NONCONVEX NONSMOOTH MULTICOMPOSITE OPTIMIZATION AND ITS APPLICATIONS TO RECURRENT NEURAL NETWORKS

LINGZI JIN*, XIAO WANG†, AND XIAOJUN CHEN‡

Abstract. We consider a class of nonconvex nonsmooth multicomposite optimization problems where the objective function consists of a Tikhonov regularizer and a composition of multiple nonconvex nonsmooth component functions. Such optimization problems arise from tangible applications in machine learning and beyond. To define and compute its first-order and second-order d(irectional)-stationary points effectively, we first derive the closed-form expression of the tangent cone for the feasible region of its constrained reformulation. Building on this, we establish its equivalence with the corresponding constrained and ℓ_1-penalty reformulations in terms of global optimality and d-stationarity. The equivalence offers indirect methods to attain the first-order and second-order d-stationary points of the original problem in certain cases. We apply our results to the training process of recurrent neural networks (RNNs).

Key words. Multicomposite optimization, tangent cone, first-order d-stationarity, second-order d-stationarity, recurrent neural network

MSC codes. 49J52, 90B10, 90C26, 90C30

1. Introduction.

In this paper, we consider the following unconstrained nonconvex nonsmooth optimization problem

$$\text{(P)} \qquad \min_{\theta\in\mathbb{R}^n} \ \Psi(\theta) + \lambda\|\theta\|^2,$$

where $\lambda > 0$, $\|\cdot\|$ is the Euclidean norm, and the mapping $\Psi : \mathbb{R}^n \to \mathbb{R}_+$ is defined by

$$\begin{aligned} &\Psi(\theta) := g(u_1, \cdots, u_L) \\ &\text{with } u_1 := \psi_0(\theta) \text{ and } u_\ell := \psi_{\ell-1}(\theta, u_1, \cdots, u_{\ell-1}), \quad \ell = 2, \cdots, L, \end{aligned}$$

for $L+1$ continuous but possibly nonconvex nonsmooth component functions

$$\psi_{\ell-1} : \mathbb{R}^{n+\bar{N}_{\ell-1}} \to \mathbb{R}^{N_\ell}, \quad \ell = 1, \ldots, L, \quad \text{and} \quad g : \mathbb{R}^{\bar{N}_L} \to \mathbb{R}_+$$

with $\bar{N}_0 := 0$ and $\bar{N}_\ell := \sum_{j=1}^{\ell} N_j$ for all $\ell = 1, \ldots, L$. Problem (P) covers a wide range of applications in machine learning where θ refers to the network parameter, Ψ is the loss function and $\lambda\|\theta\|^2$ is the regularizer to guarantee the boundedness of the solution set [21] and alleviate the overfitting [27] for (P).

In [9], Cui et al. present a novel deterministic algorithmic framework that enables the computation of a d-stationary point of the empirical deep neural network training problem formulated as a multicomposite optimization problem. The model (P) differs from the model (2.1)-(2.2) of [9] in two aspects. The first difference is that we unify parameters $\{\theta_0, \ldots, \theta_{L-1}\}$ (corresponding to $\{z_1, \ldots, z_L\}$ in [9]) as θ since the process of selecting $\theta_{\ell-1}$ from θ can be achieved by $\psi_{\ell-1}$, which facilitates the sharing of

*Department of Applied Mathematics, Hong Kong Polytechnic University, Kowloon, Hong Kong (ling-zi.jin@connect.polyu.hk).

†School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China (wangx936@mail.sysu.edu.cn). This author is partially supported by National Natural Science Foundation of China (No 11271278).

‡Department of Applied Mathematics, Hong Kong Polytechnic University, Kowloon, Hong Kong (maxjchen@polyu.edu.hk). This author is partially supported by Hong Kong Research Grant Council PolyU15300123, JLFS/P-501/24, CAS Joint Laboratory of Applied Mathematics.

This manuscript is for review purposes only.

parameters across layers simultaneously. Secondly, we explicitly articulate the information transmission across multiple layers (i.e. dependence of $\psi_{\ell-1}$ on $u_1, \ldots, u_{\ell-2}$), which is widely used in network structure, such as RNN [13] and shortcut in Resnet. In [9] it assumes that g only depends on u_L and $\psi_{\ell-1}$ only depends on $(\theta, u_{\ell-1})$. Although Ψ in (P) can be reorganized into $\Psi(\theta) = \bar{g}(\cdot) \circ \bar{\psi}_{L-1}(\theta, \cdot) \circ \cdots \circ \bar{\psi}_0(\theta)$ with some functions $\bar{g}, \{\bar{\psi}_\ell\}$ coinciding the formulation (2.1)-(2.2) in [9] if neglecting the first difference, it can be found that the required number of auxiliary variables under this decomposition is much larger than $\bar{N}_L$. We illustrate the differences by an example of RNNs in Remark 4.1 with Figure 1. Thus, model (P) encompasses the formulation (2.1)-(2.2) presented in [9].

Directly solving (P) by SGD-type methods (SGDs) is common in computer science. However, the automatic differentiation (AD), the key of SGDs, based on chain rules fails for the subdifferential of Ψ at a nondifferentiable point θ [4]. To the best of our knowledge, existing algorithms that directly solve unconstrained nonconvex nonsmooth problem (P) with rigorous convergence analysis can be roughly separated into two groups. One combines (S)GDs with smoothing techniques aiming at (approximate) Clarke stationary points [7, 20, 31]. Another approach constructs advanced AD algorithms based on chain rules for some generalized subdifferentials. The latter further branches into two distinct paths. Along the first path Nesterov [23] utilizes the chain rule of directional derivatives to define lexicographic differentiation and evaluate lexicographic subdifferential [2, 18]. However, the nice properties of lexicographic subdifferential [18, 23] seem to be mostly applied in sensitivity analysis and have not helped to develop an algorithm converging to a stationary point defined by lexicographic subdifferential or a d-stationary point. Moreover, it is mentioned in [3] that the AD method based on lexicographic differentiation is incompatible with existing AD frameworks. Therefore, Bolte and Pauwels [4] follow a path of conservative field, which is a generalization of Clarke subdifferential. Further study establishes convergence of SGDs in the sense of conservative field stationarity, which can be improved to Clarke stationarity under certain conditions. More references can be referred to [28]. From the existing literature, directly solving (P) may only be able to find a Clarke stationary point if there is no special structure, such as weak convexity [19] and weak concavity [1].

However, in general, Clarke stationarity may be an overly lenient condition in contrast with d-stationarity [11]. On the other hand, the d-stationary points of multicomposite optimization (P) are too complicated to calculate directly (Proposition 3.4). Therefore, a more practical approach is to reformulate (P) to derive a model with easily computable d-stationary points, while establishing their relationship in terms of d-stationarity. In [10, Section 9.4.2], the equivalence between (P) with $L = 1$ and its ℓ_1-penalty form in d-stationarity is established under the premise of feasibility. In [9], a one-sided relation is obtained for simplified (P) with $L > 1$ and its ℓ_1-penalty form, which provides the algorithm for calculating d-stationary points of DNN training problem. More references that establish and utilize the relationship between the simplified (P) and its different reformulations in other kinds of stationarity can be referred to [22, 26, 27].

Apart from the above first-order optimality conditions, the second-order optimality conditions for nonsmooth optimization problems have attracted widespread interest since the 1970s [25, Chapter 13]. To avoid the concept of second-order tangent cone, Cui et al. [8] use a kind of second-order subderivative [25, 13(7)] to establish second-order conditions for minimizing twice semidifferentiable and locally Lipschitz continuous functions with polyhedral constraints [8, Proposition 2.3], and apply the

results on piecewise linear-quadratic programs. Jiang and Chen [16, Lemma 3.8] further extend the second-order necessary condition to convexly constrained optimization problems with twice semidifferentiable objective functions, and apply the results on minimax problems by using generalized directional derivatives and subderivatives. For (P) with $L = 1$ and twice semidifferentiable component functions, [10, Proposition 9.4.2] offers second-order conditions by the relation between the original problem and its ℓ_1-penalty reformulation, and the structure of the reformulation. However, the aforementioned second-order conditions are inapplicable to (P) with $L > 1$ even when g and $\{\psi_\ell\}$ are all twice semidifferentiable, since the composition of such functions may not retain this property. More references that establish second-order optimality conditions by other generalized Hessians and generalized second-order directional derivatives can be referred to commentary at the end of [25, Chapter 13].

1.1. Model reformulation. Motivated by [6, 9], we reformulate (P) as a constrained optimization problem. First we introduce auxiliary variables

$$\mathbf{u}_\ell := (u_1^\top, \ldots, u_\ell^\top)^\top \in \mathbb{R}^{\bar{N}_\ell},\ \ell = 1, \ldots, L, \text{ and an empty placeholder } \mathbf{u}_0, \tag{1.1}$$

to decompose the nested structure of Ψ, obtaining the constrained form

$$\min_z\ F(z) := g(u) + \lambda\|\theta\|^2, \ \text{ subject to } \ u_\ell = \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1}), \ell = 1, \ldots, L, \tag{P0}$$

where for brevity we denote $u := \mathbf{u}_L \in \mathbb{R}^{\bar{N}_L}$,

$$z := (\theta^\top, u^\top)^\top \in \mathbb{R}^{\bar{N}}, \text{ and } \bar{N} := n + \bar{N}_L. \tag{1.2}$$

The nonconvex nonsmooth objective function and the nonsmooth equality constraints in (P0) pose significant challenges for both theoretical analysis and numerical tractability. Therefore, (P0) will only be used as an intermediary. Denote $[L] := \{1, \ldots, L\}$. As the final reformulation, the ℓ_1-penalty form of (P0) with positive penalty parameters $\{\beta_\ell, \ell \in [L]\}$ is defined as:

$$\min_z\ \Theta(z) := F(z) + \sum_{\ell=1}^{L} \beta_\ell \|u_\ell - \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})\|_1. \tag{P1}$$

We will analyze the properties of (P), (P0) and (P1) and establish the relationship between them, which makes it realistic to attain second-order stationary points of (P).

1.2. Contribution. The contributions of this paper lie in threefold.

Firstly, we obtain a full characterization of the tangent cone of the feasible region of (P0) under directional differentiability and local Lipschitz continuity of g and $\{\psi_{\ell-1}, \ell \in [L]\}$ in Theorem 3.6. In general, it is challenging to express the tangent cone of a nonconvex feasible region [10, p525 and Remark 9.2.1]. For the nonconvex feasible region constructed by nonsmooth equality constraints in (P0), it can be verified that NNAMCQ (no nonzero abnormal multiplier constraint qualification) [29, Remark 2] holds using the method similar to [22, Lemma 6]. Based on that, a subset of its tangent cone can be expressed by a superset of its normal cone [25, Corollary 10.50] using the relations between tangent and normal cones [25, Theorems 6.26 and 6.28]. However, the closed-form of its tangent cone is still difficult to obtain solely through constraint qualifications (CQs). In contrast, we provide a closed-form expression of the tangent cone of the feasible region of (P0) by directly utilizing the pull-out structure of constraints.

Secondly, we show the equivalence between (P), (P0) and (P1) regarding d-stationary points and global minimizers, which generalizes the results in [9] and Chapter 9 of [10]. As a consequence of the equivalence between (P) and (P1), the penalty form (P1) with according algorithms [11, 30] offers an alternative way to solve the original problem (P). Furthermore, we derive a unified second-order necessary condition for nonconvex nonsmooth constrained minimization with twice directionally differentiable objective functions, which extends the results in [8, 16]. Together with the equivalence between (P), (P0) and (P1), the second-order optimality conditions for (P0) and (P1) provide second-order necessary and sufficient criteria for (P), which cover the ones proposed in [10, Proposition 9.4.2].

Thirdly, we apply our theoretical results to the minimization problem for training an Elman RNN with a single unidirectional hidden layer. The equivalence in d-stationarity of (P0-RNN) and (P1-RNN) not only generalizes the result from Theorem 2.1 of [9], but also provides the explicit thresholds for penalty parameters. Moreover, we observe that every d-stationary point of (P0-RNN) is also a second-order d-stationary point for (P0-RNN) and the same result holds for (P1-RNN) under certain conditions, which makes their second-order d-stationary points computable by the methods for DC programs [11].

1.3. Organization. The rest of this paper is organized as follows. In Section 2, we introduce some basic definitions and preliminary properties of (P), (P0) and (P1). The d-stationarity of (P), (P0), (P1) and the second-order d-stationarity of (P0), (P1) are defined in Section 3. Based on the closed-form expression of the tangent cone of the feasible region of (P0) in subsection 3.1, we establish the equivalence between (P), (P0) and (P1) in terms of global optimality and d-stationarity under certain conditions in subsection 3.2. And subsection 3.3 shows that second-order d-stationarity of (P0) and (P1) provides second-order necessary conditions for (P). Subsection 3.4 offers second-order sufficient conditions for strong local minimizers of (P) through (P1). In Section 4, we apply the general theoretical results to RNNs. Concluding remarks are given in Section 5.

1.4. Notation. In the following, we denote the set of integers and nonnegative (positive) integers as $\mathbb{Z}$ and $\mathbb{Z}_+$ ($\mathbb{Z}_{++}$) respectively. For any $m \in \mathbb{Z}_{++}$, we denote $[m] := \{1, \ldots, m\}$. The accumulative multiplication is presented by $\prod$. For any sequence $\{a_j \geq 0, j \in \mathbb{Z}_+\}$ and any $j_1, j_2 \in \mathbb{Z}_+$ with $j_1 > j_2$, denote $\sum_{j=j_1}^{j_2} a_j := 0$ and $\prod_{j=j_1}^{j_2} a_j := 1$. For any vector sequence $\{u_j, j \in \mathbb{Z}_+\}$ and any $j_1, j_2 \in \mathbb{Z}_+$ with $j_1 > j_2$, denote $(u_{j_1}, \ldots, u_{j_2})$ as an empty placeholder. For any vector a and positive integer i, $[a]_i$ refers to the ith component of a. For any two sets $A, B \subseteq \mathbb{R}^m$, denote $A + B = \{a + b \mid a \in A, b \in B\}$. Denote $\mathbb{B}(\mathbf{0}; 1) := \{z \in \mathbb{R}^{\bar{N}} \mid \|z\| \leq 1\}$. For any set $\mathcal{F} \subseteq \mathbb{R}^m$, the indicator function is defined as $\delta_{\mathcal{F}}(x) = 0$, if $x \in \mathcal{F}$, and $+\infty$, otherwise. For any $m \in \mathbb{Z}_{++}, \gamma \in \mathbb{R}$ and any function $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$, the level set is defined as $lev_{\leq \gamma} f := \{x \in \mathbb{R}^m \mid f(x) \leq \gamma\}$.

Denote the optimal solution sets of (P), (P0) and (P1) by

$$\mathcal{S} := \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \ [\Psi(\theta) + \lambda \|\theta\|^2], \quad \mathcal{S}_0 := \operatorname*{argmin}_{z \in \mathcal{F}_0} F(z), \quad \mathcal{S}_1 := \operatorname*{argmin}_{z \in \mathbb{R}^{\bar{N}}} \ \Theta(z),$$

respectively, where

$$\mathcal{F}_0 := \{z \in \mathbb{R}^{\bar{N}} \mid u_\ell = \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1}), \ell \in [L]\}. \tag{1.3}$$

2. Preliminaries. In this section, we present some preliminaries that are used in subsequent sections. Let

$$\begin{aligned} &z^0 := (\mathbf{0}^\top, (u_1^0)^\top, \ldots, (u_L^0)^\top)^\top \text{ with } u_\ell^0 := \psi_{\ell-1}(\mathbf{0}, u_1^0, \ldots, u_{\ell-1}^0),\ \ell = 1, \ldots, L, \\ &\bar{\gamma} := \Theta(z^0). \end{aligned} \tag{2.1}$$

Then we have $z^0 \in \mathcal{F}_0 \neq \emptyset$ and $\bar{\gamma} = F(z^0)$. Next, we prove that $\mathcal{S}, \mathcal{S}_0$ and $\mathcal{S}_1$ are nonempty and compact under the continuity of $g : \mathbb{R}^{\bar{N}_L} \to \mathbb{R}_+$ and $\{\psi_{\ell-1} : \mathbb{R}^{n+\bar{N}_{\ell-1}} \to \mathbb{R}^{N_\ell}, \ell \in [L]\}$. Noting the nonnegativity of g, we can obtain the following result by the level-boundedness [25, Theorem 1.9] of $(\Psi(\cdot) + \lambda\|\cdot\|^2)$ and $(F + \delta_{\mathcal{F}_0})$.

LEMMA 2.1. *The optimal solution sets $\mathcal{S}$ and $\mathcal{S}_0$ are nonempty and compact.*

In fact, it can be naturally obtained that (P) is equivalent to (P0) in global optimality. If $\bar{\theta} \in \mathcal{S}$, then $\bar{z} := (\bar{\theta}^\top, \bar{u}_1^\top, \ldots, \bar{u}_L^\top)^\top \in \mathcal{S}_0$ where $\bar{u}_\ell := \psi_{\ell-1}(\bar{\theta}, \bar{u}_1, \ldots, \bar{u}_{\ell-1})$ for all $\ell \in [L]$; conversely, if $\bar{z} := (\bar{\theta}^\top, \bar{u}_1^\top, \ldots, \bar{u}_L^\top)^\top \in \mathcal{S}_0$, then $\bar{\theta} \in \mathcal{S}$.

