$$

$$\rho_4(x) \coloneqq \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad \forall x \in \Re.$$

$$\rho_5(x) \coloneqq \frac{e^{-x}}{(1+e^{-x})^2}.$$

$$\rho_6(x) \coloneqq \frac{1}{(x^2+1)^{\frac{3}{2}}}.$$

Following the discussion in Sect. 2.2, the corresponding smoothing functions of $|x|$ are given as follows:

$$\phi_1(\mu, x) \coloneqq \begin{cases} -\frac{5x^8}{128\mu^7} + \frac{7x^6}{32\mu^5} - \frac{35x^4}{64\mu^3} + \frac{35x^2}{32\mu} + \frac{35\mu}{128} & \text{if } |x| \le \mu, \\ |x| & \text{if } |x| > \mu. \end{cases}$$

$$\phi_2(\mu, x) \coloneqq \begin{cases} \frac{x^6}{16\mu^5} - \frac{5x^4}{16\mu^3} + \frac{15x^2}{16\mu} + \frac{5\mu}{16} & \text{if } |x| \le \mu, \\ |x| & \text{if } |x| > \mu. \end{cases}$$

$$\phi_3(\mu, x) \coloneqq \begin{cases} -\frac{x^4}{8\mu^3} + \frac{3x^2}{4\mu} + \frac{3\mu}{8} & \text{if } |x| \le \mu, \\ |x| & \text{if } |x| > \mu. \end{cases}$$

$$\phi_4(\mu, x) \coloneqq x\,\mathrm{erf}\left(\frac{x}{\sqrt{2}\mu}\right) + \sqrt{\frac{2}{\pi}}\mu e^{-\frac{x^2}{2\mu^2}},$$

$$\phi_5(\mu, x) \coloneqq \mu\left[\log\left(1 + e^{-\frac{x}{\mu}}\right) + \log\left(1 + e^{\frac{x}{\mu}}\right)\right].$$

$$\phi_6(\mu, x) \coloneqq \sqrt{\mu^2 + x^2}.$$

Here, the error function is defined by

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

The graphs of $|x|$ and $\phi_i(\mu, x)$, $i = 1, 2, \ldots, 6$ with $\mu = 0.25$ are illustrated in Fig. 1. From the graphs, we infer the following inequality relating the smoothing functions:

$$\begin{cases} |x| \le \phi_1(\mu, x) \le \phi_2(\mu, x) \le \phi_3(\mu, x) \le \phi_4(\mu, x) \le \phi_5(\mu, x), \phi_6(\mu, x). \\ \text{there exists } \alpha > 0 \text{ such that } \phi_6(\mu, x) \le \phi_5(\mu, x) \text{ for all } x \in [-\alpha, \alpha]. \end{cases}$$

It is not difficult to show that the relation $|x| \le \phi_1(\mu, x) \le \phi_2(\mu, x) \le \phi_3(\mu, x)$, while the proof of the relation $\phi_3(\mu, x) \le \phi_4(\mu, x) \le \phi_5(\mu, x)$ can be found in [23]. Using the same proof technique in [23], one can easily achieve the remaining inequalities. On the other hand, the graphs of the corresponding smoothing functions for $|x|^p$ where $p \in (0, 1]$ is shown in Figs. 1 and 2. We note that the smooth approximation ϕ_6 is the function used in [21] for their smoothing algorithm for (4) with $R_1(\omega) \coloneqq \sum_{i=1}^n |\omega_i|^p$ $(0 < p \le 1)$.

In this paper, we consider the six functions above and determine which approximation is the best suitable in solving (4) with $R_1(\omega)$ satisfying Assumption (A). It is easy to check that the six density functions as above satisfy Assumptions (B1)-(B2). Condition (B3), on the other hand, holds by choosing $c = 4$, and $r = 2$. Indeed,

$$1 - \frac{4}{4 + S^2} \le \sqrt{1 - \frac{4}{4 + S^2}} = 2\int_0^S \rho_3(s)\, ds \le 2\int_0^S \rho_i(s)\, ds \quad \forall i = 1, \ldots, 6.$$

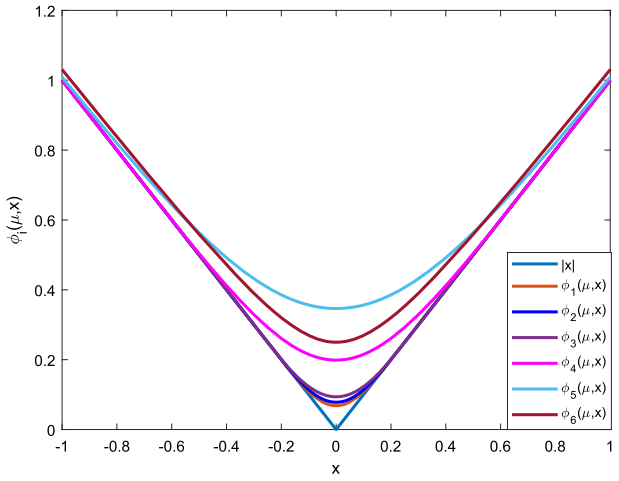

Fig. 1 Graph of $|x|$ and $\phi_i(\mu, x)$, $i = 1, 2, \ldots, 6$ with $\mu = 0.25$

