BILEVEL OPTIMIZATION WITH CONVEX MAJORANT APPROACH FOR TRAINING SPARSE NEURAL NETWORKS

XINGBANG CUI* AND XIAOJUN CHEN†

Abstract. This paper proposes a convex majorant approach for training sparse neural networks by bilevel optimization where the upper level problem minimizes a smooth nonconvex function while the lower level problem minimizes a smooth nonconvex function with a nonsmooth convex group sparse regularizer over a box set for fixed sparse regularization hyperparameters. The convex majorant function approximates the objective function of the lower level problem. We establish the relationship between the original bilevel optimization and the bilevel optimization with the convex majorant approach regarding global and local minimizers. Moreover, we use a smoothing function to approximate the convex majorant function, and derive the convergence of global minimizers to those of the corresponding nonsmooth bilevel problems with smoothing parameter converging to zero. A smoothing implicit function method is proposed to solve the smooth approximate bilevel optimization problem. Some numerical experiments including the tests on the data from machine learning repository show that the convex majorant approach performs better than the widely used Grid Search method, Random Search method and Bayesian optimization method.

Key words. Bilevel optimization, sparse regularization hyperparameter, convex majorant, smoothing method

AMS subject classifications. 90C30, 90C33, 90C90

1. Introduction.

In this paper, we consider bilevel optimization for tuning hyperparameters of L-layer sparse feed-forward neural networks with L being a positive integer. We divide the given data $\{(X^i, Y^i) \in \mathbb{R}^n \times \mathbb{R}^m, i = 1, \cdots, N\}$ into a training set $\{(X^i, Y^i) \in \mathbb{R}^n \times \mathbb{R}^m, i = 1, \cdots, N_{tr}\}$ and a validation set $\{(X^i, Y^i) \in \mathbb{R}^n \times \mathbb{R}^m, i = N_{tr}+1, \cdots, N\}$, where $N = N_{tr} + N_{va}$. Let $W_\ell \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$, $b^\ell \in \mathbb{R}^{n_\ell}$, $\alpha^\ell \in \mathbb{R}^{n_{\ell-1}}$ for $\ell = 1, \cdots, L$, where $n_0 = n$ and $n_L = m$. The bilevel optimization involves the following functions:

$$
\begin{aligned}
F(u) &= \frac{1}{N_{va}} \sum_{i=N_{tr}+1}^{N} \|W_L \sigma(\cdots \sigma(W_1 X^i + b^1) \cdots) + b^L - Y^i\|^2, \\
H(u) &= \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \|W_L \sigma(\cdots \sigma(W_1 X^i + b^1) \cdots) + b^L - Y^i\|^2, \\
Q(w; \lambda) &= \sum_{\ell=1}^{L} \sum_{j=1}^{n_{\ell-1}} \alpha_j^\ell \|(W_\ell)_{\cdot j}\|,
\end{aligned}
$$

where $w = ((W_1)_{\cdot 1}^\top, \cdots, (W_1)_{\cdot n}^\top, \cdots, (W_L)_{\cdot n_{L-1}}^\top)^\top \in \mathbb{R}^p$, $b = ((b^1)^\top, \cdots, (b^L)^\top)^\top \in \mathbb{R}^s$, $u = (w^\top, b^\top)^\top \in \mathbb{R}^q$, $\lambda = ((\alpha^1)^\top, \cdots, (\alpha^L)^\top)^\top \in \mathbb{R}^r$ with $p = \sum_{\ell=1}^{L} n_{\ell-1} n_\ell$,

*School of Mathematics and Statistics, Shaanxi Normal University, Xi'an, China; Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (cuixingbangsdu@126.com). The author is supported by Hong Kong Polytechnic University Post-doctoral Fellowship.

†Department of Applied Mathematics, The Hong Kong Polytechnic University, Kowloon, Hong Kong (maxjchen@polyu.edu.hk). The author is supported by CAS-Croucher Funding Scheme for Joint Laboratories "CAS AMSS-PolyU Joint Laboratory of Applied Mathematics", Hong Kong Research Council Grant PolyU15300123.

$s = \sum_{\ell=1}^{L} n_\ell$, $r = \sum_{\ell=1}^{L} n_{\ell-1}$, $q = p + s$, and $\sigma : \mathbb{R} \to \mathbb{R}$ is a twice continuously differentiable activation function. Here $\|\cdot\|$ denotes the Euclidean norm and $\sigma(u) := (\sigma(u_1), \cdots, \sigma(u_q))^\top$ for $u \in \mathbb{R}^q$. The functions F and H are smooth nonconvex, while the function $Q(\cdot;\lambda)$ is nonsmooth convex for any fixed sparse regularization hyperparameter $\lambda \geq 0$.

We focus on the following bilevel optimization problem:

$$\min_{\lambda,u} \quad F(u) \qquad \text{s.t.} \quad \lambda \geq 0, \quad u \in S(\lambda), \tag{1.1}$$

where $S(\lambda)$ is the solution set of the lower level problem parameterized by λ:

$$\min_{u} \quad H(u) + Q(w;\lambda) \qquad \text{s.t.} \quad u \in \Omega. \tag{1.2}$$

Here $\Omega := [\underline{u}, \overline{u}] \subseteq \mathbb{R}^q$ is a compact box set with $\underline{u} < \overline{u}$.

The feed-forward neural network is an important kind of neural networks. According to the universal approximation theorem [2, 11, 23, 37], a feed-forward neural network with a single hidden layer can approximate any continuous function to any desired accuracy as long as the activation function is not polynomial and there are sufficient hidden nodes. In many applications, the sparse neural networks have advantages for saving storage cost and computation cost [14, 40, 42]. Moreover, sparse neural networks have simpler structures and fewer parameters compared to the fully connected feed-forward neural networks, which can avoid data overfitting problems [13, 40].

The sparse regularization term $Q(w;\lambda)$ in (1.2) helps training the neural network with weight matrices W_ℓ, $\ell = 1, \cdots, L$, that have few nonzero columns. This term is based on group sparse regularization which has been extensively employed in designing compact neural networks [14, 20, 38, 40, 42, 43]. Via this regularization technique, some columns of the weight matrices are forced to be zero simultaneously. Intuitively, this means that some connections of two neurons of two adjacent different levels are eliminated, which results in sparse neural networks (see [14, Figure 1] for an example).

There is no doubt that the selection of hyperparamters is crucial in constructing the sparse neural networks (see [38, Fig. 4]). In most related papers, the hyperparameters are set via the Grid Search method [14, 38], which may not yield an optimal selection in general. A lot of evidences show that the bilevel optimization model is efficient and promising for hyperparameter selection in machine learning [15, 18, 28, 34, 35]. Hence, in this paper we study the nonsmooth nonconvex bilevel optimization (1.1) for the selection of optimal hyperparameters.

Since lower level problem (1.2) is nonsmooth and nonconvex, it is extremely challenging to solve problem (1.1). One approach for bilevel optimization problems is to reformulate the bilevel optimization problem as a single level optimization problem with optimality conditions of the lower level optimization problem as constraints (see [12, Chapter 12]). However, it has been shown in [32, Example 1] and [33, Example 1.1] that when the lower level optimization problem is nonconvex, any optimal solution of the bilevel optimization problem may not even be a stationary point of the new single level optimization problem. Another method addressing nonconvex lower level problems is to use the value function, where the bilevel program is reformulated as a single level optimization problem via the value function, which can be solved via some existing algorithms for the nonconvex and nonsmooth optimization problems, see [24, 27, 41]. There are some other methods including the bounding algorithm [33] and gradient method [29, 31]. Li and Yang [25] constructed a piecewise convex relaxation of the nonconvex lower level problem by adding a quadratic term. However, all of these works tend to be complicated and impractical for large-scale bilevel optimization problems. Moreover, the objective functions of the lower level problems in [24, 25, 27, 29, 31, 41] are assumed to be smooth, while problem (1.2) is nonsmooth. In [1, 34], the authors directly reformulated the bilevel optimization problems with nonsmooth and nonconvex lower level problems via optimality conditions of the lower level optimization problems, and employed smoothing methods to solve the resulting single level problems. In [30], the authors proposed a single-loop gradient-based algorithm by the Moreau envelope-based reformulation. However, as we have stated above, the equivalence between the original bilevel problem and the single level problem may fail due to the nonconvexity of the lower level problem.

We construct the following strongly convex majorant function with fixed $\lambda \in \mathbb{R}^r_+,\ z \in \Omega$ and $\gamma > 0$:

$$G(u;\lambda,z) := H(z) + \nabla H(z)^\top (u-z) + \frac{\gamma}{2}\|u-z\|^2 + Q(w;\lambda)$$

for $u \in \Omega$. Since H is twice continuously differentiable and Ω is a compact set, we can choose γ such that $\|\nabla^2 H(\cdot)\| \le \gamma$ over Ω. The choice of γ guarantees that given any fixed $\lambda \in \mathbb{R}^r_+,\ z \in \Omega$,

$$G(u;\lambda,z) \ge H(u) + Q(w;\lambda)$$

for $u \in \Omega$. Now we consider the following problem:

$$\min_{\lambda,z,u} \quad F(u) \qquad \text{s.t.} \quad \lambda \ge 0,\ z \in \Omega,\ u = u(\lambda,z), \tag{1.3}$$

where $u(\lambda,z)$ is the unique solution of the following lower level problem:

$$\min_{u} \quad G(u;\lambda,z) \qquad \text{s.t.} \quad u \in \Omega. \tag{1.4}$$

The convex majorant approach (1.4) is based on the second order Taylor expansion, which is different from the piecewise convex relaxation in [25]. Note that although the objective function $G(\cdot;\lambda,z)$ of problem (1.4) is nonsmooth, it can have a smoothing function with the gradient consistency (see [7] for the definition). In particular, we propose a strongly convex smoothing function

$$G_\mu(u;\lambda,z) := H(z) + \nabla H(z)^\top (u-z) + \frac{\gamma}{2}\|u-z\|^2 + Q_\mu(w;\lambda) \tag{1.5}$$

for $u \in \Omega$, where $\mu > 0$ is an arbitrarily small real number and

$$Q_\mu(w;\lambda) = \sum_{\ell=1}^{L}\sum_{j=1}^{n_{\ell-1}} \alpha_j^\ell \sqrt{\|(W_\ell)_{\cdot j}\|^2 + \mu}. \tag{1.6}$$

For any fixed λ and z, we have

$$\lim_{u\to\tilde u,\mu\downarrow 0} G_\mu(u;\lambda,z) = G(\tilde u;\lambda,z) \quad \text{and} \quad \text{conv}\left\{\lim_{u\to\tilde u,\mu\downarrow 0} \nabla G_\mu(u;\lambda,z)\right\} = \partial G(\tilde u;\lambda,z), \tag{1.7}$$

where *conv* denotes the convex hull and $\partial G(\tilde u;\lambda,z)$ is the Clarke subgradient of G at $\tilde u$ [9].

The contributions of this paper are summarized as follows.

(i) We propose a convex majorant approach (1.3) for problem (1.1) by replacing the objective function of the lower level problem (1.2) with a convex majorant function $G(\cdot;\lambda,z)$. We then derive the equivalence between the global and local optimal solutions of problem (1.1) and problem (1.3) under the assumptions on feasibility.

(ii) We use the smoothing function $G_\mu(\cdot;\lambda,z)$ to define a smooth approximation problem of problem (1.3). We prove that any accumulation point of global optimal solutions of the smooth approximation problems is the global optimal solution of problem (1.3) as the smoothing parameter μ goes to zero.

(iii) We propose a smoothing implicit function method to solve the smooth approximate problem of problem (1.3), and derive the convergence of the method to a Clarke stationary point of the smooth approximate problem.

This paper is organized as follows. In Section 2, we establish the relationship between problem (1.1) and problem (1.3) regarding global and local optimal solutions. We study the smooth approximation problem of problem (1.3) in Section 3. In Section 4, we propose a smoothing implicit function method. Numerical results are presented in Section 5. Finally, concluding remarks are drawn in Section 6.

Notation: Denote a closed ball in $\mathbb{R}^q$ with center $u \in \mathbb{R}^q$ and radius $\delta > 0$ by $B(u,\delta)$. Let I_q be the identity matrix in $\mathbb{R}^{q\times q}$, and $e_q \in \mathbb{R}^q$ be the vector with all elements equal to 1. Given function $f:\mathbb{R}^m \to \mathbb{R}^n$, $Jf(x) \in \mathbb{R}^{n\times m}$ denotes the Jacobian of f at $x \in \mathbb{R}^m$. Let $diag(v) \in \mathbb{R}^{q\times q}$ be the square matrix with elements of $v \in \mathbb{R}^q$ on the diagonal. Given a nonempty closed convex set $S \subset \mathbb{R}^q$, $N_S(x) := \{v : \langle v, y-x\rangle \le 0,\ \forall\, y \in S\}$ denotes the normal cone of S at x.