LEMMA 2.2. *The optimal solution set $\mathcal{S}_1$ is nonempty and compact.*

Proof. Since Θ is proper and continuous, we only need to show its level boundedness [25, Theorem 1.9]. For any $\gamma \in \mathbb{R}_+$ and any $z \in lev_{\leq\gamma}\Theta$, it follows from $g(\cdot) \geq 0$ and $\|\cdot\| \leq \|\cdot\|_1$ that

$$\|\theta\| \leq \sqrt{\gamma/\lambda}, \tag{2.2}$$

$$\|u_\ell - \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})\| \leq \gamma\beta_\ell^{-1},\ \forall \ell \in [L]. \tag{2.3}$$

Next, we will finish the proof in an inductive manner. For $\ell = 1$, it follows from (2.2)-(2.3) and the continuity of ψ_0 on $\mathbb{R}^n$ that

$$\|u_1\| \leq \|u_1 - \psi_0(\theta)\| + \|\psi_0(\theta)\| \leq \gamma\beta_1^{-1} + \max\{\psi_0(\theta) \mid \|\theta\| \leq \sqrt{\gamma\lambda^{-1}}\} < +\infty.$$

For any $\ell = 2, \ldots, L$, assume that $u_1, \ldots, u_{\ell-1}$ are bounded. Then it follows from (2.2) and (2.3) that

$$\begin{aligned} \|u_\ell\| &\leq \|u_\ell - \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})\| + \|\psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})\| \\ &\leq \gamma\beta_\ell^{-1} + \|\psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})\| < +\infty. \end{aligned}$$

Hence, u is bounded by induction. Together with (2.2) and arbitrariness of z, it implies the boundedness of $lev_{\leq\gamma}\Theta$. ∎

For the main analysis we need the following concepts of directional differentiability and local Lipschitz continuity.

DEFINITION 2.3 ((twice) directional differentiability, Definition 1.1.3 and (4.10) of [10]). *Given an open subset $\mathcal{O}$ of $\mathbb{R}^n$ and a scalar-valued function $f : \mathcal{O} \to \mathbb{R}$. The directional derivative of f at a point $x \in \mathcal{O}$ along a direction $d \in \mathbb{R}^n$ is defined as*

$$f'(x; d) := \lim_{\tau \downarrow 0} \frac{f(x + \tau d) - f(x)}{\tau}, \tag{2.4}$$

if the limit exists. The function f is directionally differentiable at x, if the limit (2.4) exists for all $d \in \mathbb{R}^n$. The second-order directional derivative of f at a point $x \in \mathcal{O}$ along a direction $d \in \mathbb{R}^n$ is defined as

$$f^{(2)}(x; d) := \lim_{\tau \downarrow 0} \frac{f(x + \tau d) - f(x) - \tau f'(x; d)}{\tau^2/2}, \tag{2.5}$$

if the limit and the one for (2.4) *exist. The function f is twice directionally differentiable at x, if the limits* (2.4) *and* (2.5) *exist for all $d \in \mathbb{R}^n$.*

For a vector-valued function $f : \mathcal{O} \to \mathbb{R}^m$ with component functions $\{f_i : \mathcal{O} \to \mathbb{R}, i \in [m]\}$, the directional derivative $f'(x;d)$ is defined as

$$f'(x;d) := (f_1'(x;d), \ldots, f_m'(x;d))^\top,$$

if $f_i'(x;d), i \in [m]$ exist. Furthermore, if $f_i^{(2)}(x;d), i \in [m]$ exist, then the second-order directional derivative $f^{(2)}(x;d)$ is defined as

$$f^{(2)}(x;d) := (f_1^{(2)}(x;d), \ldots, f_m^{(2)}(x;d))^\top.$$

Function f is (twice) directionally differentiable at x, if all of its component functions are (twice) directionally differentiable at x.

DEFINITION 2.4 (local Lipschitz continuity). *For any function $f : \mathcal{O} \subseteq \mathbb{R}^n \to \mathbb{R}^m$, we say f is locally Lipschitz continuous near $x \in \mathcal{O}$, if there exists a neighborhood X of x and $K \geq 0$ such that $\|f(x_1) - f(x_2)\| \leq K\|x_1 - x_2\|$ for all $x_1, x_2 \in X$. And we say f is locally Lipschitz continuous, if f is locally Lipschitz continuous near every point in its domain $\mathcal{O}$.*

If f is directionally differentiable at x and locally Lipschitz continuous near x with modulus $K \geq 0$, then it follows from Definitions 2.3 and 2.4 that for all $d \in \mathbb{R}^n$,

$$\|f'(x;d)\| = \left\| \lim_{\tau \downarrow 0} \frac{f(x+\tau d) - f(x)}{\tau} \right\| \leq \lim_{\tau \downarrow 0} \frac{K\tau\|d\|}{\tau} = K\|d\| < \infty. \tag{2.6}$$

Hence, for all $d, \bar{d} \in \mathbb{R}^n$, $\|f'(x;d) - f'(x;\bar{d})\|$ is well-defined and we can similarly obtain that

$$\|f'(x;d) - f'(x;\bar{d})\| \leq K\|d - \bar{d}\|, \forall d, \bar{d} \in \mathbb{R}^n. \tag{2.7}$$

The remaining analysis will be based on the following assumptions about the directional differentiability and local Lipschitz continuity.

ASSUMPTION 1. *Functions g and $\{\psi_{\ell-1}, \ell \in [L]\}$ are directionally differentiable on $\mathbb{R}^{\bar{N}_L}$ and $\mathbb{R}^{n+\bar{N}_{\ell-1}}, \ell \in [L]$ respectively. Functions g and $\{\psi_{\ell-1}, \ell \in [L]\}$ are locally Lipschitz continuous.*

According to Lemma 2.2, $lev_{\leq \bar{\gamma}}\Theta$ with $\bar{\gamma} := \Theta(z^0)$ defined in (2.1) is nonempty and compact. Under Assumption 1, it follows from the compactness of $lev_{\leq \bar{\gamma}}\Theta$ that there exist $K_g > 0$ and $\{K_\ell > 0, \ell \in [L-1]\}$ such that

$$|g(u) - g(\bar{u})| \leq K_g\|u - \bar{u}\|, \tag{2.8}$$

$$\|\psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1}) - \psi_{\ell-1}(\theta, \bar{\mathbf{u}}_{\ell-1})\| \leq K_{\ell-1}\|\mathbf{u}_{\ell-1} - \bar{\mathbf{u}}_{\ell-1}\|, \ \ell = 2, \ldots, L \tag{2.9}$$

for all $(\theta^\top, \mathbf{u}^\top)^\top$, $(\theta^\top, \bar{\mathbf{u}}^\top)^\top \in (lev_{\leq \bar{\gamma}}\Theta + \epsilon\mathbb{B}(\mathbf{0}; 1))$, where the positive real number ϵ is sufficiently small. In (2.9), the two terms at the left-hand side are consistent in the component θ, since the subsequent analysis only needs the Lipschitz continuity moduli of $\{\psi_\ell, \ell \in [L-1]\}$ in component u. Besides, it should be noted that K_g and $\{K_\ell, \ell \in [L-1]\}$ are non-increasing as $\{\beta_\ell, \ell \in [L]\}$ increase since functions $g, \{\psi_\ell, \ell \in [L-1]\}$ and $\bar{\gamma} = F(z^0)$ are independent of penalty parameters and for any z, $\Theta(z)$ is non-decreasing as $\{\beta_\ell, \ell \in [L]\}$ increase. Furthermore, in Example 3.13 and Section 4 we will show how to estimate K_g and $\{K_\ell, \ell \in [L-1]\}$ for specific applications.

Remark 2.5. In fact, z^0 and $\bar{\gamma}$ can be replaced by any feasible point of (P0) and the value of Θ at that point. The replacement will not affect any theoretical results in this paper. We choose z^0 and $\bar{\gamma}$ defined in (2.1) since $\mathbf{0}$ is usually not a good candidate in data fitting. Thus, the requirement $z \in lev_{\leq\bar{\gamma}}\Theta$ can be regarded as a mild condition.

For clarity, break the direction $d \in \mathbb{R}^{\bar{N}}$ according to the blocks of variable z as follows

$$\begin{aligned} &d = (d_\theta^\top, d_u^\top)^\top, \text{ where} \\ &d_u = ((d_{u_1})^\top, \ldots, (d_{u_L})^\top)^\top \text{ with } d_{u_\ell} \in \mathbb{R}^{N_\ell}, \ell \in [L]. \end{aligned} \tag{2.10}$$

Then under Assumption 1, it follows from (2.6)-(2.7) and (2.8)-(2.9) that for any $z \in lev_{\leq\bar{\gamma}}\Theta$, and any $d, \bar{d} \in \mathbb{R}^{\bar{N}}$ with $d_\theta = \bar{d}_\theta$,

$$\begin{aligned} &|g'(u; d_u) - g'(u; \bar{d}_u)| \leq K_g \|d_u - \bar{d}_u\|, \\ &\|\psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}}) - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; \bar{d}_\theta, \bar{d}_{\mathbf{u}_{\ell-1}})\| \\ &\quad \leq K_{\ell-1}\|d_{\mathbf{u}_{\ell-1}} - \bar{d}_{\mathbf{u}_{\ell-1}}\|, \ell = 2, \ldots, L, \end{aligned} \tag{2.11}$$

where $d_{\mathbf{u}_{\ell-1}} := ((d_{u_1})^\top, \ldots, (d_{u_{\ell-1}})^\top)^\top$ for all $\ell \in [L]$.

Apart from Assumption 1, the analysis concerning second-order necessary conditions also requires the following assumption about the twice directional differentiability.

ASSUMPTION 2. *Functions g and $\{\psi_{\ell-1}, \ell \in [L]\}$ are twice directionally differentiable on $\mathbb{R}^{\bar{N}_L}$ and $\mathbb{R}^{n+\bar{N}_{\ell-1}}, \ell \in [L]$ respectively.*

We use the definitions of tangent cone [25, Definition 6.1] and radial cone [5].

DEFINITION 2.6 (tangent cone and radial cone). *The tangent cone of a set $\mathcal{F} \subseteq \mathbb{R}^m$ at any point $x \in \mathcal{F}$ is defined as*

$$\mathcal{T}_{\mathcal{F}}(x) := \{d \in \mathbb{R}^m \mid \exists x^k \to x \text{ with } x^k \in \mathcal{F} \text{ and } \tau_k \downarrow 0, \text{ such that } \frac{x^k - x}{\tau_k} \to d\}.$$

The radial cone of a set $\mathcal{F} \subseteq \mathbb{R}^m$ at any point $x \in \mathcal{F}$ is defined as

$$P_{\mathcal{F}}(x) := \{d \in \mathbb{R}^m \mid \exists \tau_k \downarrow 0 \text{ such that } x + \tau_k d \in \mathcal{F}\}.$$

Then it can be observed that $P_{\mathcal{F}}(x) \subseteq \mathcal{T}_{\mathcal{F}}(x)$. When $\mathcal{F}$ is convex, $P_{\mathcal{F}}(x)$ coincides with $\mathcal{T}^\circ_{\mathcal{F}}(x)$ used in [16], and it further equals to $\mathcal{T}_{\mathcal{F}}(x)$ when $\mathcal{F}$ is polyhedral.

3. Optimality and stationarity. This section will establish the relationship between (P), (P0) and (P1) in global optimality and (second-order) d-stationarity, and discuss the byproducts regarding second-order sufficient conditions. The d-stationary points are defined by the necessary tangent condition outlined at the end of [25, Chapter 8.C] without proof. And the second-order d-stationarity extends the second-order necessary condition in [16, Lemma 3.8] from a twice semidifferentiable objective function with convex constraints to a twice directionally differentiable objective function with general constraints. We provide a detailed proof for the necessity of (second-order) d-stationarity under the assumptions of the nonemptiness of solution sets, twice directional differentiability and local Lipschitz continuity of objective functions as follows.

LEMMA 3.1. *Assume that* $\operatorname{argmin}_{x\in\mathcal{F}} f(x) \neq \emptyset$ *with* $f : \mathbb{R}^m \to \mathbb{R}$. *If* $\bar{x} \in \mathcal{F}$ *is a local minimizer of* $\min_{x\in\mathcal{F}} f(x)$, f *is directionally differentiable at* $\bar{x}$ *and locally Lipschitz continuous near* $\bar{x}$, *then* $f'(\bar{x}; d) \geq 0$ *for any* $d \in \mathcal{T}_{\mathcal{F}}(\bar{x})$. *Moreover, if* f *is twice directionally differentiable at* $\bar{x}$, *then* $f^{(2)}(\bar{x}; d) \geq 0$ *for all* $d \in P_{\mathcal{F}}(\bar{x}) \cap \{d \in \mathbb{R}^m \mid f'(\bar{x}; d) = 0\}$.

Proof. Firstly, the local optimality implies that $\bar{x} \in \mathcal{F}$ is a local minimizer of $(f + \delta_{\mathcal{F}})(x)$ over $\mathbb{R}^m$. Hence it follows from [25, Theorem 10.1] that

$$\liminf_{\tau\downarrow 0, d'\to d} \frac{(f + \delta_{\mathcal{F}})(\bar{x} + \tau d') - (f + \delta_{\mathcal{F}})(\bar{x})}{\tau} \geq 0, \forall d \in \mathbb{R}^m.$$

Then the remainder is to show that for any $d \in \mathcal{T}_{\mathcal{F}}(\bar{x})$,

$$\liminf_{\tau\downarrow 0, d'\to d} \frac{(f + \delta_{\mathcal{F}})(\bar{x} + \tau d') - (f + \delta_{\mathcal{F}})(\bar{x})}{\tau} \leq f'(\bar{x}; d).$$

For any $d \in \mathcal{T}_{\mathcal{F}}(\bar{x})$, it follows from the definition of tangent cone that there exist $x^k \to \bar{x}, x^k \in \mathcal{F}$ and $\tau_k \downarrow 0$ such that $d^k := \frac{x^k - \bar{x}}{\tau_k} \to d$ as $k \to \infty$, which implies

$$\begin{aligned}
&\liminf_{d'\to d, \tau\downarrow 0} \frac{f(\bar{x} + \tau d') - f(\bar{x}) + \delta_{\mathcal{F}}(\bar{x} + \tau d') - \delta_{\mathcal{F}}(\bar{x})}{\tau} \\
&\leq \liminf_{k\to\infty} \frac{f(\bar{x} + \tau_k d^k) - f(\bar{x}) + \delta_{\mathcal{F}}(\bar{x} + \tau_k d^k) - \delta_{\mathcal{F}}(\bar{x})}{\tau_k} \\
&= \liminf_{k\to\infty} \frac{f(\bar{x} + \tau_k d^k) - f(\bar{x})}{\tau_k} \\
&= f'(\bar{x}; d),
\end{aligned}$$

where the first equality uses $\bar{x} + \tau_k d^k = x^k \in \mathcal{F}$ and $\bar{x} \in \mathcal{F}$, the last equality comes from

$$\begin{aligned}
&\lim_{d'\to d, \tau\downarrow 0} \frac{f(\bar{x} + \tau d') - f(\bar{x})}{\tau} \\
&= \lim_{d'\to d, \tau\downarrow 0} \frac{f(\bar{x} + \tau d) - f(\bar{x})}{\tau} + \lim_{d'\to d, \tau\downarrow 0} \frac{f(\bar{x} + \tau d') - f(\bar{x} + \tau d)}{\tau} \\
&= \lim_{\tau\downarrow 0} \frac{f(\bar{x} + \tau d) - f(\bar{x})}{\tau} + 0 = f'(\bar{x}; d)
\end{aligned}$$

by the fact that f is directionally differentiable at $\bar{x}$ and locally Lipschitz continuous near $\bar{x}$.

