Appendix

A. Implementations on BCD methods

In this section, we provide several remarks to discuss the specific implementations of BCD methods. Reproducible PyTorch codes can be found at: https://github.com/timlautk/BCD-for-DNNs-PyTorch or https://github.com/yao-lab/BCD-for-DNNs-PyTorch.

Remark 1 (On the initialization of parameters) In practice, the weights $\{\boldsymbol{W}_i\}_{i=1}^N$ are generally initialized according to some Gaussian distributions with small standard deviations. The bias vectors are usually set as all one vectors scaled by some small constants. Given the weights and bias vectors, the auxiliary variables $\{\boldsymbol{U}_i\}_{i=1}^N$ and state variables $\{\boldsymbol{V}_i\}_{i=1}^N$ are usually initialized by a single forward pass through the network.

Remark 2 (On the update order) We suggest such a backward update order in this paper due to the nested structure of DNNs. Besides the update order presented in Algorithm 2, any arbitrary deterministic update order can be incorporated into the BCD methods, and our proofs still go through.

Remark 3 (On the distributed implementation) One major advantage of BCD is that it can be implemented in distributed and parallel manner like in ADMM. Specifically, given m servers, the total training data are distributed to these servers. Denotes S_j as the subset of samples at server j. Thus, $n = \sum_{j=1}^m \sharp(S_j)$, where $\sharp(S_j)$ denotes the cardinality of S_j. For each layer i, the state variable $\boldsymbol{V}_i$ is divided into m submatrices by column, i.e., $\boldsymbol{V}_i := ((\boldsymbol{V}_i)_{:S_1}, \ldots, (\boldsymbol{V}_i)_{:S_m})$, where $(\boldsymbol{V}_i)_{:S_j}$ denotes the submatrix of V_i including all the columns in the index set S_j. The auxiliary variables $\boldsymbol{U}_i$'s are decomposed similarly. From Algorithm 2, the updates of $\{\boldsymbol{V}_i\}_{i=1}^N$ and $\{\boldsymbol{U}_i\}_{i=1}^N$ do not need any communication and thus, can be computed in a parallel way. The difficult part is the update of weight $\boldsymbol{W}_i$, which is generally hard to parallelize. To deal with this part, there are some effective strategies suggested in the literature like Taylor et al. (2016).

B. Proof of Proposition 1

PROOF We verify these special cases as follows.

On the loss function ℓ: Since these losses are all nonnegative and continuous on their domains, they are proper lower semicontinuous and lower bounded by 0. In the following, we only verify that they are either real analytic or semialgebraic.

(a1) If $\ell(t)$ is the squared (t^2) or exponential (e^t) loss, then according to Krantz & Parks (2002), they are real analytic.

(a2) If $\ell(t)$ is the logistic loss ($\log(1+\mathrm{e}^{-t})$), since it is a composition of logarithm and exponential functions which both are real analytic, thus according to Lemma 3, the logistic loss is real analytic.

(a3) If $\ell(\boldsymbol{u};\boldsymbol{y})$ is the cross-entropy loss, i.e., given $\boldsymbol{y} \in \mathbb{R}^{d_N}$, $\ell(\boldsymbol{u};\boldsymbol{y}) = -\frac{1}{d_N}[\langle \boldsymbol{y}, \log \widehat{\boldsymbol{y}}(\boldsymbol{u})\rangle + \langle \boldsymbol{1}-\boldsymbol{y}, \log(\boldsymbol{1}-\widehat{\boldsymbol{y}}(\boldsymbol{u}))\rangle]$, where log is performed elementwise and $(\widehat{\boldsymbol{y}}(\boldsymbol{u})_i)_{1\le i\le d_N} := \left((1+\mathrm{e}^{-u_i})^{-1}\right)_{1\le i\le d_N}$ for any $\boldsymbol{u}\in\mathbb{R}^{d_N}$, which can be viewed as a linear combination of logistic functions, then by (a2) and Lemma 3, it is also analytic.