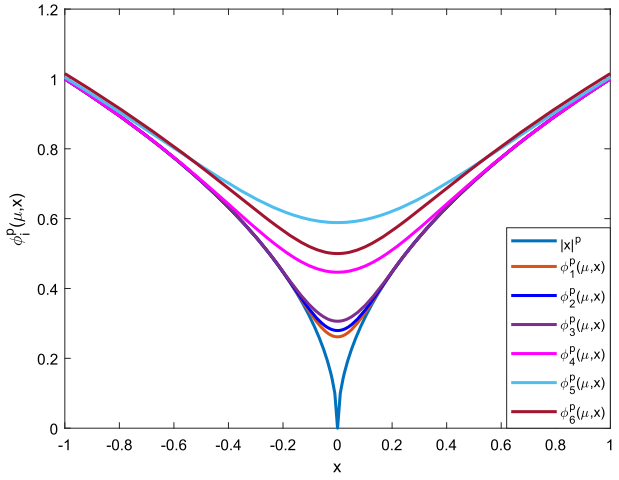

Fig. 2 Graph of $|x|^p$ and $(\phi(\mu, x))^p$, $i = 1, 2, \ldots, 6$ with $\mu = 0.25$ and $p = 0.5$

According to Assumption (B4), only the functions ρ_4, ρ_5 and ρ_6 can be used (theoretically) for the case $p = 1$.

Appendix C Proof of Lemma 4.5

In this appendix, we give a proof of Lemma 4.5.

Proof By Assumption (C), we know that there exists $\bar{d} \in \Re^{n+r}$ such that (62) and (63) hold. Meanwhile, we have from the formula of Ψ_j that

$$\nabla_{(\omega,\lambda)}\Psi_j(\omega,\lambda) = \begin{cases} \nabla_{(\omega,\lambda)}\Phi_j(\omega,\lambda) & \text{if } j \notin I(\omega^*) \\ \begin{bmatrix} e_j \\ 0_r \end{bmatrix} & \text{if } j \in I(\omega^*), \end{cases} \tag{C1}$$

where e_j is the jth standard unit vector in $\Re^n$ and 0_r denotes the zero vector in $\Re^r$. It is then clear from (C1) and (62) that $\bar{d}_j = 0$ for all $j \in I(\omega^*)$. Consequently, letting $d \in \Re^{n-|I(\omega^*)|+r}$ be the vector $d := (\bar{d})_{j\notin I(\omega^*)}$, it follows from (62) and (63) that Eqs. (64) and (65) hold.

It remains to show that $\{\nabla_{(\tilde{\omega},\lambda)}\Phi_j(\omega^*,\lambda^*)\}_{j\notin I(\omega^*)}$ is linearly independent. To this end, note first that we have from Assumption (C) the linear independence of $\{\nabla_{(\omega,\lambda)}\Psi_j(\omega^*,\lambda^*)\}_{j=1}^n$, that is, the matrix

$$M := \left[\left(\nabla_{(\omega,\lambda)}\Psi_j(\omega^*,\lambda^*)\right)_{(j\notin I(\omega^*))}, \left(\nabla_{(\omega,\lambda)}\Psi_j(\omega^*,\lambda^*)\right)_{(j\in I(\omega^*))}\right] \in \Re^{(n+r)\times n}$$

has full column rank. Using Eq. (C1) and switching the rows of M so that the first $|I(\omega^*)|$ rows correspond to the index set $I(\omega^*)$, we have that the matrix

$$\begin{bmatrix} (\nabla_{\breve{\omega}}\Phi_j(\omega^*,\lambda^*))_{j\notin I(\omega^*)} & E_{|I(\omega^*)|} \\ (\nabla_{(\tilde{\omega},\lambda)}\Phi_j(\omega^*,\lambda^*))_{j\notin I(\omega^*)} & O_{(n-|I(\omega^*)|+r)\times|I(\omega^*)|} \end{bmatrix}$$

has full column rank, where $\breve{\omega} := (\omega_j)_{j\in I(\omega^*)}$, E_s denotes the identity matrix of order s, and $O_{s\times t}$ is the zero matrix of size $s \times t$. Since the upper and lower right blocks of the above matrix are the identity matrix and zero matrix, respectively, a series of elementary column operations leads us to conclude that

$$\begin{bmatrix} O_{|I(\omega^*)|\times(n-|I(\omega^*)|)} & E_{|I(\omega^*)|} \\ (\nabla_{(\tilde{\omega},\lambda)}\Phi_j(\omega^*,\lambda^*))_{j\notin I(\omega^*)} & O_{(n-|I(\omega^*)|+r)\times|I(\omega^*)|} \end{bmatrix}$$

also has full column rank. As a consequence, $\{\nabla_{(\tilde{\omega},\lambda)}\Phi_j(\omega^*,\lambda^*)\}_{j\notin I(\omega^*)}$ is linearly independent, as desired. □

Acknowledgements The authors are very grateful to the anonymous referees for their valuable comments and suggestions which have significantly improved the quality of this paper. Part of this work was conducted while J. H. Alcantara was a postdoctoral fellow at National Taiwan Normal University, and C. T. Nguyen was a research assistant at National Taiwan Normal University. T. Okuno was partially supported by JSPS KAKENHI Nos. 20K19748 and 20H04145; A. Takeda was partially supported by JSPS KAKENHI No. 23H03351; J.-S. Chen was partially supported by NSTC 112-2115-M-003-012-MY2.