2. Relationship between problems (1.1) and (1.3). In this section, we investigate the relationship between problem (1.1) and problem (1.3). We assume that the solution sets of problems (1.1) and (1.3) are nonempty. The following lemma indicates the relationship in regard to the feasibility. As for problem (1.1), $(\tilde{\lambda},\tilde{u})$ is a feasible point of problem (1.1) if $\tilde{\lambda} \ge 0$, $\tilde{u} \in \Omega$, and $\tilde{u}$ solves lower level problem (1.2) globally for the fixed hyperparameter $\tilde{\lambda}$. The feasibility of problem (1.3) can be defined similarly.

LEMMA 2.1. *If $(\tilde{\lambda},\tilde{u})$ is a feasible point of problem (1.1), then $(\tilde{\lambda},\tilde{u},\tilde{u})$ is a feasible point of problem (1.3).*

Proof. It suffices to prove that $G(\tilde{u};\tilde{\lambda},\tilde{u}) \le G(u;\tilde{\lambda},\tilde{u})$ for any $u \in \Omega$. Note that

$$G(\tilde{u};\tilde{\lambda},\tilde{u}) = H(\tilde{u}) + Q(\tilde{w};\tilde{\lambda}) \le H(u) + Q(w;\tilde{\lambda}) \le G(u;\tilde{\lambda},\tilde{u}),$$

since $\tilde{u} \in S(\tilde{\lambda})$. The conclusion is obvious. ∎

From Lemma 2.1, the following two theorems give some properties of global and local optimal solutions of problem (1.3) related to problem (1.1).

THEOREM 2.2. *Let $(\tilde{\lambda},\tilde{z},\tilde{u})$ be a global optimal solution of problem (1.3). Then the following statements hold.*

(i) $F(\tilde{u}) \le F(u)$, for any feasible point $u \in S(\lambda), \lambda \ge 0$ of problem (1.1).

(ii) If $(\tilde{\lambda},\tilde{u})$ is a feasible point for problem (1.1), then $(\tilde{\lambda},\tilde{u})$ is a global optimal solution of (1.1).

Proof. (i) According to Lemma 2.1, for any feasible point $u \in S(\lambda), \lambda \ge 0$ of problem (1.1), (λ,u,u) is a feasible point of problem (1.3). Since $(\tilde{\lambda},\tilde{z},\tilde{u})$ is a global optimal solution of problem (1.3), we have $F(\tilde{u}) \le F(u)$.

(ii) Assume by contradiction that $(\tilde{\lambda},\tilde{u})$ is not a global optimal solution of (1.1). Then there exists a feasible point (λ^*,u^*) of problem (1.1) such that $F(u^*) < F(\tilde{u})$. Due to Lemma 2.1, we know that (λ^*,u^*,u^*) is a feasible point of problem (1.3).

However, the fact that $F(u^*) < F(\tilde{u})$ contradicts the hypothesis that $(\tilde{\lambda}, \tilde{z}, \tilde{u})$ is a global optimal solution of (1.3). ∎

THEOREM 2.3. *Let $(\tilde{\lambda}, \tilde{u}, \tilde{u})$ be a local optimal solution of problem (1.3). If $(\tilde{\lambda}, \tilde{u})$ is a feasible point of problem (1.1), then $(\tilde{\lambda}, \tilde{u})$ is a local optimal solution of problem (1.1).*

Proof. Assume by contradiction that $(\tilde{\lambda}, \tilde{u})$ is not a local optimal solution of problem (1.1). Then there exists a sequence of feasible points (λ^k, u^k), $k = 1, 2, \cdots$, of problem (1.1) satisfying that

$$\lim_{k\to\infty} (\lambda^k, u^k) = (\tilde{\lambda}, \tilde{u}), \text{ and } F(u^k) < F(\tilde{u}),\ k = 1, 2, \cdots .$$

Based on Lemma 2.1, we know that (λ^k, u^k, u^k), $k = 1, 2, \cdots$, are feasible points of problem (1.3). Hence, for any neighborhood of $(\tilde{\lambda}, \tilde{u}, \tilde{u})$, we can find some (λ^k, u^k, u^k) in this neighborhood such that $F(u^k) < F(\tilde{u})$, which incurs a contradiction with the hypothesis that $(\tilde{\lambda}, \tilde{u}, \tilde{u})$ is a local optimal solution of problem (1.3). Thus we have proved that $(\tilde{\lambda}, \tilde{u})$ is a local optimal solution of problem (1.1). ∎

Now we give a property of global optimal solutions of problem (1.1) related to problem (1.3).

THEOREM 2.4. *Let $(\tilde{\lambda}, \tilde{u})$ be a global (or local) optimal solution of (1.1). Then $(\tilde{\lambda}, \tilde{u}, \tilde{u})$ is a global (or local) optimal solution of (1.3) on $S_1 := \{(\lambda, u, u) : u \in S(\lambda), \lambda \geq 0\}$.*

Proof. We first prove the conclusion corresponding to the global optimal solution. Due to Lemma 2.1, it is obvious that $(\tilde{\lambda}, \tilde{u}, \tilde{u})$ is a feasible point of problem (1.3). According to the definition of S_1, (λ^*, u^*) is a feasible point of problem (1.1) when $(\lambda^*, u^*, u^*) \in S_1$. Then we have $F(u^*) \geq F(\tilde{u})$ since $(\tilde{\lambda}, \tilde{u})$ is a global optimal solution of problem (1.1), which indicates that $(\tilde{\lambda}, \tilde{u}, \tilde{u})$ is a global optimal solution of problem (1.3) on S_1. The conclusion corresponding to the local optimal solution can be proved like the proof for Theorem 2.3, which is omitted here. ∎

In the following, we investigate properties of the solution function $u(\cdot, \cdot)$ of problem (1.4).

PROPOSITION 2.5. *The solution function $u : \mathbb{R}^r_+ \times \Omega \to \mathbb{R}^q$ is Lipschitz continuous with Lipschitz constant $\kappa := \max\{2, \frac{\sqrt{r}}{\gamma}\}$, i.e., for any $(\lambda^1, z^1), (\lambda^2, z^2) \in \mathbb{R}^r_+ \times \Omega$,*

$$\|u(\lambda^1, z^1) - u(\lambda^2, z^2)\| \leq \kappa(\|z^1 - z^2\| + \|\lambda^1 - \lambda^2\|). \tag{2.1}$$

Proof. Given $(\lambda^1, z^1), (\lambda^2, z^2) \in \mathbb{R}^r_+ \times \Omega$, denote $u^1 := u(\lambda^1, z^1)$ and $u^2 := u(\lambda^2, z^2)$. According to the first order optimality condition, we have

$$\left\langle \nabla H(z^i) + \gamma(u^i - z^i) + \xi^i, z - u^i \right\rangle \geq 0,\ \forall\ z \in \Omega,\ i = 1, 2,$$

where $\xi^1 = ((\zeta^1)^\top, 0^\top)^\top \in \mathbb{R}^q$ with $\zeta^1 \in \partial Q(w^1; \lambda^1)$ and $\xi^2 = ((\zeta^2)^\top, 0^\top)^\top \in \mathbb{R}^q$ with $\zeta^2 \in \partial Q(w^2; \lambda^2)$. By setting $z = u^2$ and $z = u^1$ in the above two inequalities respectively and combining them, we have

$$\left\langle \nabla H(z^1) - \nabla H(z^2) + \gamma(u^1 - u^2) - \gamma(z^1 - z^2) + \xi^1 - \xi^2, u^2 - u^1 \right\rangle \geq 0,$$

which is equivalent to

$$\left\langle \nabla H(z^1) - \nabla H(z^2) - \gamma(z^1 - z^2) + \xi^1 - \xi^2, u^2 - u^1 \right\rangle \geq \gamma\|u^1 - u^2\|^2.$$

We analyze the three terms on the left hand one by one. Since H is twice continuously differentiable and $\|\nabla^2 H(z)\| \le \gamma$ over compact Ω, we have

$$\left\langle \nabla H(z^1) - \nabla H(z^2), u^2 - u^1 \right\rangle \le \|\nabla H(z^1) - \nabla H(z^2)\|\|u^1 - u^2\| \le \gamma\|z^1 - z^2\|\|u^1 - u^2\|. \tag{2.2}$$

In addition, we also have

$$\left\langle -\gamma(z^1 - z^2), u^2 - u^1 \right\rangle \le \gamma\|z^1 - z^2\|\|u^1 - u^2\|. \tag{2.3}$$

Now we turn to the third term. Let $w^i = ((W_1^i)_{\cdot 1}^\top, \cdots, (W_1^i)_{\cdot n}^\top, \cdots, (W_L^i)_{\cdot n_{L-1}}^\top)^\top$ and $\lambda^i = (((\alpha^1)^i)^\top, \cdots, ((\alpha^L)^i)^\top)^\top$, $i = 1, 2$. It is not difficult to see that

$$\begin{aligned}\left\langle \xi^1 - \xi^2, u^2 - u^1 \right\rangle &= \left\langle \zeta^1 - \zeta^2, w^2 - w^1 \right\rangle \\ &= \sum_{\ell=1}^{L}\sum_{j=1}^{n_{\ell-1}} \left\langle (\alpha_j^\ell)^1 \zeta_{\ell,j}^1 - (\alpha_j^\ell)^2 \zeta_{\ell,j}^2, (W_\ell^2)_{\cdot j} - (W_\ell^1)_{\cdot j} \right\rangle,\end{aligned}$$

where $\zeta_{\ell,j}^i \in \partial\|\cdot\|((W_\ell^i)_{\cdot j})$, $i = 1, 2$. We can consider each item of the third term separately. For $1 \le j \le n_{\ell-1}$, we have

$$\begin{aligned}&\left\langle (\alpha_j^\ell)^1 \zeta_{\ell,j}^1 - (\alpha_j^\ell)^2 \zeta_{\ell,j}^2, (W_\ell^2)_{\cdot j} - (W_\ell^1)_{\cdot j} \right\rangle \\ =&\left\langle (\alpha_j^\ell)^1 \zeta_{\ell,j}^1 - (\alpha_j^\ell)^1 \zeta_{\ell,j}^2 + (\alpha_j^\ell)^1 \zeta_{\ell,j}^2 - (\alpha_j^\ell)^2 \zeta_{\ell,j}^2, (W_\ell^2)_{\cdot j} - (W_\ell^1)_{\cdot j} \right\rangle \\ =&(\alpha_j^\ell)^1 \left\langle \zeta_{\ell,j}^1 - \zeta_{\ell,j}^2, (W_\ell^2)_{\cdot j} - (W_\ell^1)_{\cdot j} \right\rangle + \left\langle ((\alpha_j^\ell)^1 - (\alpha_j^\ell)^2) \zeta_{\ell,j}^2, (W_\ell^2)_{\cdot j} - (W_\ell^1)_{\cdot j} \right\rangle \\ \le&|(\alpha_j^\ell)^1 - (\alpha_j^\ell)^2|\|\zeta_{\ell,j}^2\|\|(W_\ell^2)_{\cdot j} - (W_\ell^1)_{\cdot j}\| \\ \le&|(\alpha_j^\ell)^1 - (\alpha_j^\ell)^2|\|u^1 - u^2\|,\end{aligned} \tag{2.4}$$

where the first inequality is from the convexity of the Euclidean norm and the second inequality is from the fact that $\|\zeta_{\ell,j}^2\| \le 1$. Combining (2.2), (2.3) and (2.4), we have

$$\begin{aligned}\|u^1 - u^2\| &\le 2\|z^1 - z^2\| + \frac{1}{\gamma}\sum_{\ell=1}^{L}\sum_{j=1}^{n_{\ell-1}} |(\alpha_j^\ell)^1 - (\alpha_j^\ell)^2| \\ &\le \kappa(\|z^1 - z^2\| + \|\lambda^1 - \lambda^2\|),\end{aligned} \tag{2.5}$$

where $\kappa := \max\{2, \frac{\sqrt{r}}{\gamma}\}$. Hence (2.1) holds. □

3. Smooth approximation of problem (1.3).

The nonsmoothness of (1.3) comes from the group sparse regularization term Q in the objective function of its lower level problem (1.4). In this paper, we use the smoothing function Q_μ in (1.6) and G_μ in (1.5) as smoothing functions of Q and G, respectively, where $\mu > 0$ is the smoothing parameter. Properties of the continuity and differentiability of smoothing function Q_μ can be directly derived from some existing literature (see for example [36]), and readily extended to G_μ.