(a4) If ℓ is the hinge loss, i.e., given $\boldsymbol{y} \in \mathbb{R}^{d_N}$, $\ell(\boldsymbol{u};\boldsymbol{y}) := \max\{0, 1-\langle \boldsymbol{u},\boldsymbol{y}\rangle\}$ for any $\boldsymbol{u}\in\mathbb{R}^{d_N}$, by Lemma 4(1), it is semialgebraic, because its graph is $\mathrm{cl}(\mathcal{D})$, the closure of the set $\mathcal{D}$, where

$$\mathcal{D} = \{(\boldsymbol{u}, z) : 1-\langle \boldsymbol{u},\boldsymbol{y}\rangle - z = 0, \boldsymbol{1}-\boldsymbol{u} \succ 0\} \cup \{(\boldsymbol{u}, z) : z = 0, \langle \boldsymbol{u},\boldsymbol{y}\rangle - 1 > 0\}.$$

On the activation function σ_i: Since all the considered specific activations are continuous on their domains, they are Lipschitz continuous on any bounded set. In the following, we only need to check that they are either real analytic or semialgebraic.

(b1) If σ_i is a linear or polynomial function, then according to Krantz & Parks (2002), σ_i is real analytic.

(b2) If $\sigma_i(t)$ is sigmoid, $(1+\mathrm{e}^{-t})^{-1}$, or hyperbolic tangent, $\tanh t := \frac{\mathrm{e}^t-\mathrm{e}^{-t}}{\mathrm{e}^t+\mathrm{e}^{-t}}$, then the sigmoid function is a composition $g\circ h$ of these two functions where $g(u) = \frac{1}{1+u}, u>0$ and $h(t) = \mathrm{e}^{-t}$ (resp. $g(u) = 1-\frac{2}{u+1}, u>0$ and $h(t) = \mathrm{e}^{2t}$ in the hyperbolic tangent case). According to Krantz & Parks (2002), g and h in both cases are real analytic. Thus, according to Lemma 3, sigmoid and hyperbolic tangent functions are real analytic.

(b3) If σ_i is ReLU, i.e., $\sigma_i(u) := \max\{0, u\}$, then we can show that ReLU is semialgebraic since its graph is $\mathrm{cl}(\mathcal{D})$, the closure of the set $\mathcal{D}$, where

$$\mathcal{D} = \{(u, z) : u - z = 0, u > 0\} \cup \{(u, z) : z = 0, -u > 0\}.$$

(b4) Similar to the ReLU case, if σ_i is leaky ReLU, i.e., $\sigma_i(u) = u$ if $u > 0$, otherwise $\sigma_i(u) = au$ for some $a > 0$, then we can similarly show that leaky ReLU is semialgebraic since its graph is $\mathrm{cl}(\mathcal{D})$, the closure of the set $\mathcal{D}$, where

$$\mathcal{D} = \{(u, z) : u - z = 0, u > 0\} \cup \{(u, z) : au - z = 0, -u > 0\}.$$

(b5) If σ_i is polynomial as used in Liao & Poggio (2017), then according to Krantz & Parks (2002), it is real analytic.

(b6) If σ_i is softplus, i.e., $\sigma_i(u) = \frac{1}{t}\log(1 + \mathrm{e}^{tu})$ for some $t > 0$, since it is a composition of two analytic functions $\frac{1}{t}\log(1 + u)$ and e^{tu}, then according to Krantz & Parks (2002), it is real analytic.

On $r_i(\boldsymbol{W}_i)$, $s_i(\boldsymbol{V}_i)$: By the specific forms of these regularizers, they are nonnegative, lower semicontinuous and continuous on their domain. In the following, we only need to verify they are either real analytic and semialgebraic.

(c1) **the squared ℓ_2 norm $\|\cdot\|_2^2$:** According to Bochnak et al. (1998), the ℓ_2 norm is semialgebraic, so is its square according to Lemma 4(2), where $g(t) = t^2$ and $h(\boldsymbol{W}) = \|\boldsymbol{W}\|_2$.

(c2) **the squared Frobenius norm $\|\cdot\|_F^2$:** The squared Frobenius norm is semiaglebraic since it is a finite sum of several univariate squared functions.

(c3) **the elementwise 1-norm $\|\cdot\|_{1,1}$:** Note that $\|\boldsymbol{W}\|_{1,1} = \sum_{i,j} |\boldsymbol{W}_{ij}|$ is the finite sum of absolute functions $h(t) = |t|$. According to Lemma 4(1), the absolute value function is semialgebraic since its graph is the closure of the following semialgebraic set

$$\mathcal{D} = \{(t, s) : t + s = 0, -t > 0\} \cup \{(t, s) : t - s = 0, t > 0\}.$$

Thus, the elementwise 1-norm is semialgebraic.