For the second-order optimality condition, since $\bar{x} \in \mathcal{F}$ is a local minimizer of $\min_{x\in\mathbb{R}^m}(f + \delta_{\mathcal{F}})(x)$, it follows from [25, Theorem 13.24 (a)] that for all $d \in \mathbb{R}^m$,

$$0 \leq \liminf_{\tau\downarrow 0, d'\to d} \frac{(f + \delta_{\mathcal{F}})(\bar{x} + \tau d') - (f + \delta_{\mathcal{F}})(\bar{x})}{\tau^2/2},$$

which implies that for all $d \in P_{\mathcal{F}}(\bar{x}) \cap \{d \in \mathbb{R}^m \mid f'(\bar{x}; d) = 0\}$,

$$\begin{aligned}
0 &\leq \liminf_{\tau\downarrow 0} \frac{(f + \delta_{\mathcal{F}})(\bar{x} + \tau d) - (f + \delta_{\mathcal{F}})(\bar{x})}{\tau^2/2} \leq \liminf_{k\to\infty} \frac{f(\bar{x} + \tau_k d) - f(\bar{x})}{\tau_k^2/2} \\
&= \liminf_{k\to\infty} \frac{f(\bar{x} + \tau_k d) - f(\bar{x}) - \tau_k f'(\bar{x}; d)}{\tau_k^2/2} = f^{(2)}(\bar{x}; d),
\end{aligned}$$

where the second inequality uses the definition of $P_{\mathcal{F}}(\bar{x})$, the first equality holds due to $f'(\bar{x};d)=0$, and the last equality comes from the twice directional differentiability of f at $\bar{x}$ and $\tau_k \downarrow 0$. ∎

The first-order condition for convexly constrained optimization problems with a semidifferentiable objective function is established in [16, Lemma 3.8]. Lemma 3.1 extends this condition to nonconvexly constrained optimization problems with a directionally differentiable objective function. Actually, the two first-order conditions are completely identical in form since the directional derivative equals to the subderivative used in [16, Lemma 3.8] under directional differentiability and local Lipschitz continuity. However, for the second-order condition, the nonconvexity of the feasible region and non-twice-semidifferentiability of the objective function entail us to narrow the range of directions from $\mathcal{T}_{\mathcal{F}}^{\circ}(x)$ in [16] to $P_{\mathcal{F}}(x)$ and relax the nonnegativity of second-order subderivatives to the nonnegativity of second-order directional derivatives. Based on Lemma 3.1, we can define unified first-order and second-order necessary conditions as follows.

DEFINITION 3.2 (second-order d-stationary point). *For any $f:\mathbb{R}^m\to\mathbb{R}$ that is directionally differentiable on $\mathbb{R}^m$ and locally Lipschitz continuous, and any $\mathcal{F}\subseteq\mathbb{R}^m$ such that $\emptyset \neq \operatorname{argmin}_{x\in\mathcal{F}} f(x)$, we call $\bar{x}\in\mathcal{F}$ a d(irectional)-stationary point of $\min_{x\in\mathcal{F}} f(x)$ if $f'(\bar{x};d)\geq 0$ for all $d\in\mathcal{T}_{\mathcal{F}}(\bar{x})$. And we further call $\bar{x}$ a second-order d-stationary point of $\min_{x\in\mathcal{F}} f(x)$ if f is also twice directionally differentiable at $\bar{x}$ with $f^{(2)}(\bar{x};d)\geq 0$ for all $d\in P_{\mathcal{F}}(\bar{x})\cap\{d\in\mathbb{R}^m \mid f'(\bar{x};d)=0\}$.*

Example 3.3. To illustrate Definition 3.2, we consider $\min_{x\in\mathcal{F}} f(x)$ with $f(x)=\max\{-1,x_1x_2\}+0.1\|x\|^2$ and $\mathcal{F}=[-1,1]^2$. This example has only three first-order d-stationary points $(0,0)^\top,(-1,1)^\top$ and $(1,-1)^\top$, while f is differentiable at the first point, but not differentiable at the other two points. At $\bar{x}:=(0,0)^\top$, $f'(\bar{x};d)=\nabla f(\bar{x})^\top d\equiv 0$ for all $d\in\mathcal{T}_{\mathcal{F}}(\bar{x})=\mathbb{R}^2$, whereas $f^{(2)}(\bar{x};d)=2d_1d_2+0.2\|d\|^2<0$ for any $d:=(d_1,d_2)^\top\in\{d\in P_{\mathcal{F}}(\bar{x})\mid f'(\bar{x};d)=0\}=\mathbb{R}^2$ with $d_2=-d_1\neq 0$. Hence $(0,0)^\top$ is not a second-order d-stationary point, and thus not a local minimizer.

In contrast, at $\bar{x}:=(-1,1)^\top$, $f'(\bar{x};d)=(d_1-d_2)-0.2(d_1-d_2)\geq 0$ for all $d\in\mathcal{T}_{\mathcal{F}}(\bar{x})=\{(d_1,d_2)^\top \mid d_1\geq 0,d_2\leq 0\}$, and $f^{(2)}(\bar{x};d)=0\geq 0$ for any $d\in\{d\in P_{\mathcal{F}}(\bar{x})\mid f'(\bar{x};d)=0\}=\{d\in\mathcal{T}_{\mathcal{F}}(\bar{x})\mid f'(\bar{x};d)=0\}=\{\mathbf{0}\}$. Hence $(-1,1)^\top$ is a second-order d-stationary point. Similarly, we can verify that $(1,-1)^\top$ is also a second-order d-stationary point. From the boundedness of the feasible set $\mathcal{F}$ and the continuity of the objective function f, the optimal solution set of this example is nonempty. Since $f((-1,1)^\top)=f((1,-1)^\top)$, the two points are optimal solutions.

Since the local Lipschitz continuity of Ψ, F and Θ naturally holds under Assumption 1, the corresponding d-stationary points can be defined once we have checked their directional differentiability.

PROPOSITION 3.4. *Under Assumption 1, Ψ is directionally differentiable on $\mathbb{R}^n$, and the directional derivative of the objective function of* (P) *along any $d_\theta\in\mathbb{R}^n$ is*

$$g'\left(u_1,\ldots,u_L;d_{u_1},\ldots,d_{u_L}\right)+2\lambda\theta^\top d_\theta, \tag{3.1}$$

where $u_\ell:=\psi_{\ell-1}(\theta,u_1,\ldots,u_{\ell-1})$ for all $\ell\in[L]$, and $d_{u_\ell}:=\psi'_{\ell-1}(\theta,u_1,\ldots,u_{\ell-1};d_\theta,d_{u_1},\ldots,d_{u_{\ell-1}})$ for all $\ell\in[L]$; F and Θ are directionally differentiable on $\mathbb{R}^{\bar{N}}$, and for any direction $d\in\mathbb{R}^{\bar{N}}$ defined in (2.10),

$$F'(z;d)=g'(u;d_u)+2\lambda\theta^\top d_\theta, \tag{3.2}$$

$$\Theta'(z;d) = F'(z;d) + \sum_{\ell=1}^{L} \beta_\ell \left(\sum_{i \in I_+^\ell(z)} [d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i \right. \tag{3.3}$$

$$- \sum_{i \in I_-^\ell(z)} [d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i$$

$$\left. + \sum_{i \in I_0^\ell(z)} \left| [d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i \right| \right),$$

where for all $\ell \in [L]$,

$$I_+^\ell(z) := \{i \in [N_\ell] \mid [u_\ell - \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})]_i > 0\},$$
$$I_-^\ell(z) := \{i \in [N_\ell] \mid [u_\ell - \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})]_i < 0\},$$
$$I_0^\ell(z) := [N_\ell] \backslash \left(I_+^\ell(z) \cup I_-^\ell(z) \right).$$

Proof. Firstly, applying [10, Proposition 4.1.2] sequentially on $\psi_1(\cdot, \psi_0(\cdot)), \ldots,$ $\psi_{L-1}(\cdot, \psi_0(\cdot), \psi_1(\cdot, \psi_0(\cdot)), \cdots)$ and Ψ with directionally differentiable and locally Lipschitz continuous $\{\psi_{\ell-1}, \ell \in [L]\}$ and g, we can obtain

$$\Psi'(\theta; d_\theta) = g' \left(\psi_0(\theta), \quad \psi_1(\theta, \psi_0(\theta)), \ldots; \right.$$
$$\left. \psi'_0(\theta; d_\theta), \psi'_1(\theta, \psi_0(\theta); d_\theta, \psi'_0(\theta; d_\theta)), \ldots \right),$$

which can be reorganized as (3.1) with $\{u_\ell, d_{u_\ell}, \ell \in [L]\}$ defined as above. Then the result about F can be directly obtained from Assumption 1 and Definition 2.3. And for Θ, it is sufficient to show the directional differentiability of each penalty term according to Definition 2.3, which can be obtained by the directional differentiability and local Lipschitz continuity of $\|\cdot\|_1$ and $\{\psi_{\ell-1}, \ell \in [L]\}$ [10, Proposition 4.1.2]. ∎

Together with the nonemptiness of optimal solution sets $\mathcal{S}$, $\mathcal{S}_0$ and $\mathcal{S}_1$ and Definition 3.2, it implies that under Assumption 1,

- $\mathcal{D} := \{\theta \in \mathbb{R}^n \mid [\Psi(\cdot) + \lambda\|\cdot\|^2]'(\theta; d_\theta) \geq 0 \text{ for all } d_\theta \in \mathbb{R}^n\}$ is the set of d-stationary points of (P);
- $\mathcal{D}_0 := \{z \in \mathbb{R}^{\bar{N}} \mid z \in \mathcal{F}_0 \text{ and } F'(z;d) \geq 0 \text{ for all } d \in \mathcal{T}_{\mathcal{F}_0}(z)\}$ is the set of d-stationary points of (P0);
- $\mathcal{D}_1 := \{z \in \mathbb{R}^{\bar{N}} \mid \Theta'(z;d) \geq 0 \text{ for all } d \in \mathbb{R}^{\bar{N}}\}$ is the set of d-stationary points of (P1).

Although the d-stationary point of (P) can be defined as above, the complicated nested structure in Ψ' makes it challenging to compute. Notably, through (P0), we can express its d-stationarity more clearly and further establish (P)'s relationship with (P1) in Sections 3.1 and 3.2, which facilitates computation.

And for the second-order necessary conditions of (P0) and (P1), the twice directional differentiability of F and Θ can be verified under Assumptions 1 and 2.

PROPOSITION 3.5. *Under Assumptions 1 and 2, F and Θ are twice directionally*

differentiable on $\mathbb{R}^{\bar{N}}$ *and for any direction* $d \in \mathbb{R}^{\bar{N}}$ *defined in* (2.10),

$$F^{(2)}(z;d) = g^{(2)}(u;d_u) + 2\lambda\|d_\theta\|^2,$$

$$\Theta^{(2)}(z;d) = F^{(2)}(z;d) + \sum_{\ell=1}^{L} \beta_\ell \left(- \sum_{i \in I_+^\ell(z) \cup I_{0,+}^\ell(z;d)} [\psi_{\ell-1}^{(2)}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i \right.$$

$$+ \sum_{i \in I_-^\ell(z) \cup I_{0,-}^\ell(z;d)} [\psi_{\ell-1}^{(2)}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i$$

$$\left. + \sum_{i \in I_{0,0}^\ell(z;d)} \left| [\psi_{\ell-1}^{(2)}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i \right| \right),$$

where for all $\ell \in [L]$, $I_+^\ell(z), I_-^\ell(z)$ *and* $I_0^\ell(z)$ *are defined as in Proposition 3.4 and*

$$I_{0,+}^\ell(z;d) := \{i \in I_0^\ell(z) \mid [d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i > 0\},$$
$$I_{0,-}^\ell(z;d) := \{i \in I_0^\ell(z) \mid [d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i < 0\},$$
$$I_{0,0}^\ell(z;d) := I_0^\ell(z) \backslash \left(I_{0,+}^\ell(z;d) \cup I_{0,-}^\ell(z;d) \right).$$

Proof. The result about F can be directly obtained from Assumptions 1, 2 and Definition 2.3. And for Θ, it is sufficient to show the twice directional differentiability of each penalty term according to Definition 2.3, which can be obtained by the twice directional differentiability and local Lipschitz continuity of $\{\psi_{\ell-1}, \ell \in [L]\}$ and the max-structure of $|\cdot|$ (i.e. for any $x \in \mathbb{R}, |x| = \max\{x, -x\}$) [10, Example 4.2.1]. ∎

Together with Proposition 3.4, it implies that under Assumptions 1 and 2,

- $\mathcal{SD}_0 := \{z \in \mathcal{D}_0 \mid F^{(2)}(z;d) \geq 0$ for all $d \in P_{\mathcal{F}_0}(z)$ satisfying $F'(z;d) = 0\}$ is the set of second-order d-stationary points of (P0);
- $\mathcal{SD}_1 := \{z \in \mathcal{D}_1 \mid \Theta^{(2)}(z;d) \geq 0$ for all $d \in \mathbb{R}^{\bar{N}}$ satisfying $\Theta'(z;d) = 0\}$ is the set of second-order d-stationary points of (P1).

Although it seems immature to define the second-order d-stationary point of (P) under Assumptions 1 and 2, we could next see how the sets $\mathcal{SD}_0, \mathcal{SD}_1$ help to provide second-order necessary conditions for (P).

3.1. Closed-form of $\mathcal{T}_{\mathcal{F}_0}$. Here we give the closed-form of the tangent cone of $\mathcal{F}_0$ in (1.3) based on directional derivatives of constraints of (P0), which plays an important role in the following subsections.

THEOREM 3.6. *Under Assumption 1, it holds that*

$$\mathcal{T}_{\mathcal{F}_0}(z) = \{d \in \mathbb{R}^{\bar{N}} \mid d_{u_\ell} = \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}}), \ell \in [L]\},$$

for any feasible point $z \in \mathcal{F}_0$.

Proof. We firstly prove the one-sided inclusion

$$\mathcal{T}_{\mathcal{F}_0}(z) \subseteq \{d \in \mathbb{R}^{\bar{N}} \mid d_{u_\ell} = \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}}), \ell \in [L]\}.$$

According to Definition 2.6, for any $d \in \mathcal{T}_{\mathcal{F}_0}(z)$, there exists a sequence $\{z^k \in \mathcal{F}_0, k \in \mathbb{Z}_{++}\}$ converging to z and a sequence $\tau_k \downarrow 0$ such that $d = \lim_{k\to\infty} \frac{z^k - z}{\tau_k}$. Denoting $d^k := \frac{z^k - z}{\tau_k}$, we have $z^k = z + \tau_k d^k$ for all k and $d^k \to d$ as $k \to \infty$. For $\ell = 1$, it

follows from the definition of ψ_0', $\tau_k \downarrow 0$, local Lipschitz continuity of ψ_0 and $d_\theta^k \to d_\theta$ that

$$\begin{aligned} d_{u_1} - \psi_0'(\theta; d_\theta) &= \lim_{k\to\infty} \left[\frac{u_1^k - u_1}{\tau_k} - \frac{\psi_0(\theta + \tau_k d_\theta^k) - \psi_0(\theta)}{\tau_k} \right] \\ &= \lim_{k\to\infty} \frac{[u_1^k - \psi_0(\theta^k)] - [u_1 - \psi_0(\theta)]}{\tau_k} = 0, \end{aligned}$$

where the second equality uses $\theta^k = \theta + \tau_k d_\theta^k$, the last equality comes from $z^k, z \in \mathcal{F}_0$. For $\ell = 2, \ldots, L$, it follows from the definition of $\psi_{\ell-1}'$, $\tau_k \downarrow 0$, local Lipschitz continuity of $\psi_{\ell-1}$ and $(d_\theta^k, d_{\mathbf{u}_{\ell-1}}^k) \to (d_\theta, d_{\mathbf{u}_{\ell-1}})$ that

$$\begin{aligned} & d_{u_\ell} - \psi_{\ell-1}'(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}}) \\ &= \lim_{k\to\infty} \left[\frac{u_\ell^k - u_\ell}{\tau_k} - \frac{\psi_{\ell-1}(\theta + \tau_k d_\theta^k, \mathbf{u}_{\ell-1} + \tau_k d_{\mathbf{u}_{\ell-1}}^k) - \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})}{\tau_k} \right] \\ &= \lim_{k\to\infty} \frac{[u_\ell^k - \psi_{\ell-1}(\theta^k, \mathbf{u}_{\ell-1}^k)] - [u_\ell - \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})]}{\tau_k} = 0, \end{aligned}$$

where the second equality holds due to $z^k = z + \tau_k d^k$, the last equality comes from $z^k, z \in \mathcal{F}_0$.