We consider the following smooth approximation of problem (1.3):

$$\min_{\lambda,z,u} \quad F(u) \qquad \text{s.t.} \quad \lambda \ge 0,\ z \in \Omega,\ u = u_\mu(\lambda, z), \tag{3.1}$$

where $u_\mu(\lambda, z)$ is the unique solution of the following lower level problem:

$$\min_{u} \quad G_\mu(u; \lambda, z) \qquad \text{s.t.} \quad u \in \Omega. \tag{3.2}$$

Obviously, $u(\lambda, z) = u_\mu(\lambda, z)$ when $\mu = 0$. Since for any fixed $\mu \geq 0$, $\lambda \geq 0$ and $z \in \Omega$, $G_\mu(\cdot; \lambda, z)$ is a strongly convex function and Ω is a convex compact set, $u_\mu(\cdot, \cdot)$ is the unique solution of (3.2). In the following, we investigate properties of the solution function $u_\mu(\cdot, \cdot)$ of problem (3.2) for $\mu > 0$.

PROPOSITION 3.1. *For any $\mu > 0$, the solution function $u_\mu : \mathbb{R}^r_+ \times \Omega \to \mathbb{R}^q$ is Lipschitz continuous with Lipschitz constant $\kappa := \max\{2, \frac{\sqrt{r}}{\gamma}\}$, which is independent of μ, i.e., for any $(\lambda^1, z^1), (\lambda^2, z^2) \in \mathbb{R}^r_+ \times \Omega$,*

$$\|u_\mu(\lambda^1, z^1) - u_\mu(\lambda^2, z^2)\| \leq \kappa(\|z^1 - z^2\| + \|\lambda^1 - \lambda^2\|). \tag{3.3}$$

Proof. The proof can be directly derived following the proof of Proposition 2.5 with

$$\zeta^i_{\ell,j} = \frac{(W^i_\ell)_{\cdot j}}{\sqrt{\|(W^i_\ell)_{\cdot j}\|^2 + \mu}},$$

and $\|\zeta^i_{\ell,j}\| \leq 1$. ∎

PROPOSITION 3.2. *For any $(\tilde{\lambda}, \tilde{z}, \tilde{\mu}) \in \mathbb{R}^r_+ \times \Omega \times [0, 1]$, we have*

$$\lim_{(\lambda, z, \mu) \to (\tilde{\lambda}, \tilde{z}, \tilde{\mu})} u_\mu(\lambda, z) = u_{\tilde{\mu}}(\tilde{\lambda}, \tilde{z}). \tag{3.4}$$

Proof. Since $G_\mu(u; \lambda, z)$ is continuous with respect to (λ, z, μ) and Ω is a compact set, we know that for the lower level problem (3.2), the solution set mapping denoted by $\hat{S} : \mathbb{R}^r_+ \times \Omega \times [0, 1] \rightrightarrows \Omega$ with $\hat{S}(\lambda, z, \mu) = \{u_\mu(\lambda, z)\}$ is upper semicontinuous with respect to (λ, z, μ) according to [5, Proposition 4.4]. Since for any $\lambda \in \mathbb{R}^r_+$, $z \in \Omega$, $\mu \in [0, 1]$, $\hat{S}(\lambda, z, \mu)$ is singleton, by the definition of upper semicontinuous multifunction [5, Section 4.1], we obtain the continuity of $u_\mu(\lambda, z)$. ∎

The following proposition is based on Proposition 3.2, and will be used in the proof of Theorem 3.4.

PROPOSITION 3.3. *If $(\lambda_\mu, z_\mu, u_\mu)$ is a feasible point of (3.1), then any accumulation point of $(\lambda_\mu, z_\mu, u_\mu)$ when $\mu \downarrow 0$ is a feasible point of (1.3).*

THEOREM 3.4. *If $(\lambda_\mu, z_\mu, u_\mu)$ is a global optimal solution of problem (3.1), then any accumulation point of $(\lambda_\mu, z_\mu, u_\mu)$ when $\mu \downarrow 0$ is a global optimal solution of problem (1.3).*

Proof. Let (λ^*, z^*, u^*) be an accumulation point of $(\lambda_\mu, z_\mu, u_\mu)$ when $\mu \downarrow 0$. According to Proposition 3.3, (λ^*, z^*, u^*) is a feasible point of (1.3). Assume that there exists a feasible point $(\tilde{\lambda}, \tilde{z}, \tilde{u})$ of problem (1.3) such that $F(\tilde{u}) < F(u^*)$. Due to the continuity of F, there exist δ_1, δ_2 such that for all $u^1 \in B(\tilde{u}, \delta_1)$ and $u^2 \in B(u^*, \delta_2)$, we have $F(u^1) < F(u^2)$. Notice that the solution $u_\mu(\tilde{\lambda}, \tilde{z})$ of lower level problem (3.2) converges to $\tilde{u}$ when $\mu \downarrow 0$, where $(\tilde{\lambda}, \tilde{z})$ is fixed. Letting $\tilde{\mu}$ be sufficiently small such that $\tilde{u}_{\tilde{\mu}} := u_{\tilde{\mu}}(\tilde{\lambda}, \tilde{z}) \in B(\tilde{u}, \delta_1)$ and $u_{\tilde{\mu}} \in B(u^*, \delta_2)$, we have $F(\tilde{u}_{\tilde{\mu}}) < F(u_{\tilde{\mu}})$, which obviously contradicts the global optimality of $(\lambda_{\tilde{\mu}}, z_{\tilde{\mu}}, u_{\tilde{\mu}})$. ∎

4. Smoothing implicit function method for problem (3.1). According to Theorems 2.2 and 2.3, the global (or local) optimal solutions of problem (1.3) correspond to the global (or local) optimal solutions of (1.1) under some assumptions. Further, due to Theorem 3.4, any accumulation point of global optimal solutions of problem (3.1) is the global optimal solution of problem (1.3) as the smoothing parameter μ goes to zero. Thus we focus on solving problem (3.1) with sufficiently small μ hereafter. For the ease of statement, we let $y = (\lambda^\top, z^\top)^\top$ and omit subscript μ.

Obviously, problem (3.1) can be equivalently transformed to

$$\min_{y,u} \quad F(u) \qquad \text{s.t. } y \in \mathbb{R}^r_+ \times \Omega,\ \Phi(y,u) = 0, \tag{4.1}$$

where $\Phi(y,u) := u - \Pi_\Omega(u - \tau(\nabla H(z) + \gamma(u-z) + \nabla_u Q_\mu(w;\lambda)))$ with fixed $\tau > 0$, and $\Pi_\Omega : \mathbb{R}^q \to \Omega$ is the projection operator.

By substituting unique solution $u(y)$ (subscript μ is omitted for brevity) into objective function F, problem (3.1) can be equivalently transformed to

$$\min_{y} \quad \tilde{F}(y) \qquad \text{s.t. } y \in \mathbb{R}^r_+ \times \Omega, \tag{4.2}$$

where $\tilde{F}(y) := F(u(y))$.

4.1. Smoothing approximation of problem (4.1). Since operator Π_Ω is not differentiable, we use the smoothing function proposed in [4] to approximate Φ, and consider

$$\min_{y,u} \quad F(u) \qquad \text{s.t. } y \in \mathbb{R}^r_+ \times \Omega,\ \Phi_\nu(y,u) = 0, \tag{4.3}$$

where Φ_ν is a smoothing function of Φ with smoothing parameter $\nu > 0$. The detailed formulation of Φ_ν can be found in Appendix.

According to Lemma 7.3(iii) and implicit function theorem, there exists a unique solution denoted by $u_\nu(y)$ to $\Phi_\nu(y,u) = 0$ for any fixed $y \in \mathbb{R}^r_+ \times \Omega$. Thus problem (4.3) can be equivalently transformed to

$$\min_{y} \quad \tilde{F}_\nu(y) \qquad \text{s.t. } y \in \mathbb{R}^r_+ \times \Omega, \tag{4.4}$$

where $\tilde{F}_\nu(y) := F(u_\nu(y))$.

Function Φ_ν based on the smoothing function in [4] enjoys impressive properties, which are presented as follows. Accordingly, $\Phi(y,u) = 0$ and $\Phi_\nu(y,u) = 0$ can have the same solution for a positive smoothing parameter ν.

PROPOSITION 4.1. *For any fixed $y \in \mathbb{R}^r_+ \times \Omega$, we have*

$$\|\Phi(y, u_\nu(y)) - \Phi_\nu(y, u_\nu(y))\| \le \frac{\sqrt{q}}{2}\nu, \tag{4.5}$$

for any $\nu \in (0,1]$. Moreover, for any fixed $y \in \mathbb{R}^r_+ \times \Omega$, there is $\tilde{\nu}$ such that

$$u_\nu(y) = u(y) \text{ and } \Phi(y, u_\nu(y)) = \Phi_\nu(y, u_\nu(y)) = 0, \tag{4.6}$$

for any $\nu \in (0, \tilde{\nu}]$.

Proof. From Lemma 7.3(i), we can obtain (4.5). Then we prove (4.6). Denote $\bar{\phi}(\tilde{y},\tilde{u}) := \tilde{u} - \tau\phi(\tilde{y},\tilde{u})$ for any $(\tilde{y},\tilde{u}) \in \mathbb{R}^r_+ \times \Omega \times \Omega$, where ϕ is defined in Appendix. Given any fixed $y \in \mathbb{R}^r_+ \times \Omega$, let $I_1 := \{i : \underline{u}_i > \bar{\phi}_i(y,u(y))\}$, $I_2 := \{i : \underline{u}_i \le \bar{\phi}_i(y,u(y)) \le \overline{u}_i\}$, $I_3 := \{i : \overline{u}_i < \bar{\phi}_i(y,u(y))\}$, $\rho_1 = \min\{3, \underline{u}_i - \bar{\phi}_i(y,u(y)) : i \in I_1\}$, $\rho_2 = \min\{3, \bar{\phi}_i(y,u(y)) - \overline{u}_i : i \in I_3\}$. Denote

$$\tilde{\nu} = \min\{(\rho_1/3)^2, (\rho_2/3)^2\}. \tag{4.7}$$

In order to prove (4.6), it suffices to show that

$$\psi^i_\nu(\bar{\phi}_i(y,u(y))) = \Pi_{[\underline{u}_i,\overline{u}_i]}(\bar{\phi}_i(y,u(y))) \tag{4.8}$$

holds for $\nu \in (0, \tilde{\nu}]$ and $i = 1, \cdots, q$. Actually, it is obvious that (4.8) holds for $i \in I_2$. Next, for $i \in I_1$, since $\tilde{\nu} \le (\rho_1/3)^2 \le 1$, we have

$$\bar{\phi}_i(y, u(y)) \le \underline{u}_i - \rho_1 \le \underline{u}_i - 3\sqrt{\tilde{\nu}} \le \underline{u}_i - \nu - 2\sqrt{\nu},$$

for $\nu \in (0, \tilde{\nu}]$. According to Lemma 7.1(ii), (4.8) holds for $i \in I_1$. Similarly, we can prove that (4.8) holds for $i \in I_3$. Therefore, (4.8) holds for $\nu \in (0, \tilde{\nu}]$ and $i = 1, \cdots, q$.□

PROPOSITION 4.2. *If (y_ν, u_ν) is a global optimal solution of problem (4.3), then any accumulation point of (y_ν, u_ν) when $\nu \downarrow 0$ is a global optimal solution of problem (4.1).*

Proof. Let (y^*, u^*) be an accumulation point of (y_ν, u_ν) when $\nu \downarrow 0$. For the ease of statement, we do not take the subsequence in the proof. Firstly, we prove that (y^*, u^*) is feasible for problem (4.1). Let $y = y_\nu$ in (4.5). Noting that $\Phi_\nu(y_\nu, u_\nu) = 0$ for $\nu > 0$, we have

$$\|\Phi(y_\nu, u_\nu)\| = \|\Phi(y_\nu, u_\nu) - \Phi_\nu(y_\nu, u_\nu)\| \le \frac{\sqrt{q}}{2}\nu. \tag{4.9}$$