(c4) **the elastic net:** Note that the elastic net is the sum of the elementwise 1-norm and the squared Frobenius norm. Thus, by (c2), (c3) and Lemma 4(3), the elastic net is semialgebraic.

(c5) If r_i or s_i is the indicator function of nonnegative closed half space or a closed interval (box constraints), by Lemma 4(1), any polyhedral set is semialgebraic such as the nonnegative orthant $\mathbb{R}^{p\times q}_+ = \{\boldsymbol{W} \in \mathbb{R}^{p\times q}, \boldsymbol{W}_{ij} \geq 0, \forall i, j\}$, and the closed interval. Thus, by Lemma 4(4), r_i or s_i is semialgebraic in this case. □

C. Proof of Theorem 1

To prove Theorem 1, we first show that the Kurdyka-Łojasiewicz (KŁ) property holds for the considered DNN training models (see Proposition 2), then establish the function value convergence of the BCD methods (see Theorem 4), followed by establishing their global convergence as well as the $\mathcal{O}(1/k)$ convergence rate to a critical point as shown in Theorem 5. Combining Proposition 2, Theorems 4 and 5 yields Theorem 1.

C.1. The Kurdyka-Łojasiewicz Property in Deep Learning

Before giving the definition of the KŁ property, we first introduce some notions and notations from variational analysis, which can be found in Rockafellar & Wets (1998).

The notion of subdifferential plays a central role in the following definition of the KŁ property. For each $\boldsymbol{x} \in \mathrm{dom}(h)$, the *Fréchet subdifferential* of h at $\boldsymbol{x}$, written $\widehat{\partial} h(\boldsymbol{x})$, is the set of vectors $\boldsymbol{v} \in \mathbb{R}^p$ which satisfy

$$\liminf_{\boldsymbol{y}\neq\boldsymbol{x}, \boldsymbol{y}\to\boldsymbol{x}} \frac{h(\boldsymbol{y}) - h(\boldsymbol{x}) - \langle \boldsymbol{v}, \boldsymbol{y} - \boldsymbol{x}\rangle}{\|\boldsymbol{x} - \boldsymbol{y}\|} \geq 0.$$

When $\boldsymbol{x} \notin \mathrm{dom}(h)$, we set $\widehat{\partial} h(\boldsymbol{x}) = \varnothing$. The *limiting-subdifferential* (or simply *subdifferential*) of h introduced in Mordukhovich (2006), written $\partial h(\boldsymbol{x})$ at $\boldsymbol{x} \in \mathrm{dom}(h)$, is defined by

$$\partial h(\boldsymbol{x}) := \{\boldsymbol{v} \in \mathbb{R}^p : \exists \boldsymbol{x}^k \to \boldsymbol{x},\ h(\boldsymbol{x}^k) \to h(\boldsymbol{x}),\ \boldsymbol{v}^k \in \widehat{\partial} h(\boldsymbol{x}^k) \to \boldsymbol{v}\}. \tag{C.1}$$

A necessary (but not sufficient) condition for $\boldsymbol{x} \in \mathbb{R}^p$ to be a minimizer of h is $\mathbf{0} \in \partial h(\boldsymbol{x})$. A point that satisfies this inclusion is called *limiting-critical* or simply *critical*. The distance between a point $\boldsymbol{x}$ to a subset $\mathcal{S}$ of $\mathbb{R}^p$, written $\mathrm{dist}(\boldsymbol{x}, \mathcal{S})$, is defined by $\mathrm{dist}(\boldsymbol{x}, \mathcal{S}) = \inf\{\|\boldsymbol{x} - \boldsymbol{s}\| : \boldsymbol{s} \in \mathcal{S}\}$, where $\|\cdot\|$ represents the Euclidean norm.

The KŁ property (Łojasiewicz, 1963; 1993; Kurdyka, 1998; Bolte et al., 2007a;b) plays a central role in the convergence analysis of nonconvex algorithms (see e.g., Attouch et al., 2013; Xu & Yin, 2013; Wang et al., 2019). The following definition is adopted from Bolte et al. (2007a).