Next we deduce the reverse inclusion

$$\mathcal{T}_{\mathcal{F}_0}(z) \supseteq \{d \in \mathbb{R}^{\bar{N}} \mid d_{u_\ell} = \psi_{\ell-1}'(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}}), \ell \in [L]\}.$$

By Definition 2.6, it is equivalent to show, for any $d \in \mathbb{R}^{\bar{N}}$ satisfying

$$d_{u_\ell} = \psi_{\ell-1}'(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}}),\ \ell \in [L], \tag{3.4}$$

there exist sequences $\{\tau_k \downarrow 0\}$ and $\{d^k \to d\}$ such that $\{z + \tau_k d^k\} \subseteq \mathcal{F}_0$. For any d satisfying (3.4) and any decreasing sequence $\{\tau_k \downarrow 0\}$, define

$$d_\theta^k := d_\theta,\ d_{u_1}^k := \frac{\psi_0(\theta + \tau_k d_\theta^k) - \psi_0(\theta)}{\tau_k},\ \text{and}$$

$$d_{u_\ell}^k := \frac{\psi_{\ell-1}(\theta + \tau_k d_\theta^k, \mathbf{u}_{\ell-1} + \tau_k d_{\mathbf{u}_{\ell-1}}^k) - \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})}{\tau_k}\ \text{in the order of } \ell = 2, \ldots, L,$$

for all k. We prove that $d^k \to d$ and $z^k := (z + \tau_k d^k) \in \mathcal{F}_0$ for all k. Firstly, $\lim_{k\to\infty} d_\theta^k = d_\theta$. Then it follows from the definition of $d_{u_1}^k$, $d_\theta^k = d_\theta$, $\tau_k \downarrow 0$ and the directional differentiability of ψ_0 that $\lim_{k\to\infty} d_{u_1}^k = \psi_0'(\theta; d_\theta) = d_{u_1}$, where the last equality holds due to (3.4). For any $\ell = 2, \ldots, L$, assume that $d_{\mathbf{u}_{\ell-1}}^k \to d_{\mathbf{u}_{\ell-1}}$ has been verified. Then it follows from the definition of $d_{u_\ell}^k$ and local Lipschitz continuity of $\psi_{\ell-1}$ that

$$\begin{aligned} \lim_{k\to\infty} d_{u_\ell}^k &= \lim_{k\to\infty} \frac{\psi_{\ell-1}(\theta + \tau_k d_\theta, \mathbf{u}_{\ell-1} + \tau_k d_{\mathbf{u}_{\ell-1}}) - \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})}{\tau_k} \\ &= \psi_{\ell-1}'(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}}) = d_{u_\ell}, \end{aligned}$$

where the second equality comes from $\tau_k \downarrow 0$ and directional differentiability of $\psi_{\ell-1}$, the last equality holds due to (3.4). By induction and (2.10), we obtain that $d^k \to d$. From $z^k := (z + \tau_k d^k)$ and (1.2), (2.10), we first have

$$\theta^k := \theta + \tau_k d_\theta^k,\ \text{and}\ u_\ell^k = u_\ell + \tau_k d_{u_\ell}^k,\ \forall \ell \in [L]. \tag{3.5}$$

For $\ell = 1$, (3.5) and the definition of $d^k_{u_1}$ imply that $u^k_1 = u_1 + \psi_0(\theta^k) - \psi_0(\theta) = \psi_0(\theta^k)$, where the last equality follows from $z \in \mathcal{F}_0$. For $\ell = 2, \ldots, L$, (3.5) and the definition of $d^k_{u_\ell}$ imply that $u^k_\ell = u_\ell + \psi_{\ell-1}(\theta^k, \mathbf{u}^k_{\ell-1}) - \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1}) = \psi_{\ell-1}(\theta^k, \mathbf{u}^k_{\ell-1})$, where the last equality follows from $z \in \mathcal{F}_0$. Combining the results for all $\ell \in [L]$, we obtain that $z^k \in \mathcal{F}_0$ for all k. ∎

It is noteworthy that Theorem 3.6 provides the expression for $\mathcal{T}_{\mathcal{F}_0}(z)$, where $\mathcal{F}_0$ is defined by nonsmooth constraints. In general, such an expression for the tangent cone of the feasible region is only achievable for smooth constraints under the Linear Independence Constraint Qualification (LICQ) [24, Lemma 12.2]. The one-sided inclusion for $\mathcal{T}_{\mathcal{F}_0}(z)$ can only guarantee one-sided implication between (P0) and (P1) [10, Theorem 9.2.1 and Remark 9.2.1], while the closed-form in Theorem 3.6 can guarantee the equivalence (see Theorem 3.9). If we use constraint qualifications for nonsmooth constraints named NNAMCQ [22] and relations between tangent and normal cones [25, Theorems 6.26 and 6.28], we can only obtain a subset of $\mathcal{T}_{\mathcal{F}_0}(z)$ presented by $\{\psi'_{\ell-1}, \ell \in [L]\}$, which fails to imply the full characterization of $\mathcal{T}_{\mathcal{F}_0}(z)$ in certain cases such as $\mathcal{T}_{\mathcal{F}^{RNN}_0}(z)$ in Section 4. Noting that $P_{\mathcal{F}_0}(z) \subseteq \mathcal{T}_{\mathcal{F}_0}(z)$ for any $z \in \mathcal{F}_0$, Theorem 3.6 also provides the expression of a superset of $P_{\mathcal{F}_0}(z)$, which helps to obtain its closed-form in certain cases (see Section 4).

3.2. Equivalence in optimality and d-stationarity. Here we show the equivalence of (P), (P0) and (P1) in global optimality and d-stationarity. Firstly, (P) and (P0) are equivalent in global optimality as we discussed after Lemma 2.1. Similarly, according to Proposition 3.4 and Theorem 3.6, (P) and (P0) are equivalent in d-stationarity when neglecting dimension lifting.

LEMMA 3.7. *If $\theta \in \mathcal{D}$, then $z := (\theta^\top, u_1^\top, \ldots, u_L^\top)^\top \in \mathcal{D}_0$ where $u_\ell := \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})$ for all $\ell \in [L]$. Conversely, if $z := (\theta^\top, u^\top)^\top \in \mathcal{D}_0$, then $\theta \in \mathcal{D}$.*

Proof. For any $\theta \in \mathcal{D}$, it follows from $z := (\theta^\top, u_1^\top, \ldots, u_L^\top)^\top$ with $u_\ell := \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})$ for all $\ell \in [L]$ that $z \in \mathcal{F}_0$. Then for any $d \in \mathcal{T}_{\mathcal{F}_0}(z)$, we have $d_{u_\ell} = \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})$ for all $\ell \in [L]$ by Theorem 3.6. Together with Proposition 3.4, it implies that for any $d \in \mathcal{T}_{\mathcal{F}_0}(z)$,

$$F'(z; d) = \Psi'(\theta; d_\theta) + 2\lambda\theta^\top d_\theta \geq 0,$$

where the inequality comes from $\theta \in \mathcal{D}$. On the other hand, if $z := (\theta^\top, u^\top)^\top \in \mathcal{D}_0$, then $z \in \mathcal{F}_0$, i.e. $u_\ell := \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})$ for all $\ell \in [L]$. Then for any $d_\theta \in \mathbb{R}^n$, it follows from Proposition 3.4 that under the setting of $d := (d_\theta^\top, d_{u_1}^\top, \ldots, d_{u_L}^\top)^\top$ with $d_{u_\ell} := \psi'_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1}; d_\theta, d_{u_1}, \ldots, d_{u_{\ell-1}})$ for all $\ell \in [L]$,

$$\Psi'(\theta; d_\theta) + 2\lambda\theta^\top d_\theta = F'(z; d) \geq 0,$$

where the inequality uses $d \in \mathcal{T}_{\mathcal{F}_0}(z)$ from Theorem 3.6 and $z \in \mathcal{D}_0$. ∎

To establish the equivalence between (P0) and (P1), inspired by Theorem 2.1 (a) of [9], we first show that under proper setting of $\{\beta_\ell > 0, \ell \in [L]\}$ restricted by K_g and $\{K_\ell > 0, \ell \in [L-1]\}$, the d-stationary point of (P1) in $lev_{\leq \bar{\gamma}}\Theta$ must be feasible to (P0), where $\bar{\gamma}$ is defined in (2.1).

LEMMA 3.8. *Under Assumption 1, let z be a d-stationary point of* (P1) *with $\Theta(z) \leq \bar{\gamma}$ and $\{\beta_\ell > 0, \ell \in [L]\}$ satisfying*

$$\beta_\ell > K_g \prod_{j=\ell+1}^{L} (1 + K_{j-1}), \text{ for all } \ell \in [L]. \tag{3.6}$$

Then $z \in \mathcal{F}_0$.

Proof. To show $z \in \mathcal{F}_0$, we are going to prove $u_\ell = \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})$ in the order of $\ell = L, \ldots, 1$ separately.

To show $u_L = \psi_{L-1}(\theta, u_1, \ldots, u_{L-1})$ by contradiction, we use $I_+^L(z), I_-^L(z)$ and $I_0^L(z)$ defined in Proposition 3.4. Suppose $u_L \neq \psi_{L-1}(\theta, u_1, \ldots, u_{L-1})$, then $I_+^L(z) \cup I_-^L(z) \neq \emptyset$. Let $\bar{z} := (\bar{\theta}^\top, \bar{u}_1^\top, \ldots, \bar{u}_L^\top)^\top$ with

$$\bar{\theta} := \theta,\ \bar{u}_1 := u_1,\ \ \ldots,\ \bar{u}_{L-1} := u_{L-1},$$
$$\bar{u}_L := \psi_{L-1}(\theta, u_1, \ldots, u_{L-1}) \neq u_L,$$

and $d := \bar{z} - z = (\mathbf{0}, \ldots, \mathbf{0}, d_{u_L}) \neq \mathbf{0}$. Then it follows from Proposition 3.4 that

$$\begin{aligned}\Theta'(z; d) &= F'(u; d_u) + \beta_L \left(\sum_{i \in I_+^L(z)} [d_{u_L}]_i - \sum_{i \in I_-^L(z)} [d_{u_L}]_i + \sum_{i \in I_0^L(z)} |[d_{u_L}]_i| \right) \\ &= g'(u; d_u) - \beta_L \|d_{u_L}\|_1 \qquad (3.7) \\ &\leq (K_g - \beta_L)\|d_{u_L}\|_1 < 0,\end{aligned}$$

where the two equalities use the definitions of d and $I_+^L(z), I_-^L(z), I_0^L(z)$, the first inequality holds due to (2.11) at $\bar{d} = \mathbf{0}$ and $\|\cdot\| \leq \|\cdot\|_1$, and the last inequality follows from (3.6) and $d_{u_L} \neq \mathbf{0}$. However, it contradicts $\Theta'(z; d) \geq 0$ for all $d \in \mathbb{R}^{\tilde{N}}$. Hence $u_L = \psi_{L-1}(\theta, u_1, \ldots, u_{L-1})$.

For any $\ell = L-1, \ldots, 1$, we next show $u_\ell = \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})$ using $I_+^\ell(z)$, $I_-^\ell(z)$ and $I_0^\ell(z)$ in Proposition 3.4. Suppose $u_\ell \neq \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})$, then $I_+^\ell(z) \cup I_-^\ell(z) \neq \emptyset$. Let $\bar{z} := (\bar{\theta}^\top, \bar{u}_1^\top, \ldots, \bar{u}_L^\top)^\top$ with

$$\begin{aligned}&\bar{\theta} := \theta,\ \bar{u}_1 := u_1,\ \ \ldots,\ \bar{u}_{\ell-1} := u_{\ell-1}, \\ &\bar{u}_\ell = \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1}) \neq u_\ell, \\ &\bar{u}_{\ell+1} := u_{\ell+1} + \psi_\ell'(\theta, u_1, \ldots, u_\ell; \mathbf{0}, \ldots, \mathbf{0}, \bar{u}_\ell - u_\ell), \\ &\vdots \\ &\bar{u}_L := u_L + \psi_{L-1}'(\theta, u_1, \ldots, u_{L-1}; \mathbf{0}, \ldots, \mathbf{0}, \bar{u}_\ell - u_\ell, \ldots, \bar{u}_{L-1} - u_{L-1}),\end{aligned}$$

and $d := \bar{z} - z = (\mathbf{0}, \ldots, \mathbf{0}, d_{u_\ell}, \ldots, d_{u_L}) \neq \mathbf{0}$. Then it can be checked that

$$d_{u_{\ell+1}} = \psi_\ell'(\theta, \mathbf{u}_\ell; d_\theta, d_{\mathbf{u}_\ell}),\ \ \ldots,\ d_{u_L} = \psi_{L-1}'(\theta, \mathbf{u}_{L-1}; d_\theta, d_{\mathbf{u}_{L-1}}). \qquad (3.8)$$

Together with Proposition 3.4, it implies that

$$\begin{aligned}\Theta'(z; d) &= F'(u; d_u) + \beta_\ell \left(\sum_{i \in I_+^\ell(z)} [d_{u_\ell}]_i - \sum_{i \in I_-^\ell(z)} [d_{u_\ell}]_i + \sum_{i \in I_0^\ell(z)} |[d_{u_\ell}]_i| \right) \\ &= g'(u; d_u) - \beta_\ell \|d_{u_\ell}\|_1 \\ &\leq K_g \|d_u\| - \beta_\ell \|d_{u_\ell}\|_1, \qquad (3.9)\end{aligned}$$

where the two equalities use the definitions of d and $I_+^\ell(z), I_-^\ell(z), I_0^\ell(z)$, the inequality holds due to (2.11) at $\bar{d} = \mathbf{0}$. Next we give an upper bound of $\|d_u\|$ by estimating $\{d_{u_j}, j \in [L]\}$. Since $d_{u_1} = \mathbf{0}, \ldots, d_{u_{\ell-1}} = \mathbf{0}$, we only need to analyze $\{d_{u_j}, j = \ell, \ldots, L\}$. For $j = \ell$, it follows from the definitions of d_{u_ℓ} and $\bar{u}_\ell$ that

$$\|d_{u_\ell}\| = \|u_\ell - \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})\| \neq 0. \qquad (3.10)$$

For $j = \ell + 1$, it follows from (3.8) that

$$\|d_{u_{\ell+1}}\| = \|\psi'_\ell(\theta, \mathbf{u}_\ell; d_\theta, d_{\mathbf{u}_\ell})\| \le K_\ell(\|d_{u_1}\| + \cdots + \|d_{u_\ell}\|) = K_\ell \|d_{u_\ell}\|, \tag{3.11}$$

where the inequality holds due to (2.11) at $\bar{d} = \mathbf{0}, d_\theta = \mathbf{0}$, the last equality uses $d_{u_1} = \mathbf{0}, \ldots, d_{u_{\ell-1}} = \mathbf{0}$. And for $j \ge \ell + 2$, assume that

$$\|d_k\| \le \left(K_{k-1} \prod_{i=\ell+1}^{k-1} (1 + K_{i-1}) \right) \|d_{u_\ell}\| \tag{3.12}$$

for all $k = \ell + 1, \ldots, j - 1$. Then we can obtain that (3.12) also holds at $k = j$:

$$\begin{aligned} \|d_{u_j}\| &= \|\psi'_{j-1}(\theta, \mathbf{u}_{j-1}; d_\theta, d_{\mathbf{u}_{j-1}})\| \\ &\le K_{j-1}(\|d_{u_1}\| + \cdots + \|d_{u_{j-1}}\|) \\ &= K_{j-1}(\|d_{u_\ell}\| + \|d_{u_{\ell+1}}\| + \cdots + \|d_{u_{j-1}}\|) \\ &\le \left(K_{j-1} \prod_{i=\ell+1}^{j-1} (1 + K_{i-1}) \right) \|d_{u_\ell}\|, \end{aligned} \tag{3.13}$$

where the first inequality holds due to (2.11) at $\bar{d} = \mathbf{0}, d_\theta = \mathbf{0}$, the second equality uses $d_{u_1} = \mathbf{0}, \ldots, d_{u_{\ell-1}} = \mathbf{0}$, and the second inequality uses (3.12) at $k = \ell + 1, \ldots, j - 1$. By induction and (3.11), it implies that (3.12) holds for all $k = \ell + 1, \ldots, L$. Plugging these upper bounds for $\{\|d_{u_j}\|, j \in [L]\}$ into (3.9), we have

$$\begin{aligned} \Theta'(z; d) &\le K_g(\|d_{u_\ell}\| + \|d_{u_{\ell+1}}\| + \cdots + \|d_{u_L}\|) - \beta_\ell \|d_{u_\ell}\| \\ &\le \left(K_g \left(1 + K_\ell + \cdots + K_{L-1} \prod_{i=\ell+1}^{L-1} (1 + K_{i-1}) \right) - \beta_\ell \right) \|d_{u_\ell}\| \\ &\le \left(K_g \prod_{i=\ell+1}^{L} (1 + K_{i-1}) - \beta_\ell \right) \|d_{u_\ell}\| \\ &< 0, \end{aligned}$$

where the last inequality uses (3.6) and (3.10). However, it contradicts $\Theta'(z; d) \ge 0$ for all $d \in \mathbb{R}^{\bar{N}}$. Hence, $u_\ell = \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})$, which yields the result due to the arbitrariness of ℓ. ∎

In general, condition (3.6) can be satisfied under $\beta_1 = \cdots = \beta_L = \beta$ with sufficiently large $\beta > 0$, since K_g and $\{K_\ell, \ell \in [L-1]\}$ defined in (2.8) and (2.9) are non-increasing when β is increasing. For certain applications in machine learning, such as training process of RNNs to be shown in Section 4, Lipschitz moduli K_g and $\{K_\ell, \ell \in [L-1]\}$ satisfying (2.11) on $lev_{\le \bar{\gamma}}\Theta$ are easy to estimate (see (4.5)-(4.9)), which provides computable thresholds for $\{\beta_\ell, \ell \in [L]\}$. Based on Lemma 3.8, we could show the equivalence of (P0) and (P1) in terms of global optimality and d-stationarity.

THEOREM 3.9. *Under Assumption 1, set $\{\beta_\ell > 0, \ell \in [L]\}$ satisfying (3.6). Then*
(a) $\mathcal{S}_0 = \mathcal{S}_1$;
(b) for any $z \in lev_{\le \bar{\gamma}}\Theta$ with $\bar{\gamma}$ defined in (2.1), z is a d-stationary point of (P0) if and only if it is a d-stationary point of (P1).