Letting $\nu \downarrow 0$ in (4.9), we have $\Phi(y^*, u^*)$=0, which implies that $u^* = u(y^*)$ and (y^*, u^*) is feasible for problem (4.1). Then we show that (y^*, u^*) is a global optimal solution of problem (4.1). We prove this by contradiction. Assume that there exists a feasible point $(\tilde{y}, \tilde{u})$ of problem (4.1) such that $F(\tilde{u}) < F(u^*)$. Since (y_ν, u_ν) is a global optimal solution of problem (4.3), we have $F(u_\nu(\tilde{y})) \ge F(u_\nu)$. Letting $y = \tilde{y}$ and $\nu \downarrow 0$ in (4.5), we can obtain that $\lim_{\nu \downarrow 0} u_\nu(\tilde{y}) = \tilde{u}$, which implies that $F(\tilde{u}) \ge F(u^*)$. This contradicts the foregoing assumption. So we have proved the conclusion. □

If y is a local optimal solution of (4.2), then it satisfies $0 \in \partial \tilde{F}(y) + N_{\mathbb{R}^r_+ \times \Omega}(y)$. Via [9, Theorem 2.6.6], the above inclusion can be transformed to

$$0 \in (\partial u(y))^\top \nabla F(u(y)) + N_{\mathbb{R}^r_+ \times \Omega}(y). \tag{4.10}$$

Nevertheless, (4.10) involves the subdifferential of implicit function $u(\cdot)$, which is kind of elusive. So we introduce the concept of a weak Clarke stationary point for problem (4.2). Let $u = u(y)$. We call $y \in \mathbb{R}^r_+ \times \Omega$ a weak Clarke stationary point of (4.2) if there exist $V_1 \in \partial_u \Phi(y, u)$ and $V_2 \in \partial_y \Phi(y, u)$ such that (y, u) satisfies that

$$0 \in (-(V_1)^{-1} V_2)^\top \nabla F(u(y)) + N_{\mathbb{R}^r_+ \times \Omega}(y). \tag{4.11}$$

Remark 4.3. Here we give the explicit form of $\partial \Phi(y, u)$ for $(y, u) \in \mathbb{R}^r_+ \times \Omega \times \Omega$. Define

$$D(y, u) := \left\{ \operatorname{diag}(a) : a_i \in \left\{ \begin{array}{ll} \{1\}, & \text{if } u_i - \tau\phi_i(y, u) \in (\underline{u}_i, \overline{u}_i), \\ \{0\}, & \text{if } u_i - \tau\phi_i(y, u) \notin [\underline{u}_i, \overline{u}_i], \quad i = 1, \cdots, q \\ [0, 1], & \text{otherwise}, \end{array} \right. \right\},$$

where ϕ is defined in Appendix. Using the chain rule, we can derive that

$$\begin{aligned} \partial_u \Phi(y, u) &= \{(\tau\gamma - 1)D + I_q + \tau D \nabla_u^2 Q_\mu(w; \lambda) : D \in D(y, u)\}, \\ \partial_y \Phi(y, u) &= \{\tau D J_y \nabla_u Q_\mu(w; \lambda) + \tau D(0, \nabla^2 H(z) - \gamma I_q) : D \in D(y, u)\}. \end{aligned} \tag{4.12}$$

Remark 4.4. Actually, $S_\Phi := \{-(V_1)^{-1} V_2 : V_1 \in \partial_u \Phi(y, u),\ V_2 \in \partial_y \Phi(y, u)\}$ is an approximation of $\partial u(y)$ in (4.10). For example, when Φ is continuously differentiable near (y, u), we can show that $S_\Phi = \partial u(y)$. In fact, using [9, Proposition 2.2.4], we know that in this case, $\partial \Phi(y, u) = \{J\Phi(y, u)\}$, which indicates that $V_1 = J_u\Phi(y,u)$ and $V_2 = J_y\Phi(y,u)$. Further, via the implicit function theorem, we know $u(\cdot)$ is continuously differentiable near y and $\partial u(y) = \{Ju(y)\}$, where $Ju(y) = -(J_u\Phi(y,u))^{-1}J_y\Phi(y,u) = -(V_1)^{-1}V_2$.

On the other hand, $y \in \mathbb{R}^r_+ \times \Omega$ is said to be a stationary point of problem (4.4) if it satisfies

$$0 \in \nabla\tilde{F}_\nu(y) + N_{\mathbb{R}^r_+\times\Omega}(y). \tag{4.13}$$

Then we have the following proposition.

PROPOSITION 4.5. *If y_ν is a stationary point of problem (4.4), then any accumulation point of y_ν when $\nu \downarrow 0$ is a weak Clarke stationary point of problem (4.2).*

Proof. Using the implicit function theorem, we have

$$\nabla\tilde{F}_\nu(y) = -(J_y\Phi_\nu(y,u_\nu(y)))^\top(J_u\Phi_\nu(y,u_\nu(y)))^{-\top}\nabla F(u_\nu(y)). \tag{4.14}$$

Combining (4.14) with Lemma 7.3(ii), we can obtain the conclusion. ∎

4.2. Smoothing implicit function method. Motivated by Propositions 4.1, 4.2, and 4.5, problem (4.4) is a satisfying approximation of problem (4.2) for ν sufficiently small. In what follows, we will design a smoothing method where ν will eventually be small enough. The framework of the smoothing implicit function method is exhibited in Algorithm 4.1.

Algorithm 4.1 Smoothing implicit function method

Require: Choose parameters $\nu^0 \in (0,1]$, $\bar{\nu} \in (0,\nu^0]$, $\delta_1 > 0$, $\delta_2 \in (0,1)$, initial point $y^0 \in \mathbb{R}^r_+ \times \Omega$, stepsize $\theta > 0$, tolerances $\bar{\epsilon}, \epsilon_k \in (0,1)$ for $k = 0,1,2,\cdots$, and maximum number of iterations $k_{\max}$.

1: **for** $k = 0,1,2,\cdots$ **do**
2: Find u^k such that $\|\Phi_{\nu^k}(y^k,u^k)\| \le \epsilon_k$.
3: Find q^k such that $\|(J_u\Phi_{\nu^k}(y^k,u^k))^\top q^k - \nabla F(u^k)\| \le \epsilon_k$.
4: Compute $p^k = -(J_y\Phi_{\nu^k}(y^k,u^k))^\top q^k$.
5: Let
$$y^{k+1} = \Pi_{\mathbb{R}^r_+\times\Omega}(y^k - \theta p^k).$$
6: If $\left\|y^k - \Pi_{\mathbb{R}^r_+\times\Omega}(y^k-\theta p^k)\right\| \ge \delta_1\nu^k$, set $\nu^{k+1} = \nu^k$; otherwise, choose $\nu^{k+1} = \max\{\bar{\nu}, \delta_2\nu^k\}$.
7: If $\|y^{k+1} - y^k\| \le \bar{\epsilon}$ or $k = k_{\max}$, terminate, and return y^k and u^k.
8: **end for**

Note that $\{\nu^k\}$ in Algorithm 4.1 is lower bounded by $\bar{\nu}$ due to step 6, which guarantees that stepsize θ satisfying the assumptions for the convergence of Algorithm 4.1 can be found (see Proposition 4.7 and Lemma 4.9). There exists a trade-off in choosing $\bar{\nu}$. Actually, due to Propositions 4.2 and 4.5, $\bar{\nu}$ should approach 0 in terms of smoothing approximations, which, however, will lead to very small stepsize θ. In numerical experiments, $\bar{\nu}$ is tuned empirically from a set of given parameters.

The following assumption is about the boundedness of $\{\lambda^k\}$.

ASSUMPTION 4.6. *Let $\{y^k\}$ be the sequence generated by Algorithm 4.1. Assume that $\{\lambda^k\}$ is contained in a convex compact set U.*

Now we give some notations. Since F is twice continuously differentiable over Ω, F and ∇F are Lipschitz continuous over Ω with Lipschitz constants ℓ_F and L_F

respectively. Similarly, ∇Q_μ is Lipschitz continuous over $U\times\Omega$ with Lipschitz constant denoted by L_Q. Note that u_k in step 2 of Algorithm 4.1 may not be in Ω. Nevertheless, due to Lemma 7.3(vi) and the boundedness of $\{\epsilon_k\}$, there exists constant $C>0$ such that $\{u_k\}\subset\bar{\Omega}:=[\underline{u}-Ce_q,\overline{u}+Ce_q]$. Since the analysis involving Ω can be extended to $\bar{\Omega}$, we will assume that $u_k\in\Omega$ in this paper for simplicity.

Using Lemma 7.3, we can prove the following proposition.

PROPOSITION 4.7. *For $\nu\in[\bar{\nu},1]$, there exists $\tilde{L}>0$ not related to ν such that*

$$\|\nabla\tilde{F}_\nu(y^1)-\nabla\tilde{F}_\nu(y^2)\|\le\tilde{L}\left\|y^1-y^2\right\| \tag{4.15}$$

for any $y^1,y^2\in U\times\Omega$.

Proof. The Lipschitz continuity of $\nabla\tilde{F}_\nu$ is clear from (4.14) and Lemma 7.3(iv)(v). Since ν is lower bounded by $\bar{\nu}>0$, $\tilde{L}$ is not related to ν by Lemma 7.3(iv). ∎

The following lemma shows that p^k approximates $\nabla\tilde{F}_{\nu^k}(y^k)$ well.

LEMMA 4.8. *Let Assumption 4.6 hold. Assume that $\tau\gamma<1$, $\gamma>L_Q$, and $\sum_{k=0}^{\infty}\epsilon^k<\infty$ in Algorithm 4.1. Then there exists $\bar{k}_1>0$ and $\tilde{M}>0$ such that*

$$\|\nabla\tilde{F}_{\nu^k}(y^k)-p^k\|\le\tilde{M}\epsilon_k, \tag{4.16}$$

for $k\ge\bar{k}_1$.

Proof. From Algorithm 4.1, we know that $\bar{\nu}\le\nu^k\le\nu^0$ for $k\ge0$. Let

$$\begin{aligned}&J_u^k:=J_u\Phi_{\nu^k}(y^k,u_{\nu^k}(y^k)),\ \tilde{J}_u^k:=J_u\Phi_{\nu^k}(y^k,u^k),\\&J_y^k:=J_y\Phi_{\nu^k}(y^k,u_{\nu^k}(y^k)),\ \tilde{J}_y^k:=J_y\Phi_{\nu^k}(y^k,u^k),\\&f^k:=\nabla F(u_{\nu^k}(y^k)),\ \tilde{f}^k:=\nabla F(u^k).\end{aligned}$$

Due to Lemma 7.3(iv), there exists upper bound $M_1>0$ for the norms of the above terms. Since $\bar{\nu}\le\nu^k\le\nu^0$ for $k\ge0$, from Lemma 7.3(iv)(v), there exists upper bound $M_2>0$ for $\{\|(J_u^k)^{-1}\|\}$, $\{\|(\tilde{J}_u^k)^{-1}\|\}$, $\{|\ell_{\nu^k}^u|\}$ and $\{|\ell_{\nu^k}^y|\}$ as well.

Using Lemma 7.3(vi), we know that $\|u_{\nu^k}(y^k)-u^k\|\le\frac{\epsilon_k}{\tau(\gamma-L_Q)}$. Let v^k be the solution to $(J_u^k)^\top v^k=f^k$ and $\tilde{v}^k$ be the solution to $(\tilde{J}_u^k)^\top\tilde{v}^k=\tilde{f}^k$. Obviously, $\max\{\|v^k\|,\|\tilde{v}^k\|\}\le M_1M_2$ for $k\ge0$. Now we investigate $\|v^k-\tilde{v}^k\|$. Due to Lemma 7.3(v), we have

$$\|J_u^k-\tilde{J}_u^k\|\le\frac{\ell_{\nu^k}^u\epsilon_k}{\tau(\gamma-L_Q)},\ \|f^k-\tilde{f}^k\|\le\frac{L_F\epsilon_k}{\tau(\gamma-L_Q)}.$$

Since $\sum_{k=0}^{\infty}\epsilon^k<\infty$, there exists constants $\bar{k}_1,\bar{c}_1>0$ such that $\frac{\ell_{\nu^k}^u\epsilon_k\|(J_u^k)^{-\top}\|}{\tau(\gamma-L_Q)}\le\bar{c}_1<1$, for $k\ge\bar{k}_1$. Due to [19, Theorem 7.2], for $k\ge\bar{k}_1$, we have

$$\begin{aligned}\|v^k-\tilde{v}^k\|\le&\frac{\frac{\epsilon_k}{\tau(\gamma-L_Q)}}{1-\frac{\ell_{\nu^k}^u\epsilon_k\|(J_u^k)^{-\top}\|}{\tau(\gamma-L_Q)}}(L_F\|(J_u^k)^{-\top}\|+\ell_{\nu^k}^u\|v^k\|\|(J_u^k)^{-\top}\|)\\\le&\frac{\epsilon_k}{\tau(\gamma-L_Q)(1-\bar{c}_1)}(L_F\|(J_u^k)^{-\top}\|+\ell_{\nu^k}^u\|v^k\|\|(J_u^k)^{-\top}\|)\\\le&M_3\epsilon_k,\end{aligned}$$

where $M_3:=\frac{L_FM_2+M_1(M_2)^3}{\tau(\gamma-L_Q)(1-\bar{c}_1)}$.