Definition 3 (Kurdyka-Łojasiewicz property) A function $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is said to have the **Kurdyka-Łojasiewicz (KŁ) property** at $\boldsymbol{x}^* \in \mathrm{dom}(\partial h)$ if there exist a neighborhood U of $\boldsymbol{x}^*$, a constant η, and a continuous concave function $\varphi(s) = cs^{1-\theta}$ for some $c > 0$ and $\theta \in [0, 1)$ such that the Kurdyka-Łojasiewicz inequality holds: For all $\boldsymbol{x} \in U \cap \mathrm{dom}(\partial h)$ and $h(\boldsymbol{x}^*) < h(\boldsymbol{x}) < h(\boldsymbol{x}^*) + \eta$,

$$\varphi'(h(\boldsymbol{x}) - h(\boldsymbol{x}^*)) \cdot \mathrm{dist}(\mathbf{0}, \partial h(\boldsymbol{x})) \geq 1, \tag{C.2}$$

where θ is called the KŁ exponent of h at $\boldsymbol{x}^*$. Proper lower semi-continuous functions which satisfy the Kurdyka-Łojasiewicz inequality at each point of $\mathrm{dom}(\partial h)$ are called KŁ functions.

Note that we have adopted in the definition of the KŁ inequality (C.2) the following notational conventions: $0^0 = 1, \infty/\infty = 0/0 = 0$. Such property was firstly introduced by Łojasiewicz (1993) on real analytic functions (Krantz & Parks, 2002) for $\theta \in \left[\frac{1}{2}, 1\right)$, then was extended to functions defined on the o-minimal structure in Kurdyka (1998), and later was extended to nonsmooth subanalytic functions in Bolte et al. (2007a).

By the definition of the KŁ property, it means that the function under consideration is sharp up to a reparametrization (Attouch et al., 2013). Particularly, when h is smooth, finite-valued, and $h(\boldsymbol{x}^*) = 0$, the inequality (C.2) can be rewritten

$$\|\nabla(\varphi \circ h)(\boldsymbol{x})\| \geq 1,$$

for each convenient $\boldsymbol{x} \in \mathbb{R}^p$. This inequality may be interpreted as follows: up to the reparametrization of the values of h via φ, we face a *sharp function*. Since the function φ is used here to turn a singular region—a region in which the gradients are arbitrarily small—into a regular region, i.e., a place where the gradients are bounded away from zero, it is called a *desingularizing* function for h. For theoretical and geometrical developments concerning this inequality, see Bolte et al. (2007b). KŁ functions include real analytic functions (see Definition 1), semialgebraic functions (see Definition 2), tame functions defined in some o-minimal structures (Kurdyka, 1998), continuous subanalytic functions (Bolte et al., 2007a;b) and locally strongly convex functions (Xu & Yin, 2013).

In the following, we establish the KŁ properties[13] of the DNN training models with variable splitting, i.e., the functions $\mathcal{L}$ defined in (2.3) and $\overline{\mathcal{L}}$ defined in (2.5).

Proposition 2 (KŁ properties of deep learning) *Suppose that Assumption 1 hold. Then the functions $\mathcal{L}$ defined in (2.3), and $\overline{\mathcal{L}}$ defined in (2.5) when restricted to any closed set are KŁ functions.*

This proposition shows that most of the DNN training models with variable splitting have some *nice* geometric properties, i.e., they are amenable to sharpness at each point in their domains. In order to prove this theorem, we need the following lemmas. The first lemma shows some important properties of real analytic functions.

Lemma 3 (Krantz & Parks, 2002) *The sums, products, and compositions of real analytic functions are real analytic functions.*

Then we present some important properties of semialgebraic sets and mappings, which can be found in Bochnak et al. (1998).

[13] It should be pointed out that we need to use the vectorization of the matrix variables involved in $\mathcal{L}$, $\overline{\mathcal{L}}$ and $\overline{\mathcal{L}}_{\mathrm{res}}$ in order to adopt the existing definitions of KŁ property, real analytic functions and semialgebraic functions. We still use the matrix notation for the simplicity of notation.

Lemma 4 *The following hold*

(1) The finite union, finite intersection, and complement of semialgebraic sets are semialgebraic. The closure and the interior of a semialgebraic set are semialgebraic (Bochnak et al., 1998, Proposition 2.2.2).