Funding National Science and Technology Council, and Japan Society for the Promotion of Science.

References

1. Alcantara, J.H., Lee, C.H., Nguyen, C.T., Chang, Y.L., Chen, J.-S.: On construction of new NCP functions. Oper. Res. Lett. **48**(2), 115–121 (2020)
2. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**(10), 281–305 (2012)
3. Bennett, K.P., Hu, J., Ji, X., Kunapuli, G., Pang, J-S.: Model selection via bilevel optimization. In The 2006 IEEE International Joint Conference on Neural Network Proceedings, pp. 1922–1929 (2006)
4. Bennett, K.P., Hu, J., Ji, X., Kunapuli, G., Pang, J-S.: Bilevel optimization and machine learning. In: Computational Intelligence: Research Frontiers. WCCI 2008. Lecture Notes in Computer Science, vol. 5050. Springer, Berlin, Heidelberg (2008)
5. Bracken, J., McGill, J.: Mathematical programs with optimization problems in the constraints. Oper. Res. **21**(1), 37–44 (1973)
6. Chen, X.: Smoothing methods for nonsmooth, nonconvex minimization. Math. Program. **134**(1), 71–99 (2012)
7. Chen, C., Mangasarian, O.L.: A class of smoothing functions for nonlinear and mixed complementarity problems. Comput. Optim. Appl. **5**(2), 97–138 (1996)
8. 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**(5), 2832–2852 (2010)
9. Chen, X., Zhou, W.: Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization. SIAM J. Imaging Sci. **3**(4), 765–790 (2010)
10. Colson, B., Marcotte, P., Savard, G.: An overview of bilevel optimization. Ann. Oper. Res. **153**, 234–256 (2007)
11. Dempe, S.: Annotated bibliography on bilevel programming and mathematical programs with equilibrium constraints. Optimization **52**(3), 333–359 (2003)
12. De Luca, T., Facchinei, F., Kanzow, C.: A semismooth equation approach to the solution of nonlinear complementarity problems. Math. Program. **75**(3), 407–439 (1996)
13. Facchinei, F., Pang, J.-S.: Finite Dimensional Variational Inequalities and Complementarity Problems, vol. 1 and 2. Springer, New York (2003)
14. Huang, J., Horowitz, J.L., Ma, S.: Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Statist. **36**(2), 587–613 (2008)
15. Kruger, A.Y.: On Fréchet subdifferentials. J. Math. Sci. **116**(3), 3325–3358 (2003)
16. Kunisch, K., Pock, T.: A bilevel optimization approach for parameter learning in variational models. SIAM J. Imaging Sci. **6**(2), 938–983 (2013)
17. Moore, G.M., Bergeron, C., Bennett, K.P.: Nonsmooth bilevel programming for hyperparameter selection. In IEEE International Conference on Data Mining Workshops, pp. 374–381 (2009)
18. Moore, G.M., Bergeron, C., Bennett, K.P.: Model selection for primal SVM. Mach. Learn. **85**, 175–208 (2011)
19. Nikolova, M., Ng, M.K., Zhang, S., Ching, W.-K.: Efficient reconstruction of piecewise constant images using nonsmooth nonconvex minimization. SIAM J. Imaging Sci. **1**(1), 2–25 (2008)
20. Ochs, P., Ranftl, R., Brox, T., Pock, T.: Techniques for gradient-based bilevel optimization with non-smooth lower level problems. J. Math. Imaging Vis. **56**(2), 175–194 (2016)
21. Okuno, T., Takeda, A., Kawana, A., Watanabe, M.: On ℓ_p-hyperparameter learning via bilevel nonsmooth optimization. J. Mach. Learn. Res. **22**(245), 1–47 (2021)
22. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis, vol. 317. Springer, Cham (2009)
23. Saheya, B., Nguyen, C.T., Chen, J.-S.: Neural network based on systematically generated smoothing functions for absolute value equation. J. Appl. Math. Comput. **61**, 533–558 (2019)
24. Sinha, A., Malo, P., Deb, K.: A review on bilevel optimization: from classical to evolutionary approaches and applications. IEEE Trans. Evol. Comput. **22**(2), 276–295 (2018)
25. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Statist. Soc. Ser. B. **58**(1), 267–288 (1996)
26. Tso, W.W., Burnak, B., Pistikopoulos, E.N.: HY-POP: hyperparameter optimization of machine learning models through parametric programming. Comput. Chem. Eng. **139**, 106902 (2020)
27. Wang, Z., Hutter, F., Zoghi, M., Matheson, D., de Freitas, N.: Bayesian optimization in a billion dimensions via random embeddings. J. Artif. Intell. Res. **55**, 361–387 (2016)