Then we investigate $\|q^k - v^k\|$. Actually, for $k \geq \bar{k}_1$,

$$\begin{aligned}\|q^k - v^k\| =&\|q^k - \tilde{v}^k + \tilde{v}^k - v^k\| \\ \leq&\|q^k - \tilde{v}^k\| + \|v^k - \tilde{v}^k\| \\ \leq&\|(\tilde{J}_u^k)^{-\top}(\tilde{J}_u^k)^\top(q^k - \tilde{v}^k)\| + \|v^k - \tilde{v}^k\| \\ \leq&\|(\tilde{J}_u^k)^{-\top}\|\|(\tilde{J}_u^k)^\top q^k - \tilde{f}^k\| + \|v^k - \tilde{v}^k\| \\ \leq&(M_2 + M_3)\epsilon_k,\end{aligned}$$

where the last equality follows from the fact that $\|(\tilde{J}_u^k)^\top q^k - \tilde{f}^k\| \leq \epsilon_k$.

Finally, for $k \geq \bar{k}_1$, we have

$$\begin{aligned}\|\nabla \tilde{F}_{\nu^k}(y^k) - p^k\| =&\|(J_y^k)^\top v^k - ((\tilde{J}_y^k)^\top q^k)\| \\ =&\|(J_y^k)^\top v^k - (\tilde{J}_y^k)^\top v^k + (\tilde{J}_y^k)^\top v^k - (\tilde{J}_y^k)^\top q^k\| \\ \leq&\|(J_y^k)^\top v^k - (\tilde{J}_y^k)^\top v^k\| + \|(\tilde{J}_y^k)^\top v^k - (\tilde{J}_y^k)^\top q^k\| \\ \leq&\|J_y^k - \tilde{J}_y^k\|\|v^k\| + \|\tilde{J}_y^k\|\|v^k - q^k\| \\ \leq&\frac{\ell_{\nu^k}^y \epsilon_k\|v^k\|}{\tau(\gamma - L_Q)} + \|\tilde{J}_y^k\|\|v^k - q^k\| \\ \leq&\tilde{M}\epsilon_k,\end{aligned}$$

where the last but one inequality follows from Lemma 7.3(iv), and the final estimate uses $\tilde{M} := \frac{M_1(M_2)^2}{\tau(\gamma-L_Q)} + M_1M_2 + M_1M_3$. ∎

Lemma 4.9. *Let assumptions of Lemma 4.8 hold. Assume that $\theta \leq \frac{1}{\tilde{L}}$ in Algorithm 4.1, where $\tilde{L}$ is defined in Proposition 4.7. Then there exists $\bar{k}_2 > 0$ such that $\nu^k = \bar{\nu}$, for $k \geq \bar{k}_2$.*

Proof. Denote set $K := \{k : \nu^{k+1} = \max\{\bar{\nu}, \delta_2\nu^k\}\}$. It suffices to prove that set K is infinite. We prove this by contradiction. Suppose that K is finite. Then there exist $\hat{\nu} > \bar{\nu}$ and $k_0 > 0$ such that for $k \geq k_0$,

$$\nu^k = \hat{\nu} \text{ and } \|y^{k+1} - y^k\| \geq \delta_1\hat{\nu}. \tag{4.17}$$

From (4.15), we know that $\tilde{F}_{\hat{\nu}}$ satisfies that

$$\tilde{F}_{\hat{\nu}}(y_a) \leq \tilde{F}_{\hat{\nu}}(y_b) + \nabla\tilde{F}_{\hat{\nu}}(y_b)^\top(y_a - y_b) + \frac{\tilde{L}}{2}\|y_a - y_b\|^2 \tag{4.18}$$

for any $y_a, y_b \in U \times \Omega$. Due to Lemma 7.2(ii), we have

$$\|\Pi_{\mathbb{R}^r_+\times\Omega}(y_a) - \Pi_{\mathbb{R}^r_+\times\Omega}(y_b)\|^2 \leq (y_a - y_b)^\top(\Pi_{\mathbb{R}^r_+\times\Omega}(y_a) - \Pi_{\mathbb{R}^r_+\times\Omega}(y_b)).$$

Letting $y_a = y^k - \theta p^k$ and $y_b = y^k$ in the above inequality, we can obtain that

$$\|y^{k+1} - y^k\|^2 \leq -\theta(p^k)^\top(y^{k+1} - y^k). \tag{4.19}$$

Let $\bar{k}_2 = \max\{k_0, \bar{k}_1\}$ with $\bar{k}_1$ defined in Lemma 4.8. Substituting y^{k+1}, y^k into (4.18), for $k \geq \bar{k}_2$, we have

$$
\begin{aligned}
&\tilde{F}_{\hat{\nu}}(y^{k+1})\\
\leq&\tilde{F}_{\hat{\nu}}(y^k) + \nabla\tilde{F}_{\hat{\nu}}(y^k)^\top(y^{k+1}-y^k) + \frac{\tilde{L}}{2}\|y^{k+1}-y^k\|^2\\
=&\tilde{F}_{\hat{\nu}}(y^k) + (\nabla\tilde{F}_{\hat{\nu}}(y^k)-p^k)^\top(y^{k+1}-y^k) + (p^k)^\top(y^{k+1}-y^k) + \frac{\tilde{L}}{2}\|y^{k+1}-y^k\|^2\\
\leq&\tilde{F}_{\hat{\nu}}(y^k) + (\nabla\tilde{F}_{\hat{\nu}}(y^k)-p^k)^\top(y^{k+1}-y^k) - \frac{1}{\theta}\|y^{k+1}-y^k\|^2 + \frac{\tilde{L}}{2}\|y^{k+1}-y^k\|^2\\
\leq&\tilde{F}_{\hat{\nu}}(y^k) + \|\nabla\tilde{F}_{\hat{\nu}}(y^k)-p^k\|\|y^{k+1}-y^k\| - \frac{\tilde{L}}{2}\|y^{k+1}-y^k\|^2\\
\leq&\tilde{F}_{\hat{\nu}}(y^k) + M\epsilon^k - \frac{\tilde{L}}{2}\|y^{k+1}-y^k\|^2,
\end{aligned}
$$

where the second inequality holds from (4.19), the third inequality holds from the fact that $\theta \leq \frac{1}{\tilde{L}}$, the last inequality follows from Lemma 4.8 and the boundedness of $\{y^k\}$, and constant $M > 0$ is constructed based on $\tilde{M}$. So we obtain that

$$
\|y^{k+1}-y^k\|^2 \leq \frac{2}{\tilde{L}}(\tilde{F}_{\hat{\nu}}(y^k) - \tilde{F}_{\hat{\nu}}(y^{k+1}) + M\epsilon^k), \tag{4.20}
$$

for $k \geq \bar{k}_2$. Summing (4.20) for $k = \bar{k}_2, \bar{k}_2+1, \cdots$, we have

$$
\sum_{k=\bar{k}_2}^{\infty}\|y^{k+1}-y^k\|^2 \leq \frac{2}{\tilde{L}}\left(\tilde{F}_{\hat{\nu}}(y^{\bar{k}_2}) + M\sum_{k=\bar{k}_2}^{\infty}\epsilon^k\right).
$$

Since $\sum_{k=0}^{\infty}\epsilon^k < \infty$, we know that $\lim_{k\to\infty}\|y^{k+1}-y^k\| = 0$, which contradicts (4.17). So we have proved the conclusion. ∎

THEOREM 4.10. *Let assumptions of Lemma 4.9 hold. Let $(\tilde{y}, \tilde{u})$ be an accumulation point of sequence $\{(y^k, u^k)\}$ generated by Algorithm 4.1. Then $\tilde{y}$ satisfies that*

$$
0 \in \nabla\tilde{F}_{\bar{\nu}}(\tilde{y}) + N_{\mathbb{R}^r_+\times\Omega}(\tilde{y}), \tag{4.21}
$$

where $\nabla\tilde{F}_{\bar{\nu}}(\tilde{y}) = (-(J_u\Phi_{\bar{\nu}}(\tilde{y},\tilde{u}))^{-1}J_y\Phi_{\bar{\nu}}(\tilde{y},\tilde{u}))^\top\nabla F(u_{\bar{\nu}}(\tilde{y}))$.

Proof. According to the proof of Lemma 4.9, we have

$$
\lim_{k\to\infty}\|y^k - \Pi_{\mathbb{R}^r_+\times\Omega}(y^k - \theta p^k)\| = 0. \tag{4.22}
$$

Via Lemmas 4.8 and 4.9, we have

$$
\|\nabla\tilde{F}_{\bar{\nu}}(y^k) - p^k\| \leq \tilde{M}\epsilon_k, \tag{4.23}
$$

for $k \geq \bar{k}_2$ with $\bar{k}_2$ defined in Lemma 4.9. By virtue of (4.23), (4.22) can be transformed to

$$
\|\tilde{y} - \Pi_{\mathbb{R}^r_+\times\Omega}(\tilde{y} - \theta\nabla\tilde{F}_{\bar{\nu}}(\tilde{y}))\| = 0,
$$

which is equivalent to (4.21). The explicit form of $\nabla\tilde{F}_{\bar{\nu}}(\tilde{y})$ follows from (4.14). To show that $\tilde{u} = u_{\bar{\nu}}(\tilde{y})$, we utilize Lemma 7.3(vi) and obtain

$$
\|u^k - u_{\bar{\nu}}(y^k)\| \leq \frac{\epsilon_k}{\tau(\gamma - L_Q)}, \tag{4.24}
$$

for $k \geq \bar{k}_2$. Letting $k \to \infty$ in both sides of (4.24), we have $\tilde{u} = u_{\bar{\nu}}(\tilde{y})$. ∎

5. Numerical experiments. In this section, we will conduct numerical experiments on the feed-forward neural network. The synthetic data and real-life datasets from UCI machine learning repository [26] will be tested respectively.

Algorithm 4.1 will be compared with the Grid Search method, the Random Search method and the Bayesian optimization method, where Random Search method (see [18, 30]) and Bayesian optimization method (see [3, 39]) are also widely used for hyperparameter optimization in machine learning. The Grid Search method is to solve (1.2) for every grid point respectively and determine the best hyperparameter according to the validation error [34]. The Random Search method is basically the same strategy, except that the grid points are chosen randomly. To use Grid Search method and Random Search method, we denote $\alpha_j^\ell = a_0$ for $\ell = 1, 2$ and $j = 1, \cdots, n_{\ell-1}$, and choose a_0 from some given set (see [34]). The Bayesian optimization method used in this paper is from [3]. In Grid Search method, Random Search method and Bayesian optimization method, problem (1.2) with fixed λ is solved via ADADELTA [44].

5.1. Tests on synthetic data. The synthetic data are randomly generated in similar way as used in [10, Section 5.1]. We consider bilevel optimization for tuning hyperparameters of 2-layer sparse feed-forward neural networks. We first randomly generate $X^i \sim \mathcal{N}(\zeta, \Sigma_0\Sigma_0^\top)$ with $\zeta = randn(n, 1)$ and $\Sigma_0 = randn(n, 1)$. The activation function σ is the sigmoid function denoted by $\sigma(t) = \frac{1}{1+e^{-t}},\ t \in \mathbb{R}$. Truth values of W_1^*, W_2^* and $b^{1,*}, b^{2,*}$ are randomly generated as follows. Generate $\bar{W}_1 \in \mathbb{R}^{n_1 \times n}$ and $\bar{W}_2 \in \mathbb{R}^{1 \times n_1}$ from the uniform distribution $\mathcal{U}(-1, 1)$, and choose index sets $J_1 \subseteq \{1, \cdots, n\}$ of size $|J_1|$ and $J_2 \subseteq \{1, \cdots, n_1\}$ of size $|J_2|$ randomly. Let $(\bar{W}_1)_{\cdot j} = 0$ for $j \in J_1$ and $(\bar{W}_2)_{\cdot j} = 0$ for $j \in J_2$. Denote $W_1^* = \bar{W}_1$ and $W_2^* = \bar{W}_2$, and generate $b^{1,*}, b^{2,*}$ from the uniform distribution $\mathcal{U}(-1, 1)$. Then we generate

$$Y_i = W_2^* \sigma(W_1^* X^i + b^{1,*}) + b^{2,*} + \tilde{Y}_i,\ i = 1, \cdots, \bar{N},$$

where $\tilde{Y}_i \sim 0.05\mathcal{N}(0, 1)$ is the noise. The synthetic data are divided into three groups indexed by integers N_{tr}, N_{va} and N_{te}. Specifically, $\{(X^i, Y^i) : i = 1, \cdots, N_{tr}\}$ is the training group, $\{(X^i, Y^i) : i = N_{tr} + 1, \cdots, N_{tr} + N_{va}\}$ is the validation group, and $\{(X^i, Y^i) : i = N_{tr} + N_{va} + 1, \cdots, \bar{N}\}$ is the test group. We set $\overline{u} = 20 * e_q$ and $\underline{u} = -20 * e_q$.