(2) The composition $g \circ h$ of semialgebraic mappings $h : A \to B$ and $g : B \to C$ is semialgebraic (Bochnak et al., 1998, Proposition 2.2.6).

(3) The sum of two semialgebraic functions is semialgebraic (can be referred to the proof of Bochnak et al. 1998, Proposition 2.2.6).

(4) The indicator function of a semialgebraic set is semialgebraic (Bochnak et al., 1998).

Since our proof involves the sum of real analytic functions and semialgebraic functions, we still need the following lemma, of which the claims can be found in or derived directly from Shiota (1997).

Lemma 5 *The following hold:*

(1) Both real analytic functions and semialgebraic functions (mappings) are subanalytic (Shiota, 1997).

(2) Let f_1 and f_2 are both subanalytic functions, then the sum of $f_1 + f_2$ is a subanalytic function if at least one of them map a bounded set to a bounded set or if both of them are nonnegative (Shiota, 1997, p.43).

Moreover, we still need the following important lemma from Bolte et al. (2007a), which shows that the subanalytic function is a KŁ function.

Lemma 6 (Bolte et al., 2007a, Theorem 3.1) *Let $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be a subanalytic function with closed domain, and assume that h is continuous on its domain, then h is a KŁ function.*

PROOF (PROOF OF PROPOSITION 2) We first verify the KŁ property of $\overline{\mathcal{L}}$, and then similarly show that of $\mathcal{L}$. From (2.5),

$$\begin{aligned}&\overline{\mathcal{L}}(\{\boldsymbol{W}_i\}_{i=1}^N, \{\boldsymbol{V}_i\}_{i=1}^N, \{\boldsymbol{U}_i\}_{i=1}^N)\\ &\qquad := \mathcal{R}_n(\boldsymbol{V}_N; \boldsymbol{Y}) + \sum_{i=1}^N r_i(\boldsymbol{W}_i) + \sum_{i=1}^N s_i(\boldsymbol{V}_i) + \frac{\gamma}{2}\sum_{i=1}^N \left[\|\boldsymbol{V}_i - \sigma_i(\boldsymbol{U}_i)\|_F^2 + \|\boldsymbol{U}_i - \boldsymbol{W}_i\boldsymbol{V}_{i-1}\|_F^2\right],\end{aligned}$$

which mainly includes the following types of functions, i.e.,

$$\mathcal{R}_n(\boldsymbol{V}_N; \boldsymbol{Y}),\ r_i(\boldsymbol{W}_i),\ s_i(\boldsymbol{V}_i),\ \|\boldsymbol{V}_i - \sigma_i(\boldsymbol{U}_i)\|_F^2,\ \|\boldsymbol{U}_i - \boldsymbol{W}_i\boldsymbol{V}_{i-1}\|_F^2.$$

To verify the KŁ property of the function $\overline{\mathcal{L}}$, we consider the above functions one by one under the hypothesis of Proposition 2.

On $\mathcal{R}_n(\boldsymbol{V}_N; \boldsymbol{Y})$: Note that given the output data $\boldsymbol{Y}$, $\mathcal{R}_n(\boldsymbol{V}_N; \boldsymbol{Y}) := \frac{1}{n}\sum_{j=1}^n \ell((\boldsymbol{V}_N)_{:j}, \boldsymbol{y}_j)$, where $\ell : \mathbb{R}^{d_N} \times \mathbb{R}^{d_N} \to \mathbb{R}_+ \cup \{0\}$ is some loss function. If ℓ is real analytic (resp. semialgebraic), then by Lemma 3 (resp. Lemma 4(3)), $\mathcal{R}_n(\boldsymbol{V}_N; \boldsymbol{Y})$ is real-analytic (resp. semialgebraic).

On $\|\boldsymbol{V}_i - \sigma_i(\boldsymbol{U}_i)\|_F^2$: Note that $\|\boldsymbol{V}_i - \sigma_i(\boldsymbol{U}_i)\|_F^2$ is a finite sum of simple functions of the form, $|v - \sigma_i(u)|^2$ for any $u, v \in \mathbb{R}$. If σ_i is real analytic (resp. semialgebraic), then $v - \sigma_i(u)$ is real analytic (resp. semialgebraic), and further by Lemma 3 (resp. Lemma 4(2)), $|v - \sigma_i(u)|^2$ is also real analytic (resp. semialgebraic) since $|v - \sigma_i(u)|^2$ can be viewed as the composition $g \circ h$ of these two functions where $g(t) = t^2$ and $h(u, v) = v - \sigma_i(u)$.