Proof. (a). For any $z \in \mathcal{S}_1$, it follows from Lemma 3.1 that $\Theta'(z;d) \geq 0$ for all $d \in \mathbb{R}^{\bar{N}}$, i.e. z is a d-stationary point of (P1). Together with Lemma 3.8 and $\Theta(z) \leq \Theta(z^0) = \bar{\gamma}$ from the global optimality of z, it implies that $z \in \mathcal{F}_0$. Hence, $\mathcal{S}_1 \subseteq \mathcal{F}_0$. Since $\mathcal{S}_1 \neq \emptyset$, we further have

$$\mathcal{S}_1 = \operatorname*{argmin}_{z\in\mathbb{R}^{\bar{N}}} \Theta(z) = \operatorname*{argmin}_{z\in\mathcal{F}_0} \Theta(z) = \operatorname*{argmin}_{z\in\mathcal{F}_0} F(z) = \mathcal{S}_0.$$

(b). We first show that any d-stationary point of (P1) in $lev_{\leq\bar{\gamma}}\Theta$ must be a d-stationary point for (P0). Firstly, it follows from Lemma 3.8 that $z \in \mathcal{F}_0$. Hence, it follows from Proposition 3.4 and Theorem 3.6 that for any $d \in \mathcal{T}_{\mathcal{F}_0}(z)$,

$$F'(z;d) = \Theta'(z;d) \geq 0,$$

where the inequality comes from $\Theta'(z;d) \geq 0$ for all d.

Next we will show the reverse implication: any d-stationary point z of (P0) with $\Theta(z) \leq \bar{\gamma}$ is also a d-stationary point of (P1). Firstly, it follows from $z \in \mathcal{F}_0$ that $I^\ell_+(z) = I^\ell_-(z) = \emptyset$ for all $\ell \in [L]$, which are defined in Proposition 3.4. It simplifies (3.3) as

$$\Theta'(z;d) = F'(z;d) + \sum_{\ell=1}^{L} \beta_\ell \|d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\|_1. \tag{3.14}$$

Together with Theorem 3.6, it further implies that for any $d \in \mathcal{T}_{\mathcal{F}_0}(z)$,

$$\Theta'(z;d) = F'(z;d) \geq 0, \tag{3.15}$$

where the inequality holds since z is the d-stationary point of (P0). For any $d \notin \mathcal{T}_{\mathcal{F}_0}(z)$, we can construct a direction $\bar{d}$ as follows: set $\bar{d}_\theta := d_\theta$ and $\bar{d}_{u_\ell} := \psi'_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1}; \bar{d}_\theta, \bar{d}_{u_1}, \ldots, \bar{d}_{u_{\ell-1}})$ in the order of $\ell = 1, \ldots, L$. Then by Theorem 3.6, we have $\bar{d} \in \mathcal{T}_{\mathcal{F}_0}(z)$. Hence, it follows from (3.14) and (3.15) that

$$\begin{aligned}
\Theta'(z;d) &= \Theta'(z;d) - \Theta'(z;\bar{d}) + \Theta'(z;\bar{d}) \\
&\geq \Theta'(z;d) - \Theta'(z;\bar{d}) \\
&= F'(z;d) - F'(z;\bar{d}) + \sum_{\ell=1}^{L} \beta_\ell \|d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\|_1. \qquad (3.16)
\end{aligned}$$

Next we will show that the right-hand side of (3.16) is nonnegative. Note that

$$F'(z;d) - F'(z;\bar{d}) = g'(u;d_u) - g'(u;\bar{d}_u) \geq -K_g \sum_{\ell=1}^{L} \|d_{u_\ell} - \bar{d}_{u_\ell}\|, \tag{3.17}$$

where the equality follows from $\bar{d}_\theta = d_\theta$ and (3.2), and the inequality uses (2.11). We only need to estimate $\{\|d_{u_\ell} - \bar{d}_{u_\ell}\|, \ell \in [L]\}$ by induction in the following. For $\ell = 1$, it follows from the definition of $\bar{d}_{u_1}$ and $\bar{d}_\theta = d_\theta$ that

$$\|d_{u_1} - \bar{d}_{u_1}\| = \|d_{u_1} - \psi'_0(\theta; d_\theta)\|. \tag{3.18}$$

For $\ell = 2, \ldots, L$, assume that

$$\begin{aligned}
&\|d_{u_j} - \bar{d}_{u_j}\| \qquad (3.19) \\
&\leq \|d_{u_j} - \psi'_{j-1}(\theta, \mathbf{u}_{j-1}; d_\theta, d_{\mathbf{u}_{j-1}})\| \\
&\quad + K_{j-1} \sum_{k=1}^{j-1} \left[\prod_{i=k+1}^{j-1} (1 + K_{i-1}) \right] \|d_{u_k} - \psi'_{k-1}(\theta, \mathbf{u}_{k-1}; d_\theta, d_{\mathbf{u}_{k-1}})\|
\end{aligned}$$

holds for all $j = 1, \ldots, \ell - 1$. Then we can deduce that (3.19) also holds at $j = \ell$:

$$
\begin{aligned}
&\|d_{u_\ell} - \bar{d}_{u_\ell}\| \\
=&\|d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; \bar{d}_\theta, \bar{d}_{\mathbf{u}_{\ell-1}})\| \\
\le&\|d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\| \\
&+ K_{\ell-1}(\|d_{u_1} - \bar{d}_{u_1}\| + \cdots + \|d_{u_{\ell-1}} - \bar{d}_{u_{\ell-1}}\|) \qquad (3.20)\\
\le&\|d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\| \\
&+ K_{\ell-1}\begin{pmatrix}\|d_{u_1} - \psi'_0(\theta; d_\theta)\| \cdot [1 + K_1 + \cdots + K_{\ell-2}\prod_{i=2}^{\ell-2}(1 + K_{i-1})] \\ + \cdots + \|d_{u_{\ell-1}} - \psi'_{\ell-2}(\theta, \mathbf{u}_{\ell-2}; d_\theta, d_{\mathbf{u}_{\ell-2}})\|\end{pmatrix} \\
=&\|d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\| \\
&+ K_{\ell-1} \cdot \begin{pmatrix}\|d_{u_1} - \psi'_0(\theta; d_\theta)\| \cdot \prod_{i=2}^{\ell-1}(1 + K_{i-1}) + \ldots \\ + \|d_{u_{\ell-1}} - \psi'_{\ell-2}(\theta, \mathbf{u}_{\ell-2}; d_\theta, d_{\mathbf{u}_{\ell-2}})\|\end{pmatrix},
\end{aligned}
$$

where the first equality comes from the definition of $\bar{d}_{u_\ell}$, the first inequality uses $d_\theta = \bar{d}_\theta$ and (2.11), the last inequality follows from (3.19) at $j = 1, \ldots, \ell - 1$. Together with (3.18), it implies that (3.19) holds for all $\ell \in [L]$. Plugging these upper bounds for $\{\|d_{u_\ell} - \bar{d}_{u_\ell}\|, \ell \in [L]\}$ into (3.17), we have

$$
\begin{aligned}
&F'(z; d) - F'(z; \bar{d}) \\
\ge& - K_g \begin{pmatrix}\|d_{u_1} - \psi'_0(\theta; d_\theta)\| \cdot [1 + K_1 + \cdots + K_{L-1}\prod_{i=2}^{L-1}(1 + K_{i-1})] \\ + \cdots + \|d_{u_L} - \psi'_{L-1}(\theta, \mathbf{u}_{L-1}; d_\theta, d_{\mathbf{u}_{L-1}})\|\end{pmatrix} \\
=& - K_g \sum_{\ell=1}^{L}\left[\prod_{j=\ell+1}^{L}(1 + K_{j-1})\right]\|d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\|.
\end{aligned}
$$

Together with (3.16), it implies that

$$
\begin{aligned}
\Theta'(z; d) \ge& \sum_{\ell=1}^{L}\left(\beta_\ell - K_g\left[\prod_{j=\ell+1}^{L}(1 + K_{j-1})\right]\right) \cdot \|d_{u_\ell} - \psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\| \\
>& 0, \qquad (3.21)
\end{aligned}
$$

where the last inequality uses (3.6), $d \notin \mathcal{T}_{\mathcal{F}_0}(z)$ and Theorem 3.6. Therefore, it yields that $\Theta'(z; d) \ge 0$ for all d, meaning that z is a d-stationary point of (P1). ∎

Remark 3.10. Theorem 3.9 is different from Theorem 2.1 of [9] in two aspects.

(i) Theorem 2.1 of [9] only obtains one-sided implication that a d-stationary point of the penalty problem must be a d-stationary point of the original problem, while we have the equivalence. As a consequence, the penalization preserves all the d-stationary points of (P) and (P0). And when $g, \{\psi_{\ell-1}, \ell \in [L]\}$ are smooth or have DC structures of the pointwise max type [11, Condition C2], d-stationary points of (P1) can be obtained by trust region methods [30] or majorization minimization frameworks [11].

(ii) Motivated by [21, Theorem 2.5] and [22, Lemma 9], we replace the boundedness requirement of z with $\Theta(z) \le \Theta(z^0)$, which is easier to check since $\Theta(z^0)$ is easy to calculate. And the condition $\Theta(z) \le \Theta(z^0)$ also helps to obtain the threshold-like conditions expressed by $\{K_\ell, \ell \in [L - 1]\}$ and K_g, which provides the relations between penalty thresholds and the number of layers L.

3.3. Second-order d-stationarity. Here we compare the sets of second-order d-stationary points $\mathcal{SD}_0$ and $\mathcal{SD}_1$ of (P0) and (P1) to exhibit their differences in depicting second-order necessary conditions for (P).

First of all, the idea of using second-order conditions of reformulated problems to characterize optimality conditions of the original problem is motivated by [10, Proposition 9.4.2] for (P) and (P1) with $L = 1$, i.e. $\min_\theta h(G(\theta))$ with $G(\theta) := (\psi_0(\theta)^\top, \lambda\|\theta\|^2)^\top$ and $h(y) := g([y]_{1:N_1}) + [y]_{(N_1+1)}$. The second-order necessary conditions in (9.41) of [10, Proposition 9.4.2] can be reorganized as: if θ is a local minimizer of (P), then for all $\rho > \max\{\overline{K}_g, 1\}$ where $\overline{K}_g > 0$ is a Lipschitz constant of g near $u_1 := \psi_0(\theta)$,

$$
\begin{aligned}
&g'(u_1; d_{u_1}) + 2\lambda\theta^\top d_\theta \geq 0 \text{ for all } d = (d_\theta^\top, d_{u_1}^\top)^\top \text{ with } d_{u_1} = \psi_0'(\theta; d_\theta),\\
&\text{and } g^{(2)}(u_1; d_{u_1}) + 2\lambda\rho\|d_\theta\|^2 + \rho\|\psi_0^{(2)}(\theta; d_\theta)\|_1 \geq 0,\\
&\qquad \text{for all } d = (d_\theta^\top, d_{u_1}^\top)^\top \text{ with } d_{u_1} = \psi_0'(\theta; d_\theta),\ g'(u_1; d_{u_1}) + 2\lambda\theta^\top d_\theta = 0,
\end{aligned}
$$

which is actually covered by our second-order necessary conditions in constructing $\mathcal{SD}_1$ since for any $z = (\theta^\top, u_1^\top)^\top$ with $u_1 = \psi_0(\theta)$,

- for any $d = (d_\theta^\top, d_{u_1}^\top)^\top$ with $d_{u_1} = \psi_0'(\theta; d_\theta)$, it follows from Proposition 3.4 that $\Theta'(z; d) = g'(u_1; d_{u_1}) + 2\lambda\theta^\top d_\theta$;
- it follows from Proposition 3.4 that

$$
\{d \mid d_{u_1} = \psi_0'(\theta; d_\theta), g'(u_1; d_{u_1}) + 2\lambda\theta^\top d_\theta = 0\} \subseteq \{d \mid \Theta'(z; d) = 0\};
$$

- for any d with $d_{u_1} = \psi_0'(\theta; d_\theta)$ and $g'(u_1; d_{u_1}) + 2\lambda\theta^\top d_\theta = 0$, it follows from Proposition 3.5 with the setting of $\beta_1 = \rho$ and $\rho > \max\{\overline{K}_g, 1\}$ that

$$
\begin{aligned}
\Theta^{(2)}(z; d) &= g^{(2)}(u_1; d_{u_1}) + 2\lambda\|d_\theta\|^2 + \rho\|\psi_0^{(2)}(\theta; d_\theta)\|_1\\
&< g^{(2)}(u_1; d_{u_1}) + 2\lambda\rho\|d_\theta\|^2 + \rho\|\psi_0^{(2)}(\theta; d_\theta)\|_1.
\end{aligned}
$$

Hence, we will focus on our second-order necessary conditions for (P0) and (P1) rather than generalizing (9.41) of [10, Proposition 9.4.2] to the case of (P).

By the results in Sections 3.1 and 3.2, the second-order necessary conditions of (P0) and (P1) specified in Definition 3.2 are both able to characterize solutions of (P). It follows from Lemma 3.1 and Theorem 3.9 that

$$
\begin{array}{ccccc}
\mathcal{S}_0 & \subseteq & \mathcal{SD}_0 \cap lev_{\leq\bar\gamma}\Theta & \subseteq & \mathcal{D}_0 \cap lev_{\leq\bar\gamma}\Theta\\
\| & & & & \|\\
\mathcal{S}_1 & \subseteq & \mathcal{SD}_1 \cap lev_{\leq\bar\gamma}\Theta & \subseteq & \mathcal{D}_1 \cap lev_{\leq\bar\gamma}\Theta.
\end{array}
$$

Together with the bijection between $\mathcal{S}$ and $\mathcal{S}_0$ (discussed after Lemma 2.1), it implies that for any $\theta \in \mathcal{S}$, the point $z = (\theta^\top, u_1^\top, \ldots, u_L^\top)^\top$ must belong to $\mathcal{SD}_0 \cap lev_{\leq\bar\gamma}\Theta$ and $\mathcal{SD}_1 \cap lev_{\leq\bar\gamma}\Theta$ where $u_\ell := \psi_{\ell-1}(\theta, u_1, \ldots, u_{\ell-1})$ for all $\ell \in [L]$. Furthermore, we could find the latter condition $z \in \mathcal{SD}_1 \cap lev_{\leq\bar\gamma}\Theta$ is stronger by the following observation.

THEOREM 3.11. *Under Assumptions 1, 2 and* (3.6), $\mathcal{SD}_0 \cap lev_{\leq\bar\gamma}\Theta \supseteq \mathcal{SD}_1 \cap lev_{\leq\bar\gamma}\Theta$.

Proof. By the definitions of $\mathcal{SD}_0$ and $\mathcal{SD}_1$,

$$
\begin{aligned}
&\mathcal{SD}_0 \cap lev_{\leq\bar\gamma}\Theta = \{z \mid F^{(2)}(z; d) \geq 0, \forall d \in P_{\mathcal{F}_0}(z) \text{ with } F'(z; d) = 0\} \cap (\mathcal{D}_0 \cap lev_{\leq\bar\gamma}\Theta),\\
&\mathcal{SD}_1 \cap lev_{\leq\bar\gamma}\Theta = \{z \mid \Theta^{(2)}(z; d) \geq 0, \forall d \text{ with } \Theta'(z; d) = 0\} \cap (\mathcal{D}_1 \cap lev_{\leq\bar\gamma}\Theta).
\end{aligned}
$$

Together with $\mathcal{D}_0 \cap lev_{\leq\bar{\gamma}}\Theta = \mathcal{D}_1 \cap lev_{\leq\bar{\gamma}}\Theta$ from Theorem 3.9, we only need to prove that for any $z \in \mathcal{D}_1 \cap lev_{\leq\bar{\gamma}}\Theta$ satisfying $\Theta^{(2)}(z;d) \geq 0$ for all d with $\Theta'(z;d) = 0$, the inequality $F^{(2)}(z;d) \geq 0$ holds for all $d \in P_{\mathcal{F}_0}(z)$ with $F'(z;d) = 0$.

First we show

$$\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}}) = 0,\ \ell \in [L] \text{ for all } z \in \mathcal{F}_0, d \in P_{\mathcal{F}_0}(z) \tag{3.22}$$

by contradiction. If there exists $\ell \in [L]$ and $i \in [N_\ell]$, such that $[\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i > 0$, then it follows from the definition of second-order directional derivatives that for any sufficiently small positive number τ,

$$0 < [\psi_{\ell-1}(\theta + \tau d_\theta, \mathbf{u}_{\ell-1} + \tau d_{\mathbf{u}_{\ell-1}})]_i - [\psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})]_i - \tau[\psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i.$$

Together with $z \in \mathcal{F}_0, d \in P_{\mathcal{F}_0}(z) \subseteq \mathcal{T}_{\mathcal{F}_0}(z)$, it implies that

$$\begin{aligned}
&[u_\ell + \tau d_{u_\ell}]_i - [\psi_{\ell-1}(\theta + \tau d_\theta, \mathbf{u}_{\ell-1} + \tau d_{\mathbf{u}_{\ell-1}})]_i \\
=&[\psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})]_i + \tau[\psi'_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i - [\psi_{\ell-1}(\theta + \tau d_\theta, \mathbf{u}_{\ell-1} + \tau d_{\mathbf{u}_{\ell-1}})]_i \\
<&0
\end{aligned}$$

for any sufficiently small positive τ, which contradicts with $d \in P_{\mathcal{F}_0}(z)$. Similar contradiction appears if there exists $\ell \in [L]$ and $i \in [N_\ell]$, such that $[\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_i < 0$, which yields (3.22).