Denote the calculated solutions by W_1, W_2, and b^1, b^2. The test error is denoted as

$$\text{TestErr} := \frac{1}{N_{te}} \sum_{i=N_{tr}+N_{va}+1}^{\bar{N}} \|W_2\sigma(W_1X^i + b^1) + b^2 - Y^i\|^2.$$

The validation error is denoted as

$$\text{ValErr} := \frac{1}{N_{va}} \sum_{i=N_{tr}+1}^{N_{tr}+N_{va}} \|W_2\sigma(W_1X^i + b^1) + b^2 - Y^i\|^2.$$

We denote by Z_0 the number of zero columns of W_1 and W_2. Denote Z_c the number of zero columns that W_1^*, W_2^* and W_1, W_2 have in common. Here the columns of W_1 and W_2 are taken as zero vectors if their Euclidean norms are less than 10^{-3}.

In the experiments, we let $N_{tr} = \lceil 0.6\bar{N} \rceil$ and $N_{va} = \lceil 0.2\bar{N} \rceil$. The remaining data are set to be the test group. We consider nine combinations of $(\bar{N}, n, n_1, |J_1|, |J_2|)$ presented in Table 1.

In the implementation of Algorithm 4.1, we set $\nu^0 = 1$, $\delta_1 = 100$, $\delta_2 = 0.9$, and $\epsilon_k = \frac{0.1}{k^2}$ ($\epsilon_0 = 0.1$). We let $\alpha_j^\ell = 10^{-4}$ for $\ell = 1, 2$ and $j = 1, \cdots, n_{\ell-1}$, and take the

Table 1: Datatype

D1	D2	D3
(500,50,10,10,5)	(1000,100,40,30,10)	(2000,200,40,30,10)
D4	D5	D6
(3000,300,50,40,10)	(5000,500,100,80,40)	(5000,1000,100,100,50)
D7	D8	D9
(10000,1000,300,200,100)	(10000,2000,400,300,100)	(10000,3000,500,600,200)

solution of (1.2) calculated via the ADADELTA algorithm as z^0. The quasi-Newton method in [6] is employed in step 2, and q^k is obtained by the conjugate gradient method. We let $\bar{\epsilon} = 10^{-5}$ and $k_{\max} = 500$.

We set μ and $\bar{\nu}$ among $\{10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$, and employ the setting with the lowest validation error. In order to determine parameter γ, we use the Matlab built-in solver `fmincon` to solve the following problem:

$$(5.1) \qquad \max_{z} \quad \|\nabla^2 H(z)\|_F^2 \qquad \text{s.t. } z \in \Omega,$$

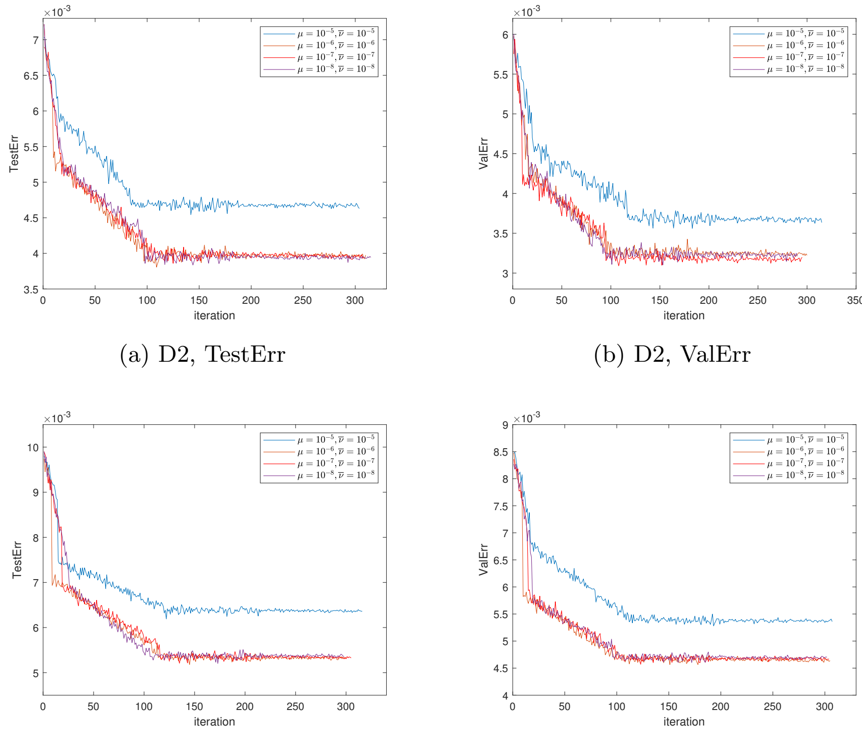
where $\|\cdot\|_F$ denotes the Frobenius norm. Denote by $\tilde{\gamma}$ the positive square root of optimal value of problem (5.1). Similarly, we can evaluate L_Q, where we set $U := [10^{-4}, 10^4]^r$. Then we let $\gamma = 2\max\{\tilde{\gamma}, L_Q\}$, and $\tau = \frac{1}{2\gamma}$. For each setting of μ and $\bar{\nu}$, it is difficult to calculate $\tilde{L}$ in practice, so we can not designate stepsize θ directly. Motivated by [17], we choose stepsize θ from $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1\}$, and accept the one with the lowest validation error.

Some numerical results about datasets $D2$ and $D3$ are exhibited in Fig. 1, where we can find that Algorithm 4.1 performs better when μ and $\bar{\nu}$ are smaller, and the performances are insensitive to the setting of μ and $\bar{\nu}$ when μ and $\bar{\nu}$ are smaller than 10^{-6}. In the implementations, the mini-batch technique [22] is employed to accelerate the computing of Algrithm 4.1, which leads to the oscillations in Fig. 1.

In the Grid Search method, we choose hyperparameter a_0 from set $\{10^{-k} : k = -4, \cdots, 4\}$. In the Random Search method, let $a_0 = 10^{-\omega}$, and generate ω 10 times from the uniform distribution $\mathcal{U}(-4, 4)$. For both methods, the hyperparameter with the smallest validation error will be accepted. In the Bayesian optimization method, for $\ell = 1, 2$ and $j = 1, \cdots, n_{\ell-1}$, we denote $\alpha_j^\ell = 10^{-\omega_j^\ell}$, and search over the transformed variable ω_j^ℓ, where the search space of ω_j^ℓ is defined as the uniform distribution $\mathcal{U}(-4, 4)$.

For every type of data, 10 examples are randomly generated, and the average results are exhibited in Table 2 and Fig. 2. Here we can see that Algorithm 4.1 performs best in regard to test error and validation error, and the gap widens with the increase of the scale of the data. All methods yield sparse neural networks, and the networks trained via Algorithm 4.1 are sparser when the size is larger. The above numerical experiments are conducted on 2-layer neural networks which can be very wide (see datatypes D8 and D9). However, considering the partially difficult computations in each iteration (solving a nonlinear system via quasi-Newton method and a linear system via conjugate gradient method), Algorithm 4.1 is more suitable for wide but not very deep neural networks.

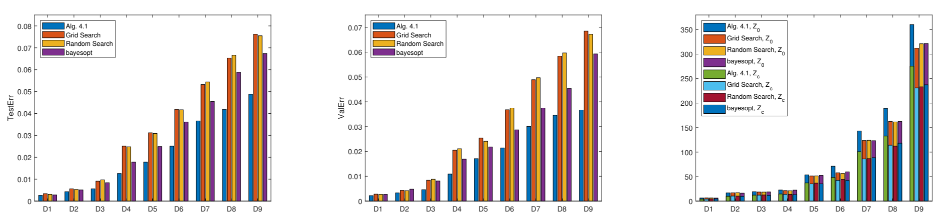
Denote StaErr $= \left\| y - \Pi_{\mathbb{R}_+^r \times \Omega}(y - \theta p) \right\|$, where y, p are obtained from the last iteration. The numerical results are presented in Table 3, where "Iter" denotes the



(a) D2, TestErr (b) D2, ValErr

(c) D3, TestErr (d) D3, ValErr

Fig. 1: Comparison of Algorithm 4.1 with varying μ and $\bar{\nu}$



(a) TestErr (b) ValErr (c) Z_0 and Z_c

Fig. 2: Numerical results for synthetic data

average number of outer iterations, and "Time" denotes the average CPU time in seconds.

5.2. Tests on real-life data. Now we conduct the experiments on the real-life datasets. These datasets are downloaded from UCI machine learning repository [26], including Higher Education Students Performance Evaluation Dataset (Student), Facebook Comment Volume Dataset (Facebook), Insurance Company Benchmark

Table 2: Numerical results for synthetic data

	Alg	TestErr	ValErr	Z_0	Z_c
D1	Alg. 4.1	**0.0026**	**0.0022**	6.3	4.1
	Grid Search	0.0034	0.0028	**6.5**	4.1
	Random Search	0.0031	0.0027	6.4	4.4
	bayesopt	0.0028	0.0027	6.2	**4.4**
D2	Alg. 4.1	**0.0043**	**0.0033**	16.7	9.9
	Grid Search	0.0056	0.0043	16.9	**10.3**
	Random Search	0.0053	0.0042	**17.2**	10.2
	bayesopt	0.0051	0.0048	16.2	9.8
D3	Alg. 4.1	**0.0056**	**0.0046**	**19.2**	12.2
	Grid Search	0.0091	0.0084	18.5	12.2
	Random Search	0.0097	0.0088	18.4	**12.5**
	bayesopt	0.0084	0.0081	18.7	11.8
D4	Alg. 4.1	**0.0126**	**0.0109**	**22.6**	**14.5**
	Grid Search	0.0251	0.0205	21.5	13.7
	Random Search	0.0248	0.0211	21.1	13.9
	bayesopt	0.0178	0.0169	22.4	13.2
D5	Alg. 4.1	**0.0178**	**0.0171**	**53.5**	**37.1**
	Grid Search	0.0312	0.0254	51.3	35.2
	Random Search	0.0309	0.0241	51.4	36.7
	bayesopt	0.0249	0.0218	52.4	35.3
D6	Alg. 4.1	**0.0251**	**0.0214**	**71.2**	**48.2**
	Grid Search	0.0419	0.0368	57.8	42.4
	Random Search	0.0417	0.0375	56.4	44.2
	bayesopt	0.0361	0.0287	59.7	42.1
D7	Alg. 4.1	**0.0366**	**0.0301**	**143.1**	**100.9**
	Grid Search	0.0532	0.0489	123.6	86.1
	Random Search	0.0544	0.0497	124.1	86.7
	bayesopt	0.0455	0.0375	123.3	88.7
D8	Alg. 4.1	**0.0419**	**0.0346**	**189.4**	**132.8**
	Grid Search	0.0653	0.0584	162.7	114.1
	Random Search	0.0666	0.0597	161.6	112.2
	bayesopt	0.0588	0.0454	162.4	118.2
D9	Alg. 4.1	**0.0488**	**0.0367**	**360.4**	**275.4**
	Grid Search	0.0762	0.0685	312.2	231.1
	Random Search	0.0755	0.0672	321.2	233.2
	bayesopt	0.0674	0.0593	321.7	237.4

Dataset (Insurance) and BlogFeedback Dataset (Blog).

We use the min-max normalization technique to rescale the data to $[0, 1]$. The settings of the algorithms and evaluation criteria are same as those in the last subsection. The numerical results are exhibited in Table 4 and Fig. 3. Here we can find that Algorithm 4.1 performs better than Grid Search method, Random Search method and Bayesian optimization method, especially in terms of Student dataset and Facebook dataset.