On $\|\boldsymbol{U}_i - \boldsymbol{W}_i\boldsymbol{V}_{i-1}\|_F^2$: Note that the function $\|\boldsymbol{U}_i - \boldsymbol{W}_i\boldsymbol{V}_{i-1}\|_F^2$ is a polynomial function with the variables $\boldsymbol{U}_i, \boldsymbol{W}_i$ and $\boldsymbol{V}_{i-1}$, and thus according to Krantz & Parks (2002) and Bochnak et al. (1998), it is both real analytic and semialgebraic.

On $r_i(\boldsymbol{W}_i)$, $s_i(\boldsymbol{V}_i)$: All r_i's and s_i's are real analytic or semialgebraic by the hypothesis of Proposition 2.

Since each part of the function $\overline{\mathcal{L}}$ is either real analytic or semialgebraic, then by Lemma 5, $\overline{\mathcal{L}}$ is a subanalytic function. Furthermore, by the continuity hypothesis of Proposition 2, $\overline{\mathcal{L}}$ is continuous in its domain. Therefore, $\overline{\mathcal{L}}$ is a KŁ function according to Lemma 6.

Similarly, we can verify the KŁ property of $\mathcal{L}$ by checking each part is either real analytic or semialgebraic. The major task is to check the KŁ properties of the functions $\|\boldsymbol{V}_i - \sigma_i(\boldsymbol{W}_i\boldsymbol{V}_{i-1})\|_F^2$ $(i = 1, \ldots, N)$. This reduces to check the function $h : \mathbb{R} \times \mathbb{R}^{d_{i-1}} \times \mathbb{R}^{d_{i-1}} \to \mathbb{R}$, $h(u, \boldsymbol{v}, \boldsymbol{w}) := |u - \sigma_i(\langle \boldsymbol{w}, \boldsymbol{v}\rangle)|^2$. Similar to the case $|v - \sigma_i(u)|^2$ for any $u, v \in \mathbb{R}$ in $\overline{\mathcal{L}}$, h is real analytic (resp. semialgebraic) if σ_i is real analytic (resp. semialgebraic) by Lemma 3 (resp. Lemma 4(2)). As a consequence, each part of the function $\mathcal{L}$ is either real analytic or semialgebraic, so $\mathcal{L}$ is a subanalytic function, and further by the continuity hypothesis of Proposition 2, $\mathcal{L}$ is a KŁ function according to Lemma 6. This completes the proof. □

C.2. Value convergence of BCD

We show the value convergence of both algorithms as follows.

Theorem 4 *Let $\{\mathcal{Q}^k := \big(\{\boldsymbol{W}_i^k\}_{i=1}^N, \{\boldsymbol{V}_i^k\}_{i=1}^N\big)\}_{k\in\mathbb{N}}$ and $\{\mathcal{P}^k := \big(\{\boldsymbol{W}_i^k\}_{i=1}^N, \{\boldsymbol{V}_i^k\}_{i=1}^N, \{\boldsymbol{U}_i^k\}_{i=1}^N\big)\}_{k\in\mathbb{N}}$ be the sequences generated by Algorithms 1 and 2, respectively. Under Assumption 1 and finite initializations $\mathcal{Q}^0$ and $\mathcal{P}^0$, then for any positive α and γ, $\{\mathcal{L}(\mathcal{Q}^k)\}_{k\in\mathbb{N}}$ (resp. $\{\overline{\mathcal{L}}(\mathcal{P}^k)\}_{k\in\mathbb{N}}$) is nonincreasing and converges to some finite $\mathcal{L}^*$ (resp. $\overline{\mathcal{L}}^*$).*

In order to prove Theorem 4, we first show the convergence of Algorithm 2 and then show that of Algorithm 1 similarly. We restate Lemma 1 precisely as follows.