Hence, for any $z \in \mathcal{SD}_1 \cap lev_{\leq\bar{\gamma}}\Theta$, we have

$$0 \leq \Theta^{(2)}(z;d) = F^{(2)}(z;d) + \sum_{\ell=1}^{L} \beta_\ell \|\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\|_1 = F^{(2)}(z;d)$$

for all $d \in P_{\mathcal{F}_0}(z)$ with $F'(z;d) = 0$, where the inequality is derived by

$$\{d \in P_{\mathcal{F}_0}(z) \mid F'(z;d) = 0\} \subseteq \{d \in \mathcal{T}_{\mathcal{F}_0}(z) \mid F'(z;d) = 0\} \subseteq \{d \mid \Theta'(z;d) = 0\}$$

from $z \in \mathcal{F}_0$, Theorem 3.6 and Proposition 3.4, the first equality uses $z \in \mathcal{F}_0$, Theorem 3.6 and Proposition 3.5, and the last equality uses (3.22). ∎

Remark 3.12. Here we discuss the conditions under which the equality in Theorem 3.11 holds. Under the premises of Theorem 3.11, it follows from (3.21) that $\Theta'(z;d) > 0$ for any $z \in \mathcal{D}_0 \cap lev_{\leq\bar{\gamma}}\Theta = \mathcal{D}_1 \cap lev_{\leq\bar{\gamma}}\Theta$ and any $d \notin \mathcal{T}_{\mathcal{F}_0}(z)$. Hence, for any $z \in \mathcal{D}_0 \cap lev_{\leq\bar{\gamma}}\Theta = \mathcal{D}_1 \cap lev_{\leq\bar{\gamma}}\Theta$,

$$\{d \mid \Theta'(z;d) = 0\} = \{d \in \mathcal{T}_{\mathcal{F}_0}(z) \mid \Theta'(z;d) = 0\} = \{d \in \mathcal{T}_{\mathcal{F}_0}(z) \mid F'(z;d) = 0\}, \tag{3.23}$$

where the last equality uses $z \in \mathcal{F}_0$, Theorem 3.6 and Proposition 3.4. Together with Proposition 3.5 and Theorem 3.6, it implies that for any d with $\Theta'(z;d) = 0$,

$$\Theta^{(2)}(z;d) = F^{(2)}(z;d) + \sum_{\ell=1}^{L} \beta_\ell \|\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\|_1,$$

which indicates that $z \in \mathcal{SD}_1 \cap lev_{\leq\bar{\gamma}}\Theta$ if and only if $z \in \mathcal{D}_1 \cap lev_{\leq\bar{\gamma}}\Theta$ and

$$F^{(2)}(z;d) + \sum_{\ell=1}^{L} \beta_\ell \|\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\|_1 \geq 0 \tag{3.24}$$

for all $d \in \mathcal{T}_{\mathcal{F}_0}(z)$ satisfying $F'(z;d) = 0$. Then it follows from (3.22) and $P_{\mathcal{F}_0}(z) \subseteq \mathcal{T}_{\mathcal{F}_0}(z)$ that, the equality in Theorem 3.11 holds if and only if

$$\begin{aligned} &F^{(2)}(z;d) + \sum_{\ell=1}^{L} \beta_\ell \|\psi_{\ell-1}^{(2)}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\|_1 \geq 0, \\ &\text{for all } d \in \mathcal{T}_{\mathcal{F}_0}(z) \backslash P_{\mathcal{F}_0}(z) \text{ satisfying } F'(z;d) = 0. \end{aligned} \tag{3.25}$$

Sufficient conditions for (3.25) include the following two conditions.

- $P_{\mathcal{F}_0}(z) = \mathcal{T}_{\mathcal{F}_0}(z)$. If $\mathcal{F}_0$ is a polyhedron or a union of finite number of polyhedrons, then $P_{\mathcal{F}_0}(z) = \mathcal{T}_{\mathcal{F}_0}(z)$ holds for any $z \in \mathcal{F}_0$. For example, under the setting of $\psi_0(\theta) := a^\top \theta, \psi_1(\theta, u_1) := [u_1]_+$ for a vector $a \in \mathbb{R}^n$,

$$\begin{aligned} \mathcal{F}_0 &= \{z = (\theta^\top, u_1, u_2)^\top \mid u_1 = a^\top \theta,\ u_2 = [u_1]_+\} \\ &= \{z \mid u_1 = a^\top \theta,\ u_1 \geq 0,\ u_2 = u_1\} \cup \{z \mid u_1 = a^\top \theta,\ u_1 \leq 0,\ u_2 = 0\} \end{aligned}$$

  is a union of two polyhedrons.
- F is convex and twice directionally differentiable. Since $F(z) = g(u) + \lambda\|\theta\|^2$, F is convex and twice directionally differentiable when g is convex and twice directionally differentiable. In this case, (3.25) naturally holds since $F^{(2)}(z;d) = \lim_{\tau \downarrow 0} \frac{F(z+\tau d) - F(z) - F'(z;\tau d)}{\tau^2/2} \geq 0$ for all d.

Inspired by Remark 3.12, we can provide an example where $\mathcal{SD}_0 \cap lev_{\leq \bar{\gamma}}\Theta \supsetneq \mathcal{SD}_1 \cap lev_{\leq \bar{\gamma}}\Theta$ under the premises of Theorem 3.11.

Example 3.13. Consider (P) with $L = 2, n = N_1 = N_2 = 1$ and $\lambda = 0.01$,

$$\psi_0(\theta) := \theta,\ \psi_1(\theta, u_1) := u_1^2,\ g(u_1, u_2) := [-u_1^2 + 0.5u_2 + 0.0001]_+.$$

On the one hand, it can be verified that $z^0 = (0, \psi_0(0), \psi_1(0, \psi_0(0)))^\top = \mathbf{0}$ is a second-order d-stationary point of (P0). Firstly, it follows from Theorem 3.6, the definition of $P_{\mathcal{F}_0}(\cdot)$ and (3.22) that $\mathcal{T}_{\mathcal{F}_0}(\mathbf{0}) = \{(d_\theta, d_{u_1}, d_{u_2})^\top \mid d_{u_1} = d_\theta,\ d_{u_2} = 0\}$ and $P_{\mathcal{F}_0}(\mathbf{0}) = \{\mathbf{0}\}$ since $\{\mathbf{0}\} \subseteq P_{\mathcal{F}_0}(\mathbf{0}) \subseteq \{d \in \mathcal{T}_{\mathcal{F}_0}(\mathbf{0}) \mid 2d_{u_1}^2 = 0\} = \{\mathbf{0}\}$. Then it can be verified that $F'(\mathbf{0};d) = 0.5d_{u_2} = 0 \geq 0$ for all $d \in \mathcal{T}_{\mathcal{F}_0}(\mathbf{0})$, and $F^{(2)}(\mathbf{0};d) = -2d_{u_1}^2 + 0.02d_\theta^2 = 0 \geq 0$ for all $d \in P_{\mathcal{F}_0}(\mathbf{0}) \cap \{d \mid F'(\mathbf{0};d) = 0\} = \{\mathbf{0}\}$. On the other hand, z^0 is not a second-order d-stationary point of (P1) with $\beta_1 = 1, \beta_2 = 0.6$ where (3.6) holds. First we can verify (3.6) holds, i.e. $\beta_1 > K_g(1 + K_1)$ and $\beta_2 > K_g$. Since $\bar{\gamma} = F(z^0) = 10^{-4}$, for all $(\theta, u_1, u_2)^\top \in lev_{\leq \bar{\gamma}}\Theta$, it can be calculated that $|u_1| \leq |\theta| + |u_1 - \theta| \leq \sqrt{10^{-4}/10^{-2}} + 10^{-4} = 0.1001$. It implies that for all $(\theta, u_1, u_2)^\top, (\theta, \bar{u}_1, \bar{u}_2)^\top \in lev_{\leq \bar{\gamma}}\Theta$,

$$\begin{aligned} &|g(u) - g(\bar{u})| \leq \sqrt{(u_1 + \bar{u}_1)^2 + 0.25}\|u - \bar{u}\| < 0.5386\|u - \bar{u}\|, \\ &|\psi_1(\theta, u_1) - \psi_1(\theta, \bar{u}_1)| \leq |u_1 + \bar{u}_1| \cdot |u_1 - \bar{u}_1| < 0.21|u_1 - \bar{u}_1|. \end{aligned}$$

Hence, there exist $K_g \in (0, 0.5386]$ and $K_1 \in (0, 0.21]$ satisfying (2.8)-(2.9), which guarantees (3.6). Then, it follows from Theorem 3.9 and $z^0 \in \mathcal{D}_0 \cap lev_{\leq \bar{\gamma}}\Theta$ that $z^0 \in \mathcal{D}_1 \cap lev_{\leq \bar{\gamma}}\Theta$. Together with Remark 3.12, it implies that $z^0 \in \mathcal{SD}_1$ if and only if (3.24) holds at z^0 for all $d \in \mathcal{T}_{\mathcal{F}_0}(z^0)$ with $F'(z^0;d) = 0$. However, for any $d = (d_\theta, d_{u_1}, d_{u_2})^\top$ with $d_{u_2} = 0$ and $d_\theta = d_{u_1} \neq 0$, we have $d \in \mathcal{T}_{\mathcal{F}_0}(z^0)$, $F'(z^0;d) = 0$ and $F^{(2)}(z^0;d) + \sum_{\ell=1}^{L} \beta_\ell \|\psi_{\ell-1}^{(2)}(\theta^0, \mathbf{u}_{\ell-1}^0; d_\theta, d_{\mathbf{u}_{\ell-1}})\|_1 = -2d_{u_1}^2 + 0.02d_\theta^2 + 1.2d_{u_1}^2 = -0.78d_\theta^2 < 0$, which violates (3.24).

In Section 4, we will provide an application of (P) where the second-order d-stationary points of corresponding (P0) and (P1) are computable by certain algorithms.

3.4. Second-order sufficient condition. Inspired by [10, Proposition 9.4.2 (b)], we provide second-order sufficient conditions for strong local minimizers [10, Section 6.4] of (P) in this subsection. For a function $f : \mathcal{F} \subseteq \mathbb{R}^m \to \mathbb{R}$, we say $x \in \mathcal{F}$ is a strong local minimizer of f on $\mathcal{F}$ if there exist $\epsilon_1, \epsilon_2 > 0$ such that $f(\bar{x}) \geq f(x) + \epsilon_1 \|\bar{x} - x\|^2$ for all $\bar{x} \in \mathcal{F}$ satisfying $\|\bar{x} - x\| \leq \epsilon_2$. To this end, we need the following assumption about twice semidifferentiability [25, Definition 13.6].

ASSUMPTION 3. *Function g and each component of vector functions $\{\psi_{\ell-1}, \ell \in [L]\}$ are twice semidifferentiable on $\mathbb{R}^{\bar{N}_L}$ and $\mathbb{R}^{n+\bar{N}_{\ell-1}}, \ell \in [L]$ respectively.*

Assumption 3 is stronger than Assumption 2. As shown in Lemma 3.1, Assumption 2 provides an upper bound of $\liminf_{\tau\downarrow 0, d'\to d} \frac{\Theta(z+\tau d')-\Theta(z)}{\tau^2/2}$ along certain directions, whereas (3.29)-(3.31) indicate that Assumption 3 can simultaneously offer a lower bound for it in all directions. For any twice semidifferentiable function f, we have $\mathrm{d}f(x)(d) := \lim_{\tau\downarrow 0, d'\to d} \frac{f(x+\tau d')-f(x)}{\tau} = f'(x; d)$ for any x, d, and

$$\begin{aligned}\lim_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{f(x+\tau d') - f(x) - \tau \mathrm{d}f(x)(d')}{\tau^2/2} &= \lim_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{f(x+\tau d') - f(x) - \tau f'(x; d')}{\tau^2/2} \\ &= \lim_{\tau\downarrow 0} \frac{f(x+\tau d) - f(x) - \tau f'(x; d)}{\tau^2/2} \\ &= f^{(2)}(x; d),\end{aligned} \tag{3.26}$$

for any x, d. For any twice semidifferentiable functions f_1, f_2 on $\mathbb{R}^n$ and any $a_1, a_2 \in \mathbb{R}$, the combination $a_1 f_1 + a_2 f_2$ is twice semidifferentiable on $\mathbb{R}^n$. Then based on previous subsections, we have the following second-order sufficient conditions for (P).

THEOREM 3.14. *Under Assumptions 1, 3 and (3.6), for any $z = (\theta^\top, u^\top)^\top \in lev_{\leq \bar{\gamma}}\Theta$, if*

$$\begin{aligned}&\Theta'(z; d) \geq 0 \text{ for all } d,\\ &\text{and } F^{(2)}(z; d) - \sum_{\ell=1}^{L} \beta_\ell \|\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\|_1 > 0,\\ &\quad \text{for all } d \neq \mathbf{0} \text{ with } \Theta'(z; d) = 0,\end{aligned} \tag{3.27}$$

then z is a strong local minimizer of (P1) and θ is a strong local minimizer of (P).

Proof. We first prove that z is a strong local minimizer of (P1). According to [25, Theorem 13.24 (c)], it is equivalent to prove that $\mathbf{0} \in \partial\Theta(z)$ and

$$\liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{\Theta(z+\tau d') - \Theta(z)}{\tau^2/2} > 0 \quad \text{for all } d \neq \mathbf{0}. \tag{3.28}$$

Since $\mathbf{0} \in \partial\Theta(z)$ can be obtained by $\hat{\partial}\Theta(z) \subseteq \partial\Theta(z)$ from [25, Theorem 8.6] and $\mathbf{0} \in \hat{\partial}\Theta(z)$ from $\Theta'(z; d) \geq 0$ for all d and [25, Exercise 8.4], we only need to prove (3.28). For any $d \neq \mathbf{0}$ satisfying $\Theta'(z; d) > 0$, it follows from Assumption 1 that $\lim_{\tau\downarrow 0, d'\to d}[\Theta(z+\tau d') - \Theta(z)]/\tau$ exists and equals to $\Theta'(z; d)$, which implies that

$$\liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{\Theta(z+\tau d') - \Theta(z)}{\tau^2/2} = \liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{[\Theta(z+\tau d') - \Theta(z)]/\tau}{\tau/2} = +\infty > 0.$$

For any $d \neq \mathbf{0}$ satisfying $\Theta'(z;d) = 0$, it follows from $\Theta'(z;d') \geq 0$ for all d' that

$$\begin{aligned}
&\liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{\Theta(z+\tau d') - \Theta(z)}{\tau^2/2}\\
&\geq \liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{\Theta(z+\tau d') - \Theta(z) - \tau\Theta'(z;d')}{\tau^2/2}\\
&\geq \liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{F(z+\tau d') - F(z) - \tau F'(z;d')}{\tau^2/2} \qquad (3.29)\\
&\quad + \sum_{\ell=1}^{L} \beta_\ell \sum_{j\in[N_\ell]} \liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{f_{\ell,j}(z+\tau d') - f_{\ell,j}(z) - \tau f'_{\ell,j}(z;d')}{\tau^2/2},
\end{aligned}$$

where $f_{\ell,j}(z) := |[u_\ell]_j - [\psi_{\ell-1}]_j(\theta, \mathbf{u}_{\ell-1})|$ for all $\ell \in [L], j \in [N_\ell]$. By the twice semidifferentiability of g and $\lambda\|\cdot\|^2$, it follows from (3.26) that

$$\liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{F(z+\tau d') - F(z) - \tau F'(z;d')}{\tau^2/2} = F^{(2)}(z;d). \qquad (3.30)$$

And for all $\ell \in [L], j \in [N_\ell]$, it follows from twice semidifferentiability of $[\psi_{\ell-1}]_j$ and [10, (4.15)] that

$$\begin{aligned}
&\liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{f_{\ell,j}(z+\tau d') - f_{\ell,j}(z) - \tau f'_{\ell,j}(z;d')}{\tau^2/2}\\
&\geq \begin{cases} -[\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_j, & \text{if } j \in I^\ell_+(z) \cup I^\ell_{0,+}(z;d),\\ [\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_j, & \text{if } j \in I^\ell_-(z) \cup I^\ell_{0,-}(z;d),\\ -|[\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_j|, & \text{if } j \in I^\ell_{0,0}(z;d), \end{cases}
\end{aligned}$$

where $I^\ell_+(z)$, $I^\ell_-(z)$, $I^\ell_0(z)$ are defined as in Proposition 3.4, $I^\ell_{0,+}(z;d)$, $I^\ell_{0,-}(z;d)$, $I^\ell_{0,0}(z;d)$ are defined as in Proposition 3.5. In fact, $I^\ell_+(z) = I^\ell_-(z) = \emptyset$ for all ℓ since $z \in \mathcal{F}_0$ according to Lemma 3.8, and furthermore it follows from (3.23) that $I^\ell_{0,+}(z;d) = I^\ell_{0,-}(z;d) = \emptyset$ for all ℓ. Thus, the inequality can be simplified as

$$\liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{f_{\ell,j}(z+\tau d') - f_{\ell,j}(z) - \tau f'_{\ell,j}(z;d')}{\tau^2/2} \geq -|[\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})]_j|. \qquad (3.31)$$

Plugging (3.30) and (3.31) into (3.29), we have

$$\liminf_{\substack{\tau\downarrow 0\\ d'\to d}} \frac{\Theta(z+\tau d') - \Theta(z)}{\tau^2/2} \geq F^{(2)}(z;d) - \sum_{\ell=1}^{L} \beta_\ell \|\psi^{(2)}_{\ell-1}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\|_1 > 0,$$

where the last inequality comes from (3.27). Thus, (3.28) holds and z is a strong local minimizer of (P1).