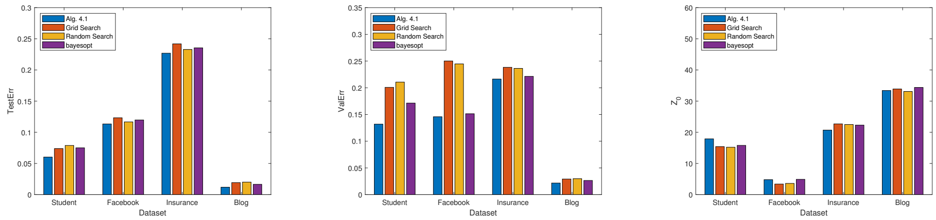
Table 3: Numerical results for synthetic data

	D1	D2	D3	D4	D5
StaErr	2.234e-06	4.162e-06	3.462e-06	1.181e-06	4.842e-06
Iter	303.4	305.8	321.6	311.4	348.5
Time	10.1	35.4	60.5	99.7	255.2
	D6	D7	D8	D9	
StaErr	9.467e-06	5.233e-06	8.462e-06	4.238e-06	
Iter	309.1	343.2	356.8	355.5	
Time	469.5	787.1	1449.9	2926.3	

Table 4: Numerical results for real-life data

Dataset	$(\bar{N}, m, n_1, n)$	Alg	TestErr	ValErr	Z_0
Student	(145,1,10,31)	Alg. 4.1	**0.0603**	**0.1319**	**17.9**
		Grid Search	0.0739	0.2008	15.4
		Random Search	0.0789	0.2106	15.2
		bayesopt	0.0751	0.1713	15.8
Facebook	(40949,1,10,53)	Alg. 4.1	**0.1134**	**0.1458**	4.8
		Grid Search	0.1233	0.2501	3.4
		Random Search	0.1167	0.2447	3.6
		bayesopt	0.1198	0.1514	**4.9**
Insurance	(5822,1,20,85)	Alg. 4.1	**0.2269**	**0.2163**	20.7
		Grid Search	0.2419	0.2383	**22.7**
		Random Search	0.2329	0.2363	22.5
		bayesopt	0.2355	0.2214	22.3
Blog	(52397,1,50,280)	Alg. 4.1	**0.0119**	**0.0217**	33.4
		Grid Search	0.0193	0.0292	33.9
		Random Search	0.0201	0.0298	33.1
		bayesopt	0.0165	0.0265	**34.4**

6. Conclusion. In the bilevel optimization problem (1.1) for tuning hyperparameters of sparse neural networks, lower level problem (1.2) is nonconvex and nonsmooth, which makes the problems computationally intractable. By using the structure of the objective function in (1.2), a convex majorant approach with smooth approximations is proposed in this paper. In particular, we introduce a convex majorant function $G(\cdot;\lambda,z)$ to approximate the objective function of the lower level problem (1.2), and establish the relationship between the original bilevel optimization (1.1) and the bilevel optimization (1.3) with $G(\cdot;\lambda,z)$ regarding global and local minimizers. Then we use smoothing function $G_\mu(\cdot;\lambda,z)$ to approximate $G(\cdot;\lambda,z)$, and derive the convergence of global minimizers to those of problem (1.3) with smoothing parameter μ converging to zero. The approximate bilevel optimization problem (3.1) with $G_\mu(\cdot;\lambda,z)$ is solved via the smoothing implicit function method. The numerical experiments including the tests on the data from machine learning repository indicate that the convex majorant approach performs better than the Grid Search method, the Random Search method and the Bayesian optimization method.



(a) TestErr (b) ValErr (c) Z_0

Fig. 3: Numerical results for real-life data

Acknowledgement The authors are grateful to the associate editor and two anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] J.H. Alcantara, C.T. Nguyen, T. Okuno, A. Takeda and J.S. Chen, *Unified smoothing approach for best hyperparameter selection problem using a bilevel optimization strategy*, Math. Program., (2024), pp. 1–40.

[2] A.R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inf. Theory, 39 (1993), pp. 930–945.

[3] J. Bergstra, D. Yamins and D. Cox, *Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures*, in International Conference on Machine Learning, 2013, pp. 115-123.

[4] W. Bian and X. Chen, *Anderson acceleration for nonsmooth fixed point problems*, SIAM J. Numer. Anal., 60 (2022), pp. 2565–2591.

[5] J.F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer, New York, 2013.

[6] C. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Math. Comp. 19 (1965), pp. 577–593.

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

[8] X. Chen, L. Qi and D. Sun, *Global and superlinear convergence of the smoothing Newton method and its application to general box constrained variational inequalities*, Math. Comput., 67 (1998), pp. 519–540.

[9] F. H. Clarke, *Optimization and Nonsmooth Analysis*, SIAM Publisher, Philadelphia, 1990.

[10] Y. Cui, Z. He and J.-S. Pang, *Multicomposite nonconvex optimization for training deep neural networks*, SIAM J. Optim., 30 (2020), pp. 1693–1723.

[11] G. Cybenko, *Approximation by superpositions of a sigmoidal function*, Math. Control Signal Syst., 2 (1989), pp. 303–314.

[12] S. Dempe and A.B. Zemkoho, *Bilevel Optimization: Advances and Next Challenges, Vol. 161*, Springer Optimization and its Applications, Berlin, 2020.

[13] P. Domingos, *A few useful things to know about machine learning*, Commun. ACM, 55 (2012), pp. 78–87.

[14] J. Feng and N. Simon, *Sparse-input neural networks for high-dimensional nonparametric regression and classification*, preprint, arxiv:1711.07592, (2017).

[15] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi and M. Pontil, *Bilevel programming for hyperparameter optimization and meta-learning*, in International Conference on Machine Learning, 2018, pp. 1568–1577.

[16] F. Facchinei and J.S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer, New York, 2003.

[17] R. Grazzi, L. Franceschi, M. Pontil and S. Salzo, *On the iteration complexity of hypergradient computation*, in International Conference on Machine Learning, 2020, pp. 3748–3758.

[18] L. Gao, J.J. Ye, H. Yin, S. Zeng and J. Zhang, *Value function based difference-of-convex algorithm for bilevel hyperparameter selection problems*, in International Conference on

Machine Learning, 2022, pp. 7164-7182.
[19] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 2002.
[20] T. Hoefler, D. Alistarh, T. Ben-Nun, N. Dryden and A. Peste, *Sparsity in deep learning: pruning and growth for efficient inference and training in neural networks*, J. Mach. Learn. Res., 22 (2021), pp. 1–124.
[21] R.A. Horn and C.R. Johnson, *Matrix Analysis*, Cambridge University Press, New York, 2013.
[22] H. Kasai, *SGDLibrary: A MATLAB library for stochastic optimization algorithms*, J. Mach. Learn. Res., 18 (2018), pp. 1–5.
[23] M. Leshno, V.Y. Lin, A. Pinkus and S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, Neural Netw., 6 (1993), pp. 861–867.
[24] Y. Li, G. Lin, J. Zhang and X. Zhu, *A novel approach for bilevel programs based on Wolfe duality*, preprint, arxiv:2302.06838, (2023).
[25] G. Li and X. Yang, *Convexification method for bilevel programs with a nonconvex follower's problem*, J. Optim. Theory Appl., 188 (2021), pp. 724–743.
[26] M. Lichman, *UCI machine learning repository*, (2013), URL http://archive.ics.uci.edu/ml.
[27] G. Lin, M. Xu and J.J. Ye, *On solving simple bilevel programs with a nonconvex lower level program*, Math. Program., 144 (2014), pp. 277–305.
[28] R. Liu, J. Gao, J. Zhang, D. Meng and Z. Lin, *Investigating bi-level optimization for learning and vision from a unified perspective: a survey and beyond*, IEEE Trans. Pattern Anal. Mach. Intell., 44 (2021), pp. 10045–10067.
[29] R. Liu, X. Liu, X. Yuan, S. Zeng and J. Zhang, *A value-function-based interior-point method for non-convex bi-level optimization*, in International Conference on Machine Learning, 2021, pp. 6882–6892.
[30] R. Liu, Z. Liu, W. Yao, S. Zeng and J. Zhang, *Moreau envelope for nonconvex bi-level optimization: a single-loop and Hessian-free solution strategy*, preprint, arxiv:2405.09927, (2024).
[31] R. Liu, Y. Liu, S. Zeng and J. Zhang, *Towards gradient-based bilevel optimization with non-convex followers and beyond*, Adv. Neural Inf. Process. Syst., (34) 2021, pp. 8662–8675.
[32] J.A. Mirrlees, *The theory of moral hazard and unobservable behaviour: Part I*, Rev. Econ. Stud., 66 (1999), pp. 3–21.
[33] A. Mitsos, P. Lemonidis and P.I. Barton, *Global solution of bilevel programs with a non-convex inner program*, J. Glob. Optim., 42 (2008), pp. 475–513.
[34] T. Okuno, A. Takeda, A. Kawana and M. Watanabe, *On l_p-hyperparameter learning via bilevel nonsmooth optimization*, J. Mach. Learn. Res., 22 (2021), pp. 1–47.
[35] F. Pedregosa, *Hyperparameter optimization with approximate gradient*, in International Conference on Machine Learning, 2016, pp. 737–746.
[36] D. Peng and X. Chen, *Computation of second-order directional stationary points for group sparse optimization*, Opt. Methods Softw., 35 (2020), pp. 348–376.
[37] A. Pinkus, *Approximation theory of the MLP model in neural networks*, Acta Numer., 8 (1999), pp. 143–195.
[38] S. Scardapane, D. Comminiello, A. Hussain and A. Uncini, *Group sparse regularization for deep neural networks*, Neurocomputing, 241 (2017), pp. 81–89.
[39] B. Shahriari, K. Swersky, Z. Wang, R.P. Wang and N. De Freitas, *Taking the human out of the loop: A review of Bayesian optimization*, Proc. IEEE, 104 (2015), pp. 148–175.
[40] W. Wen, C. Wu, Y. Wang, Y. Chen and H. Li, *Learning structured sparsity in deep neural networks*, Adv. Neural Inf. Process. Syst., 29 (2016), pp. 2074-2082.
[41] M. Xu and J.J. Ye, *A smoothing augmented Lagrangian method for solving simple bilevel programs*, Comput. Optim. Appl., 59 (2014), pp. 353–377.
[42] J. Yoon and S.J. Hwang, *Combined group and exclusive sparsity for deep neural networks*, in International Conference on Machine Learning, 2017, pp. 3958–3966.
[43] M. Yuan and Y. Lin, *Model selection and estimation in regression with grouped variables*, J. R. Stat. Soc. Ser. B-Stat. Methodol., 68 (2006), pp. 49–67.
[44] M.D. Zeiler, *Adadelta: an adaptive learning rate method*, preprint, arxiv:1212.5701, (2012).

7. APPENDIX. Note that the nonsmoothness of Φ is due to projection operator Π_Ω. For any $u \in \mathbb{R}^q$, from [4, equation (1.6)], we have

$$\Pi_\Omega(u) = \max\{\underline{u} - u, 0\} + u - \max\{u - \overline{u}, 0\},$$

where "max" has to be understood in componentwise fashion. Hence, the core of the smoothing method is to introduce a surrogate smoothing function of $\max\{\cdot, 0\}$

with some nice properties. In [4, Section 3.2], a smoothing function of $\max\{\cdot, 0\}$ is proposed as follows:

$$
\varphi_\nu(t) = \begin{cases} 0, & \text{if } t \le 0, \\ \dfrac{t^2}{2\nu}, & \text{if } 0 < t \le \nu, \\ \dfrac{1}{4}(t-\nu)^2 + t - \dfrac{1}{2}\nu, & \text{if } \nu < t \le \nu + \sqrt{\nu}, \\ -\dfrac{1}{4}(t-\nu-2\sqrt{\nu})^2 + t, & \text{if } \nu + \sqrt{\nu} < t \le \nu + 2\sqrt{\nu}, \\ t, & \text{if } t > \nu + 2\sqrt{\nu}, \end{cases} \tag{7.1}
$$

where $\nu > 0$.

Using φ_ν, a smoothing function of the projection operator Π_Ω can be defined as follows:

$$
\Psi_\nu(u) = (\psi_\nu^1(u_1), \cdots, \psi_\nu^2(u_2), \cdots, \psi_\nu^q(u_q))^\top, \tag{7.2}
$$

where, for any $t \in \mathbb{R}$,

$$
\psi_\nu^i(t) = \varphi_\nu(\underline{u}_i - t) + t - \varphi_\nu(t - \overline{u}_i),\ i = 1, 2, \cdots, q. \tag{7.3}
$$

Denote $\phi(y, u) := \nabla H(z) + \gamma(u - z) + \nabla_u Q_\mu(w; \lambda)$. Then the smoothing function of Φ can be defined as

$$
\Phi_\nu(y, u) = u - \Psi_\nu(u - \tau\phi(y, u)). \tag{7.4}
$$

Then we present some properties of smoothing functions from [4, Proposition 3.4], which are used in this paper.

LEMMA 7.1. *For any fixed $\nu \in (0, 1]$, functions ψ_ν^i in (7.2), $i = 1, 2, \cdots, q$, are continuously differentiable and satisfy the following properties:*

(i) $|\psi_\nu^i(t) - \Pi_{[\underline{u}_i, \overline{u}_i]}(t)| \le \frac{1}{2}\nu$, for any $t \in \mathbb{R}$.

(ii) $\psi_\nu^i(t) = \Pi_{[\underline{u}_i, \overline{u}_i]}(t)$ if $t \le \underline{u}_i - \nu - 2\sqrt{\nu}$ or $\underline{u}_i \le t \le \overline{u}_i$ or $t \ge \overline{u}_i + \nu + 2\sqrt{\nu}$.

(iii) $|(\psi_\nu^i)'(t)| \le 1$, for any $t \in \mathbb{R}$, where $(\psi_\nu^i)'(t)$ denotes the derivative of ψ_ν^i at t.

(iv) $|\psi_\nu^i(t^1) - \psi_\nu^i(t^2)| \le |t^1 - t^2|$, for any $t^1, t^2 \in \mathbb{R}$.

(v) There exists constant L_ν^i such that, for any $t^1, t^2 \in \mathbb{R}$, $|(\psi_\nu^i)'(t^1) - (\psi_\nu^i)'(t^2)| \le L_\nu^i |t^1 - t^2|$. Moreover, there exists $M^i > 0$ such that $L_\nu^i \le \frac{M^i}{\nu}$ for $\nu \in (0, 1]$.

Before introducing the properties of Φ_ν, we give some basic properties of the projection operator, which can be found in [16, Theorem 1.5.5].

LEMMA 7.2. *Let $\Gamma \subset \mathbb{R}^s$ be a nonempty closed convex set. Then we have the following conclusions.*

(i) For any $x_a, x_b \in \mathbb{R}^s$, $\|\Pi_\Gamma(x_a) - \Pi_\Gamma(x_b)\| \le \|x_a - x_b\|$.

(ii) For any $x_a, x_b \in \mathbb{R}^s$, $(\Pi_\Gamma(x_a) - \Pi_\Gamma(x_b))^\top (x_a - x_b) \ge \|\Pi_\Gamma(x_a) - \Pi_\Gamma(x_b)\|^2$.

LEMMA 7.3. *For any $\nu \in (0, 1]$, Φ_ν is continuously differentiable over $\mathbb{R}_+^r \times \Omega \times \Omega$, and satisfies the following porperties:*

(i) $\|\Phi_\nu(\tilde{y}, \tilde{u}) - \Phi(\tilde{y}, \tilde{u})\| \le \frac{\sqrt{q}}{2}\nu$, for any $(\tilde{y}, \tilde{u}) \in \mathbb{R}_+^r \times \Omega \times \Omega$.

(ii) $\lim\limits_{(y,u)\to(\tilde{y},\tilde{u}), \nu\downarrow 0} \operatorname{dist}(J\Phi_\nu(y, u), \partial\Phi(\tilde{y}, \tilde{u})) = 0$, for any $(\tilde{y}, \tilde{u}) \in \mathbb{R}_+^r \times \Omega \times \Omega$, where dist *denotes the distance.*

(iii) $J_u\Phi_\nu(\tilde{y}, \tilde{u})$ is invertible, for any $(\tilde{y}, \tilde{u}) \in \mathbb{R}_+^r \times \Omega \times \Omega$.

(iv) There exist constants $b^u, b^y, \tilde{b}^u_\nu > 0$ such that for any $(\tilde{y}, \tilde{u}) \in U \times \Omega \times \Omega$,

$$\|J_u\Phi_\nu(\tilde{y},\tilde{u})\| \le b^u,\ \|J_y\Phi_\nu(\tilde{y},\tilde{u})\| \le b^y,\ \|(J_u\Phi_\nu(\tilde{y},\tilde{u}))^{-1}\| \le \tilde{b}^u_\nu,$$

where U is a compact set introduced in Assumption 4.6. Moreover, for any $0 < \tilde{c} < 1$, there exists $\tilde{b}^u > 0$ such that $\|(J_u\Phi_\nu(\tilde{y},\tilde{u}))^{-1}\| \le \tilde{b}^u$ for $(\tilde{y},\tilde{u}) \in U \times \Omega \times \Omega$ and $\nu \in [\tilde{c}, 1]$.

(v) $J_u\Phi_\nu(\cdot,\cdot)$, $J_y\Phi_\nu(\cdot,\cdot)$ and $(J_u\Phi_\nu(\cdot,\cdot))^{-1}$ are Lipschitz continuous over $U\times\Omega\times\Omega$, i.e., there exist constants $\ell^u_\nu, \ell^y_\nu, \tilde{\ell}^u_\nu$ such that for any $(y^1,u^1),(y^2,u^2) \in U\times\Omega\times\Omega$, we have

$$\begin{aligned}\|J_u\Phi_\nu(y^1,u^1) - J_u\Phi_\nu(y^2,u^2)\| &\le \ell^u_\nu\|(y^1,u^1)-(y^2,u^2)\|,\\ \|J_y\Phi_\nu(y^1,u^1) - J_y\Phi_\nu(y^2,u^2)\| &\le \ell^y_\nu\|(y^1,u^1)-(y^2,u^2)\|,\\ \|(J_u\Phi_\nu(y^1,u^1))^{-1} - (J_u\Phi_\nu(y^2,u^2))^{-1}\| &\le \tilde{\ell}^u_\nu\|(y^1,u^1)-(y^2,u^2)\|.\end{aligned}$$

Moreover, there exists $M_\ell > 0$ such that $\ell^u_\nu \le \frac{M_\ell}{\nu}$ and $\ell^y_\nu \le \frac{M_\ell}{\nu}$.

(vi) Assume that $\tau\gamma < 1$ and $\gamma > L_Q$, where L_Q is the Lipschitz constant of ∇Q_μ over $U \times \Omega$. Given any $\nu \in [0,1]$ and $\tilde{y} \in U \times \Omega$,

$$\|\Phi_\nu(\tilde{y},u^1) - \Phi_\nu(\tilde{y},u^2)\| \ge \tau(\gamma - L_Q)\|u^1 - u^2\|$$

for any $u^1, u^2 \in \Omega$.

Proof. The continuous differentiability of Φ_ν is due to Lemma 7.1. Conclusion (i) is a simple consequence of Lemma 7.1(i), and conclusion (ii) is from the gradient consistency of smoothing functions.

(iii) Given any $\tilde{y} \in \mathbb{R}^r_+ \times \Omega$, we have $J_u\phi(\tilde{y},\tilde{u}) = \gamma I_q + \nabla^2_u Q_\mu(\tilde{w};\tilde{\lambda})$ for any $\tilde{u} \in \Omega$. From [16, Proposition 2.3.2], $\phi(\tilde{y},\cdot)$ is strongly monotone over Ω. By virtue of a proof similar to that of [8, Proposition 4.2] and Lemma 7.1(iii), we can obtain the conclusion.

(iv) From the compactness of $U \times \Omega \times \Omega$, we can find that $J_u\Phi_\nu$ and $J_y\Phi_\nu$ are bounded over $U \times \Omega \times \Omega$. Bounds b^u and b^y are not related to ν because of Lemma 7.1(iii). Now we prove the boundedness of $(J_u\Phi_\nu(\cdot,\cdot))^{-1}$. Let $A = J_u\Phi_\nu(\tilde{y},\tilde{u})$. Using [21, Example 5.6.6], we can prove that

$$\|A^{-1}\| = \frac{1}{\sigma_q(A)} \le \frac{(\sigma_1(A))^{q-1}}{\sigma_1(A)\cdots\sigma_q(A)} = \frac{\|A\|^{q-1}}{|\det(A)|},$$

where $\sigma_k(A)$, $k = 1,\cdots,q$, denotes the k-th largest singlular value of A. Since $U\times\Omega\times\Omega$ is compact, there exists some $(\hat{y},\hat{u}) \in U\times\Omega\times\Omega$ such that $|\det(J_u\Phi_\nu(\hat{y},\hat{u}))| \le |\det(J_u\Phi_\nu(\tilde{y},\tilde{u}))|$ for any $(\tilde{y},\tilde{u}) \in U\times\Omega\times\Omega$. Denote $\tilde{b}^u_\nu := \frac{(b^u)^{q-1}}{|\det(J_u\Phi_\nu(\hat{y},\hat{u}))|}$. Noting that b^u is the upper bound for $\|J_u\Phi_\nu(\cdot,\cdot)\|$, we have $\|(J_u\Phi_\nu(\tilde{y},\tilde{u}))^{-1}\| \le \tilde{b}^u_\nu$ for $(\tilde{y},\tilde{u}) \in U\times\Omega\times\Omega$. Then we prove the other conclusion. Let $g(\tilde{y},\tilde{u},\nu) = |\det(J_u\Phi_\nu(\tilde{y},\tilde{u}))|$ for $(\tilde{y},\tilde{u},\nu) \in U\times\Omega\times\Omega\times[\tilde{c},1]$. From the definition of $\varphi_\nu(t)$ in (7.1), we know that g is continuous over compact set $U\times\Omega\times\Omega\times[\tilde{c},1]$. So there exists $(\hat{y},\hat{u},\hat{\nu}) \in U\times\Omega\times\Omega\times[\tilde{c},1]$ such that $0 < g(\hat{y},\hat{u},\hat{\nu}) \le g(\tilde{y},\tilde{u},\nu)$ for any $(\tilde{y},\tilde{u},\tilde{\nu}) \in U\times\Omega\times\Omega\times[\tilde{c},1]$. Denote $\tilde{b}^u := \frac{(b^u)^{q-1}}{|\det(J_u\Phi_{\hat{\nu}}(\hat{y},\hat{u}))|}$. Then we have $\|(J_u\Phi_\nu(\tilde{y},\tilde{u}))^{-1}\| \le \tilde{b}^u$ for $(\tilde{y},\tilde{u}) \in U\times\Omega\times\Omega$ and $\nu \in [\tilde{c},1]$.

(v) From Lemma 7.1(iii)(iv)(v) and the compactness of $U\times\Omega\times\Omega$, we can find that $J_u\Phi_\nu$ and $J_y\Phi_\nu$ are Lipschitz continuous over $U\times\Omega\times\Omega$, and constant $M_\ell > 0$

is constructed from M^i in Lemma 7.1(v). Thus it suffices to prove that $(J_u\Phi_\nu(\cdot,\cdot))^{-1}$ is Lipschitz continuous over $U\times\Omega\times\Omega$. Let $J_u^k = J_u\Phi_\nu(y^k,u^k)$, $k=1,2$. Actually,

$$\begin{aligned}\|(J_u^1)^{-1}-(J_u^2)^{-1}\| =&\|(J_u^2)^{-1}(J_u^1-J_u^2)(J_u^1)^{-1}\|\\ \le&\|(J_u^2)^{-1}\|\|J_u^1-J_u^2\|\|(J_u^1)^{-1}\|\\ \le&(\tilde{b}_\nu^u)^2\ell_\nu^u\|(y^1,u^1)-(y^2,u^2)\|.\end{aligned}$$

Letting $\tilde{\ell}_\nu^u := (\tilde{b}_\nu^u)^2\ell_\nu^u$, we can prove the conclusion.

(vi) We firstly show that the conclusion holds with $\nu=0$. Actually, from $\Phi_0=\Phi$, we have

$$\begin{aligned}&\|\Phi(\tilde{y},u^1)-\Phi(\tilde{y},u^2)\|\\ \ge&\|u^1-u^2\|-\|\Pi_\Omega(u^1-\tau\phi(\tilde{y},u^1))-\Pi_\Omega(u^2-\tau\phi(\tilde{y},u^2))\|\\ \ge&\|u^1-u^2\|-\|(u^1-\tau\phi(\tilde{y},u^1))-(u^2-\tau\phi(\tilde{y},u^2))\|\\ \ge&\|u^1-u^2\|-(1-\tau\gamma+\tau L_Q)\|u^1-u^2\|\\ =&\tau(\gamma-L_Q)\|u^1-u^2\|,\end{aligned}$$

where the second inequality follows from Lemma 7.2(i). For the case that $\nu>0$, from Lemma 7.1(iii), note that $\|\Psi_\nu(\tilde{y},u^1)-\Psi_\nu(\tilde{y},u^2)\|\le\|u^1-u^2\|$ for any $u^1,u^2\in\mathbb{R}^q$. Hence, the conclusion for $\nu>0$ can be proved similarly. ∎