Next we prove θ, the component of z, is a strong local minimizer of (P) by contradiction. If θ is not a strong local minimizer of (P), then there exists a sequence $\{\theta^k, k \geq 1\}$ converging to θ such that

$$\Psi(\theta^k) + \lambda\|\theta^k\|^2 < \Psi(\theta) + \lambda\|\theta\|^2 + \frac{1}{k}\|\theta^k - \theta\|^2.$$

Based on $\{\theta^k, k \geq 1\}$, we can construct $\{z^k, k \geq 1\} \subseteq \mathcal{F}_0$ by setting $z^k = ((\theta^k)^\top, (u_1^k)^\top, \ldots, (u_L^k)^\top)^\top$ with $u_\ell^k := \psi_{\ell-1}(\theta^k, u_1^k, \ldots, u_{\ell-1}^k)$ for all $\ell \in [L]$. Together with $z \in \mathcal{F}_0$ from Lemma 3.8, it implies that

$$\Theta(z^k) = \Psi(\theta^k) + \lambda\|\theta^k\|^2 < \Psi(\theta) + \lambda\|\theta\|^2 + \frac{1}{k}\|\theta^k - \theta\|^2 = \Theta(z) + \frac{1}{k}\|\theta^k - \theta\|^2$$
$$\leq \Theta(z) + \frac{1}{k}\|z^k - z\|^2. \tag{3.32}$$

Meanwhile, it follows from the continuity of $\{\psi_{\ell-1}, \ell \in [L]\}$ and $\theta^k \to \theta$ that $z^k \to z$. Together with the strict inequality in (3.32), it contradicts the fact that z is a strong local minimizer of (P1). Hence, θ is a strong local minimizer of (P). ∎

Remark 3.15. According to (3.23), the sufficient condition (3.27) is equivalent to $\Theta'(z;d) \geq 0$ for all d and $F^{(2)}(z;d) - \sum_{\ell=1}^L \beta_\ell \|\psi_{\ell-1}^{(2)}(\theta, \mathbf{u}_{\ell-1}; d_\theta, d_{\mathbf{u}_{\ell-1}})\|_1 > 0$ for all $d \neq \mathbf{0}$ with $d \in \mathcal{T}_{\mathcal{F}_0}(z)$ and $F'(z;d) = 0$. Hence, it can be observed that for the case $L = 1$, (3.27) is milder than [10, (9.42)] for $\min_\theta h(G(\theta))$ with $G(\theta) := (\psi_0(\theta)^\top, \lambda\|\theta\|^2)^\top$ and $h(y) := g([y]_{1:N_1}) + [y]_{(N_1+1)}$, since under $\rho := \beta_1$ and $J_+ := \{j \in [N_1] \mid [\psi_0^{(2)}(\theta; d_\theta)]_j \leq 0\}$, $J_- := \{N_1 + 1\} \cup ([N_1] \backslash J_+)$,

$$h^{(2)}(G(\theta); G'(\theta; d_\theta)) + \rho\,[\sum_{j \in J_+} G_j^{(2)}(\theta; d_\theta) - \sum_{j \in J_-} G_j^{(2)}(\theta; d_\theta)]$$
$$= g^{(2)}(\psi_0(\theta); \psi_0'(\theta; d_\theta)) - 2\lambda\beta_1\|d_\theta\|^2 - \beta_1\|\psi_0^{(2)}(\theta; d_\theta)\|_1$$
$$< F^{(2)}(z;d) - \beta_1\|\psi_0^{(2)}(\theta; d_\theta)\|_1,$$

where the inequality uses $F^{(2)}(z;d) = g^{(2)}(\psi_0(\theta); \psi_0'(\theta; d_\theta)) + 2\lambda\|d_\theta\|^2$ and $d_\theta \neq \mathbf{0}$ for all $d \in \mathcal{T}_{\mathcal{F}_0}(z)$ with $d \neq \mathbf{0}$.

Theorem 3.14 enables us to determine whether a d-stationary point of (P1) is a strong local minimizer for (P).

4. Application: RNNs. The recurrent neural network (RNN) is a kind of feedforward neural networks for sequential processing. Different RNN variants, such as Elman networks [12], Jordan networks [17], and LSTM [15], have been widely applied on language modelling like ChatGPT and protein secondary structure prediction [13]. Due to the universal approximation property and the fundamental significance for the other RNN variants [14], we focus on the training of the Elman RNN with a single unidirectional hidden layer in this section. Without loss of generality, we consider the case where the number of sequences is $N = 1$ and the number of time steps in the sequence is $T = 3$. Given a sequence of inputs $\{x_t \in \mathbb{R}^{N_0}, t \in [3]\}$ and an associated sequence of labels $\{y_t \in \mathbb{R}^{N_2}, t \in [3]\}$, the model can be formulated as the following constrained optimization problem

$$\text{(P0-RNN)} \quad \begin{aligned} \min_{\substack{A,V,W,b,c,\\ s,w,r,v}} \ & \frac{\|r - y\|^2}{6} + \lambda(\|A\|_F^2 + \|V\|_F^2 + \|W\|_F^2 + \|b\|^2 + \|c\|^2), \\ \text{subject to} \ & w_t = W s_{t-1} + A x_t + b,\ s_t = \sigma(w_t), \\ & v_t = V s_t + c,\ r_t = \sigma(v_t),\ t \in [3], \end{aligned}$$

where $s_0 = \mathbf{0} \in \mathbb{R}^{N_1}$ and other notations are defined as follows.

1. Vector y refers to $y = (y_1^\top, y_2^\top, y_3^\top)^\top \in \mathbb{R}^{3N_2}$.

2. Vectors $s_t \in \mathbb{R}^{N_1}$ and $r_t \in \mathbb{R}^{N_2}$ refer to the hidden value and output at time t, respectively. For brevity, we denote $s = (s_1^\top, s_2^\top, s_3^\top)^\top \in \mathbb{R}^{3N_1},\ r = (r_1^\top, r_2^\top, r_3^\top)^\top \in \mathbb{R}^{3N_2}$.
3. Vectors $w_t \in \mathbb{R}^{N_1}$ and $v_t \in \mathbb{R}^{N_2}$ refer to the auxiliary hidden value and auxiliary output at time t, respectively. We denote $w = (w_1^\top, w_2^\top, w_3^\top)^\top \in \mathbb{R}^{3N_1},\ v = (v_1^\top, v_2^\top, v_3^\top)^\top \in \mathbb{R}^{3N_2}$.
4. Matrices $W \in \mathbb{R}^{N_1 \times N_1}, A \in \mathbb{R}^{N_1 \times N_0}, V \in \mathbb{R}^{N_2 \times N_1}$ and vectors $b \in \mathbb{R}^{N_1}, c \in \mathbb{R}^{N_2}$ are network parameters independent of t. And we aggregate those parameters as

$$\theta := \left(\mathrm{vec}(A)^\top, \mathrm{vec}(V)^\top, \mathrm{vec}(W)^\top, b^\top, c^\top\right)^\top \in \mathbb{R}^n, \tag{4.1}$$

where $\mathrm{vec}(A) := (a_1^\top, \ldots, a_q^\top)^\top$ for any matrix $A = (a_1, \ldots, a_q) \in \mathbb{R}^{p \times q}$ with $\{a_j \in \mathbb{R}^p, j \in [q]\}$, and $n := N_0N_1 + N_1N_2 + N_1^2 + N_1 + N_2$ in this case.
5. Function $\sigma(u) := \max\{u, \alpha u\}$ for all $u \in \mathbb{R}$ is (leaky) ReLU activator with $\alpha \in [0, 1)$. For brevity, we will not distinguish whether $\sigma(\cdot)$ applies on a scalar or on a vector componentwisely when there is no ambiguity.

To reconcile the notations in (P0-RNN) with those in (P0), we could first define $L := 8$ and

$$u_{2t-1} := w_t,\ u_{2t} := s_t,\ t \in [3],\quad u_7 := v,\ u_8 := r;$$

then define $\mathbf{u}_0$ to be an empty placeholder, $\mathbf{u}_\ell := (u_1^\top, \ldots, u_\ell^\top)^\top$ for all $\ell \in [8]$ as in (1.1). Thereby, we have

$$u := \mathbf{u}_L = (w_1^\top, s_1^\top, w_2^\top, s_2^\top, w_3^\top, s_3^\top, v^\top, r^\top)^\top \in \mathbb{R}^{6(N_1+N_2)},$$

which aggregates all the auxiliary variables s, w, r, v in (P0-RNN). Together with (4.1), we have

$$z := (\theta^\top, u^\top)^\top \in \mathbb{R}^{\bar{N}},\ \text{where } \bar{N} := n + 6(N_2 + N_1), \tag{4.2}$$

so that the objective function of (P0-RNN) can be denoted as

$$F(z) := g(u) + \lambda \|\theta\|^2,\ \text{where } g(u) := \|r - y\|^2 / 6.$$

And for the constraints of (P0-RNN), using Kronecker product $\otimes$, we denote

$$\psi_{2t-2}(\theta, \mathbf{u}_{2t-2}) := (x_t^\top \otimes I_{N_1} \quad \mathbf{0} \quad s_{t-1}^\top \otimes I_{N_1} \quad I_{N_1} \quad \mathbf{0})\,\theta,\ \psi_{2t-1}(\theta, \mathbf{u}_{2t-1}) := \sigma(w_t)$$

for all $t \in [3]$, and

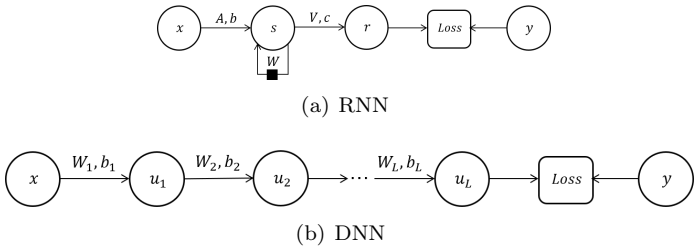
$$\psi_6(\theta, \mathbf{u}_6) := \begin{pmatrix} \mathbf{0} & s_1^\top \otimes I_{N_2} & \mathbf{0} & \mathbf{0} & I_{N_2} \\ \mathbf{0} & s_2^\top \otimes I_{N_2} & \mathbf{0} & \mathbf{0} & I_{N_2} \\ \mathbf{0} & s_3^\top \otimes I_{N_2} & \mathbf{0} & \mathbf{0} & I_{N_2} \end{pmatrix} \theta,\ \psi_7(\theta, \mathbf{u}_7) := \sigma(v).$$

Then it can be checked that

$$u_\ell = \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1}),\ \ell \in [8] \Leftrightarrow \begin{cases} w_t = W s_{t-1} + A x_t + b,\ s_t = \sigma(w_t), \\ v_t = V s_t + c,\ r_t = \sigma(v_t),\ t \in [3], \end{cases}$$

and the above functions g and $\{\psi_{\ell-1}, \ell \in [8]\}$ are continuous. Hence, (P0-RNN) is an application of (P0). Naturally, (P0-RNN) has a reformulation corresponding to (P). As noted at the beginning of Section 3.2, they are equivalent when neglecting dimension lifting.

Remark 4.1. (P0-RNN) provides an example illustrating the differences between (P) and (2.1)-(2.2) of [9]. Firstly, unifying A, V, W, b, c as θ makes it convenient to sharing parameters A, W, b in ψ_0, ψ_2 and ψ_4. Secondly, ψ_6 not only depends on θ and u_6 (i.e. s_3), but also depends on u_2, u_4 (i.e. s_1, s_2), which transmits the information across multiple layers. In contrast, DNNs in [9] demand distinct parameters in different layers, which lacks a mechanism to maintain parameter consistency among the layers sharing parameters during the training process. Figure 1 shows the architectures of RNN in (P0-RNN) and DNN in (1.1) of [9].



(a) RNN

(b) DNN

FIG. 1. *Architectures of RNN in* (P0-RNN) *and DNN in [9]*

Next, we will explore the benefits of results in Section 3 for RNN training based on (P0-RNN). For simplicity, we merge penalty parameters $\{\beta_\ell, \ell \in [L]\}$ into $(\beta_1, \beta_2) > 0$ in the ℓ_1-penalized form of (P0-RNN):

$$\text{(P1-RNN)} \qquad \min_{z \in \mathbb{R}^{\bar{N}}} \Theta(z),$$

where

$$\begin{aligned}\Theta(z) &:= F(z) + \beta_1 \sum_{\ell=1}^{6} \|u_\ell - \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})\|_1 + \beta_2 \sum_{\ell=7}^{8} \|u_\ell - \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})\|_1 \\ &= F(z) + \beta_1 \sum_{t=1}^{3} (\|w_t - W s_{t-1} - A x_t - b\|_1 + \|s_t - \sigma(w_t)\|_1) \\ &\quad + \beta_2 \sum_{t=1}^{3} (\|v_t - V s_t - c\|_1 + \|r_t - \sigma(v_t)\|_1)\,.\end{aligned}$$

For (P0-RNN) and (P1-RNN), denote

$$\begin{aligned}\mathcal{F}_0^{RNN} &:= \left\{ z \,\middle|\, \begin{array}{l} w_t = W s_{t-1} + A x_t + b,\ s_t = \sigma(w_t), \\ v_t = V s_t + c,\ r_t = \sigma(v_t),\ t \in [3] \end{array} \right\}, \\ \mathcal{S}_0^{RNN} &:= \operatorname*{argmin}_{z \in \mathcal{F}_0^{RNN}} F(z),\ \mathcal{S}_1^{RNN} := \operatorname*{argmin}_{z \in \mathbb{R}^{\bar{N}}} \Theta(z),\end{aligned}$$

and break the direction $d \in \mathbb{R}^{\bar{N}}$ according to the blocks of variable z defined in (4.2):

$$\begin{aligned} & d = (d_\theta^\top, d_u^\top)^\top, \text{ with} \\ & d_\theta = \left(d_A^\top, d_V^\top, d_W^\top, d_b^\top, d_c^\top\right)^\top, \\ & d_u = (d_{u_1}^\top, \ldots, d_{u_8}^\top)^\top = (d_{w_1}^\top, d_{s_1}^\top, d_{w_2}^\top, d_{s_2}^\top, d_{w_3}^\top, d_{s_3}^\top, d_v^\top, d_r^\top)^\top, \end{aligned} \tag{4.3}$$

where all dimensions are consistent with the corresponding variables. And for any $d \in \mathbb{R}^{\bar{N}}$, we define $d_r := (d_{r_1}^\top, d_{r_2}^\top, d_{r_3}^\top)^\top$, $d_v := (d_{v_1}^\top, d_{v_2}^\top, d_{v_3}^\top)^\top$, $d_s := (d_{s_1}^\top, d_{s_2}^\top, d_{s_3}^\top)^\top$, $d_w := (d_{w_1}^\top, d_{w_2}^\top, d_{w_3}^\top)^\top$, and $d_{s_0} := \mathbf{0} \in \mathbb{R}^{N_1}$. For any vector $a \in \mathbb{R}^{pq}$, denote

$$\mathrm{mat}_{p,q}(a) := \begin{pmatrix} a_1 & a_{p+1} & \cdots & a_{p(q-1)+1} \\ \vdots & \vdots & \vdots & \vdots \\ a_p & a_{2p} & \cdots & a_{pq} \end{pmatrix}.$$

Applying the results in Sections 2 and 3 on (P0-RNN) and (P1-RNN), we obtain the following corollary.

COROLLARY 4.2. *(i) The optimal solution sets $\mathcal{S}_0^{RNN}$ and $\mathcal{S}_1^{RNN}$ are nonempty and compact for all $(\beta_1, \beta_2) > 0$. (ii) Any local minimizer z of* (P0-RNN) *is a second-order d-stationary point of* (P0-RNN)*, that is*

$$z \in \mathcal{F}_0^{RNN} \text{ and } [\nabla F(z)]^\top d \geq 0,\ \forall d \in \mathcal{T}_{\mathcal{F}_0^{RNN}}(z)$$
$$\text{and } d^\top \nabla^2 F(z) d \geq 0,\ \forall d \in P_{\mathcal{F}_0^{RNN}}(z) \text{ with } [\nabla F(z)]^\top d = 0,$$

and for all $(\beta_1, \beta_2) > 0$, any local minimizer z of (P1-RNN) *is a second-order d-stationary point of* (P1-RNN)*, that is*

$$\Theta'(z; d) \geq 0,\ \forall d \in \mathbb{R}^{\bar{N}} \text{ and } \Theta^{(2)}(z; d) \geq 0,\ \forall d \text{ with } \Theta'(z; d) = 0,$$

where for any $z \in \mathcal{F}_0^{RNN}$,

$$\mathcal{T}_{\mathcal{F}_0^{RNN}}(z) = \left\{ d \in \mathbb{R}^{\bar{N}} \,\middle|\, \begin{array}{c} d_{v_t} = D_V s_t + V d_{s_t} + d_c,\ d_{r_t} = \sigma'(v_t; d_{v_t}), \\ d_{w_t} = D_W s_{t-1} + W d_{s_{t-1}} + D_A x_t + d_b, \\ d_{s_t} = \sigma'(w_t; d_{w_t}),\ t \in [3] \end{array} \right\},$$
$$P_{\mathcal{F}_0^{RNN}}(z) = \{ d \in \mathcal{T}_{\mathcal{F}_0^{RNN}}(z) \mid D_V d_{s_t} = 0,\ D_W d_{s_{t-1}} = 0,\ t \in [3] \}$$

with $D_A := \mathrm{mat}_{N_1,N_0}(d_A)$, $D_V := \mathrm{mat}_{N_2,N_1}(d_V)$, $D_W := \mathrm{mat}_{N_1,N_1}(d_W)$.
(iii) Under the setting of

$$\beta_1 > \gamma_1 \gamma_y \sqrt{2/(3\lambda)}, \quad \beta_2 > \sqrt{2\gamma_y/3}, \tag{4.4}$$

where $\gamma_y := \Theta(\mathbf{0}) = \|y\|^2/6, \gamma_1 := \sum_{i=0}^{2} (\sqrt{\gamma_y/\lambda})^i$, we have that

(a) $\mathcal{S}_0^{RNN} = \mathcal{S}_1^{RNN}$; and

(b) for any $z \in lev_{\leq \gamma_y} \Theta$, z is d-stationary point of (P0-RNN) *if and only if it is a d-stationary point of* (P1-RNN).

Proof. (i). The nonemptiness and compactness of $\mathcal{S}_0^{RNN}$ and $\mathcal{S}_1^{RNN}$ come from Lemmas 2.1 and 2.2.

(ii). Since g and $\{\psi_{\ell-1}, \ell \in [8]\}$ satisfy Assumptions 1 and 2, we attain the necessity of second-order d-stationarity from Lemma 3.1, and the expression of $\mathcal{T}_{\mathcal{F}_0^{RNN}}(z)$ from Theorem 3.6. The expression of $P_{\mathcal{F}_0^{RNN}}(z)$ is further derived by its definition, $P_{\mathcal{F}_0^{RNN}}(z) \subseteq \mathcal{T}_{\mathcal{F}_0^{RNN}}(z)$ and (3.22).

(iii). The thresholds (4.4) can be obtained by refining the proof of Lemma 3.8 and Theorem 3.9 as follows. Note that

$$\|r - y\| \leq \sqrt{6\gamma_y},\ \|V\| \leq \sqrt{\gamma_y/\lambda},\ \|W\| \leq \sqrt{\gamma_y/\lambda},$$

for all $z \in lev_{\leq \gamma_y}\Theta$. Next, we will estimate the constants $K_g > 0$ and $\{K_\ell > 0, \ell \in [L-1]\}$ satisfying (2.11). By the definition of $g(u)$ in (P0-RNN) and (P1-RNN), it implies that

$$|g'(u;d_u) - g'(u;\bar{d}_u)| = |\frac{(r-y)^\top(d_r - \bar{d}_r)}{3}| \leq \frac{\|r-y\|}{3}\|d_r - \bar{d}_r\| \leq \sqrt{\frac{2\gamma_y}{3}}\|d_r - \bar{d}_r\| \tag{4.5}$$

for any $z \in lev_{\leq \gamma_y}\Theta$ and any $d, \bar{d} \in \mathbb{R}^{\bar{N}}$. Then for $\{\psi_\ell, \ell \in [7]\}$, we divide them into four groups. For $\{\psi_{2t-2}, t = 2,3\}$, it follows from $\|W\| \leq \sqrt{\gamma_y/\lambda}$ for all $z \in lev_{\leq \gamma_y}\Theta$ that

$$\begin{aligned} &\|\psi'_{2t-2}(\theta, \mathbf{u}_{2t-2}; d_\theta, d_{\mathbf{u}_{2t-2}}) - \psi'_{2t-2}(\theta, \mathbf{u}_{2t-2}; \bar{d}_\theta, \bar{d}_{\mathbf{u}_{2t-2}})\| \\ &= \big\|W(d_{s_{t-1}} - \bar{d}_{s_{t-1}})\big\| \leq \sqrt{\gamma_y/\lambda}\|d_{s_{t-1}} - \bar{d}_{s_{t-1}}\| \end{aligned} \tag{4.6}$$

for all $d, \bar{d} \in \mathbb{R}^{\bar{N}}$ with $d_\theta = \bar{d}_\theta$. For $\{\psi_{2t-1}, t \in [3]\}$, it follows from the definition of σ that

$$\begin{aligned} &\|\psi'_{2t-1}(\theta, \mathbf{u}_{2t-1}; d_\theta, d_{\mathbf{u}_{2t-1}}) - \psi'_{2t-1}(\theta, \mathbf{u}_{2t-1}; \bar{d}_\theta, \bar{d}_{\mathbf{u}_{2t-1}})\| \\ &= \|\sigma'(w_t; d_{w_t}) - \sigma'(w_t; \bar{d}_{w_t})\| \leq \|d_{w_t} - \bar{d}_{w_t}\| \end{aligned} \tag{4.7}$$

for all $d, \bar{d} \in \mathbb{R}^{\bar{N}}$. For ψ_6, it follows from $\|V\| \leq \sqrt{\gamma_y/\lambda}$ that for all $z \in lev_{\leq \gamma_y}\Theta$ and for all $d, \bar{d} \in \mathbb{R}^{\bar{N}}$ with $d_\theta = \bar{d}_\theta$,

$$\begin{aligned} &\|\psi'_6(\theta, \mathbf{u}_6; d_\theta, d_{\mathbf{u}_6}) - \psi'_6(\theta, \mathbf{u}_6; \bar{d}_\theta, \bar{d}_{\mathbf{u}_6})\| \\ &= \left\| \begin{pmatrix} V(d_{s_1} - \bar{d}_{s_1}) \\ V(d_{s_2} - \bar{d}_{s_2}) \\ V(d_{s_3} - \bar{d}_{s_3}) \end{pmatrix} \right\| \leq \sqrt{\gamma_y/\lambda} \sum_{t\in[3]} \|d_{s_t} - \bar{d}_{s_t}\|. \end{aligned} \tag{4.8}$$

For ψ_7, it follows from the definition of σ that

$$\|\psi'_7(\theta, \mathbf{u}_7; d_\theta, d_{\mathbf{u}_7}) - \psi'_7(\theta, \mathbf{u}_7; \bar{d}_\theta, \bar{d}_{\mathbf{u}_7})\| = \|\sigma'(v; d_v) - \sigma'(v; \bar{d}_v)\| \leq \|d_v - \bar{d}_v\| \tag{4.9}$$

for all $d, \bar{d} \in \mathbb{R}^{\bar{N}}$. Then we can yield (a) and (b) under the thresholds (4.4) by replacing (2.11) used in Lemma 3.8 and Theorem 3.9 with (4.5)-(4.9) as follows.

- For Lemma 3.8, prove $u_\ell = \psi_{\ell-1}(\theta, \mathbf{u}_{\ell-1})$ in the order of $\ell = 8, \ldots, 1$ by contradiction, but without the use of induction (3.12), under same definitions of $\bar{z}$. During the process, plug (4.5) with $\bar{d} = \mathbf{0}$ into (3.7) and (3.9); in (3.11) and (3.13), use
  - (4.6) with $d_\theta = \mathbf{0}, \bar{d} = \mathbf{0}$, $t = 2, 3$,
  - (4.7) with $d_\theta = \mathbf{0}, \bar{d} = \mathbf{0}$, $t = 1, 2, 3$,
  - (4.8) with $d_\theta = \mathbf{0}, \bar{d} = \mathbf{0}$,
  - (4.9) with $d_\theta = \mathbf{0}, \bar{d} = \mathbf{0}$.
- For Theorem 3.9, keep the analysis before (3.19) unchanged except for plugging (4.5) into (3.17). Then repeat (3.20) for $\ell = 2, \ldots, 8$ instead of using the induction (3.19). During the process, we use (4.6), (4.7), (4.8) and (4.9) when the subscript of ψ belongs to $\{2,4\}$, $\{1,3,5\}$, $\{6\}$ and $\{7\}$ in (3.20) respectively. The calculations after (3.20) are also kept without changes. □

The results in Corollary 4.2 can be easily extended to the case where $N > 1$, $T > 3$ and s_t, w_t, r_t, v_t aggregate corresponding components for all samples at tth time step with thresholds

$$\beta_1 > \gamma_1\gamma_y\sqrt{2/(\lambda NT)}, \qquad \beta_2 > \sqrt{2\gamma_y/(NT)}, \tag{4.10}$$

under $\gamma_y := \Theta(\mathbf{0}) = \|y\|^2/(2NT)$, $\gamma_1 := \sum_{i=0}^{T-1}(\sqrt{\gamma_y/\lambda})^i$. And the exact penalty in d-stationarity can be generalized to more scenarios including but not limited to more complicated variants in RNNs (such as LSTM and GRU) with any locally Lipschitz continuous and directionally differentiable activator (such as tanh and ELU).

Besides, it follows from the convexity of F in (P0-RNN) and (P1-RNN) that for all $z \in \mathcal{F}_0^{RNN}$, $d^\top\nabla^2 F(z)d \geq 0$ for all $d \in P_{\mathcal{F}_0^{RNN}}(z) \cap \{d \mid [\nabla F(z)]^\top d = 0\}$. Hence, every d-stationary point of (P0-RNN) is a second-order d-stationary point for (P0-RNN). Similarly, according to (3.24), every d-stationary point of (P1-RNN) in $lev_{\leq\gamma_y}\Theta$ is a second-order d-stationary point for (P1-RNN) under (4.10). In fact, it follows from Remark 3.12 that $\mathcal{SD}_0 \cap lev_{\leq\gamma_y}\Theta = \mathcal{SD}_1 \cap lev_{\leq\gamma_y}\Theta = \mathcal{D}_0 \cap lev_{\leq\gamma_y}\Theta = \mathcal{D}_1 \cap lev_{\leq\gamma_y}\Theta$ in this case. As a consequence, one can obtain a second-order d-stationary point of (P0-RNN) and (P1-RNN) by applying the algorithms in [9, 11] on (P1-RNN) with (4.10).

5. Conclusions. The paper investigates a class of nonconvex nonsmooth multicomposite optimization problems (P) with an objective function comprised of a regularization term and a multi-layer composite function with twice directionally differentiable and locally Lipschitz continuous components. The d-stationarity of (P) is hard to attain directly, and its second-order d-stationarity is vague without additional assumptions on the objective function. Based on the closed-form expression of the tangent cone $\mathcal{T}_{\mathcal{F}_0}(\cdot)$, we prove the equivalence between (P), the constrained form (P0) and the ℓ_1-penalty formulation (P1) in terms of global optimality and d-stationarity. The equivalence offers an indirect way to compute the d-stationary points of (P). Furthermore, it provides second-order necessary and sufficient conditions for (P) through (P0) and (P1). The theoretical results are also applied to the training process of recurrent neural networks.

Acknowledgments. We would like to thank two referees for their constructive and helpful comments.

REFERENCES

[1] J. H. Alcantara, C. pei Lee, and A. Takeda, *A four-operator splitting algorithm for nonconvex and nonsmooth optimization*, to appear in SIAM J. Optim., (2025).
[2] P. I. Barton, K. A. Khan, P. Stechlinski, and H. A. Watson, *Computationally relevant generalized derivatives: theory, evaluation and applications*, Optim. Methods Softw., 33 (2018), pp. 1030–1072.
[3] J. Bolte, T. Le, E. Pauwels, and T. Silveti-Falls, *Nonsmooth implicit differentiation for machine-learning and optimization*, in Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 13537–13549.
[4] J. Bolte and E. Pauwels, *Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning*, Math. Program., 188 (2021), pp. 19–51.
[5] J. Borwein, *Necessary and sufficient conditions for quadratic minimality*, Numer. Funct. Anal. Optim., 5 (1982), pp. 127–140.
[6] M. Carreira-Perpinan and W. Wang, *Distributed optimization of deeply nested systems*, in Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, vol. 33, PMLR, 2014, pp. 10–19.

[7] X. Chen, *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Program., 134 (2012), pp. 71–99.
[8] Y. Cui, T.-H. Chang, M. Hong, and J.-S. Pang, *A study of piecewise linear-quadratic programs*, J. Optim. Theory Appl., 186 (2020), pp. 523–553.
[9] Y. Cui, Z. He, and J.-S. Pang, *Multicomposite nonconvex optimization for training deep neural networks*, SIAM J. Optim., 30 (2020), pp. 1693–1723.
[10] Y. Cui and J.-S. Pang, *Modern Nonconvex Nondifferentiable Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021.
[11] Y. Cui, J.-S. Pang, and B. Sen, *Composite difference-max programs for modern statistical estimation problems*, SIAM J. Optim., 28 (2018), pp. 3344–3374.
[12] J. L. Elman, *Finding structure in time*, Cogn. Sci., 14 (1990), pp. 179–211.
[13] A. Graves, *Supervised sequence labelling with recurrent neural networks*, in Studies in Computational Intelligence, 2012.
[14] B. Hammer, *On the approximation capability of recurrent neural networks*, Neurocomputing, 31 (2000), pp. 107–123.
[15] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural Comput., 9 (1997), pp. 1735–1780.
[16] J. Jiang and X. Chen, *Optimality conditions for nonsmooth nonconvex-nonconcave min-max problems and generative adversarial networks*, SIAM J. Math. Data Sci., 5 (2023), pp. 693–722.
[17] M. I. Jordan, *Attractor dynamics and parallelism in a connectionist sequential machine*, IEEE Press, Artificial Neural Networks: Concept Learning, 1990, pp. 112–127.
[18] K. A. Khan and P. I. Barton, *A vector forward mode of automatic differentiation for generalized derivative evaluation*, Optim. Methods Softw., 30 (2015), pp. 1185–1212.
[19] J. Liang, R. D. C. Monteiro, and H. Zhang, *Proximal bundle methods for hybrid weakly convex composite optimization problems*, arxiv:2303.14896, (2024).
[20] T. Lin, Z. Zheng, and M. Jordan, *Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization*, in Advances in Neural Information Processing Systems, vol. 35, Curran Associates, Inc., 2022, pp. 26160–26175.
[21] W. Liu, X. Liu, and X. Chen, *Linearly constrained nonsmooth optimization for training autoencoders*, SIAM J. Optim., 32 (2022), pp. 1931–1957.
[22] W. Liu, X. Liu, and X. Chen, *An inexact augmented Lagrangian algorithm for training leaky ReLU neural network with group sparsity*, J. Mach. Learn. Res., 24 (2023), pp. 1–43.
[23] Y. Nesterov, *Lexicographic differentiation of nonsmooth functions*, Math. Program., 104 (2005), pp. 669–700.
[24] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, New York, 2006.
[25] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer Science & Business Media, Berlin, 2009.
[26] K. Wang and L. T. Biegler, *MPCC strategies for nonsmooth nonlinear programs*, Optim. Eng., 24 (2023), pp. 1883–1929.
[27] Y. Wang, C. Zhang, and X. Chen, *An augmented Lagrangian method for training recurrent neural networks*, SIAM J. Sci. Comput., 47 (2025), pp. C22–C51.
[28] N. Xiao, X. Hu, X. Liu, and K.-C. Toh, *Adam-family methods for nonsmooth optimization with convergence guarantees*, J. Mach. Learn. Res., 25 (2024), pp. 1–53.
[29] J. J. Ye, *Multiplier rules under mixed assumptions of differentiability and Lipschitz continuity*, SIAM J. Control Optim., 39 (2000), pp. 1441–1460.
[30] Y. Yuan, *Conditions for convergence of trust region algorithms for nonsmooth optimization*, Math. Program., 31 (1985), pp. 220–228.
[31] J. Zhang, H. Lin, S. Jegelka, S. Sra, and A. Jadbabaie, *Complexity of finding stationary points of nonsmooth nonconvex functions*, in Proceedings of the 37th International Conference on Machine Learning, vol. 119, PMLR, 2020, pp. 11173–11182.