Lemma 7 (Restate of Lemma 1) *Let $\{\mathcal{P}^k\}_{k\in\mathbb{N}}$ be a sequence generated by the BCD method (Algorithm 2), then*

$$\overline{\mathcal{L}}(\mathcal{P}^k) \le \overline{\mathcal{L}}(\mathcal{P}^{k-1}) - a\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F^2, \tag{C.3}$$

where

$$a := \min\left\{\frac{\alpha}{2}, \frac{\gamma}{2}\right\} \tag{C.4}$$

for the case that $\boldsymbol{V}_N$ is updated via the proximal strategy, or

$$a := \min\left\{\frac{\alpha}{2}, \frac{\gamma}{2}, \alpha + \frac{\gamma - L_R}{2}\right\} \tag{C.5}$$

for the case that $\boldsymbol{V}_N$ is update via the prox-linear strategy.

According to Algorithm 2, the decreasing property of the sequence $\{\overline{\mathcal{L}}(\mathcal{P}^k)\}_{k\in\mathbb{N}}$ is obvious. However, establishing the sufficient descent inequality (C.3) for the sequence $\{\overline{\mathcal{L}}(\mathcal{P}^k)\}_{k\in\mathbb{N}}$ is nontrivial. To achieve this, we should take advantage of the specific update strategies and also the form of $\overline{\mathcal{L}}$ as shown in the following proofs.

PROOF The descent quantity in (C.3) can be developed via considering the descent quantity along the update of each block variable. From Algorithm 2, each block variable is updated either by the proximal strategy with parameter $\alpha/2$ (say, updates of $\boldsymbol{V}_N^k, \{\boldsymbol{U}_i^k\}_{i=1}^{N-1}, \{\boldsymbol{W}_i^k\}_{i=1}^N$-blocks in Algorithm 2) or by minimizing a strongly convex function[14] with parameter $\gamma > 0$ (say, updates of $\{\boldsymbol{V}_i^k\}_{i=1}^{N-1}, \boldsymbol{U}_N^k$-blocks in Algorithm 2), we will consider both cases one by one.

(a) Proximal update case: In this case, we take the $\boldsymbol{W}_i^k$-update case for example. By Algorithm 2, $\boldsymbol{W}_i^k$ is updated according to the following

$$\boldsymbol{W}_i^k \leftarrow \operatorname*{argmin}_{\boldsymbol{W}_i}\left\{r_i(\boldsymbol{W}_i) + \frac{\gamma}{2}\|\boldsymbol{U}_i^k - \boldsymbol{W}_i\boldsymbol{V}_{i-1}^{k-1}\|_F^2 + \frac{\alpha}{2}\|\boldsymbol{W}_i - \boldsymbol{W}_i^{k-1}\|_F^2\right\}. \tag{C.6}$$

Let $h^k(\boldsymbol{W}_i) = r_i(\boldsymbol{W}_i) + \frac{\gamma}{2}\|\boldsymbol{U}_i^k - \boldsymbol{W}_i\boldsymbol{V}_{i-1}^{k-1}\|_F^2$ and $\bar{h}^k(\boldsymbol{W}_i) = r_i(\boldsymbol{W}_i) + \frac{\gamma}{2}\|\boldsymbol{U}_i^k - \boldsymbol{W}_i\boldsymbol{V}_{i-1}^{k-1}\|_F^2 + \frac{\alpha}{2}\|\boldsymbol{W}_i - \boldsymbol{W}_i^{k-1}\|_F^2$. By the optimality of $\boldsymbol{W}_i^k$, the following holds

$$\bar{h}^k(\boldsymbol{W}_i^{k-1}) \ge \bar{h}^k(\boldsymbol{W}_i^k),$$

which implies

$$h^k(\boldsymbol{W}_i^{k-1}) \ge h^k(\boldsymbol{W}_i^k) + \frac{\alpha}{2}\|\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1}\|_F^2. \tag{C.7}$$

[14] The function h is called a strongly convex function with parameter $\gamma > 0$ if $h(u) \ge h(v) + \langle \nabla h(v), u - v\rangle + \frac{\gamma}{2}\|u - v\|^2$.

Note that the $\boldsymbol{W}_i^k$-update (C.6) is equivalent to the following original proximal BCD update, i.e.,

$$\boldsymbol{W}_i^k \leftarrow \underset{\boldsymbol{W}_i}{\operatorname{argmin}} \overline{\mathcal{L}}(\boldsymbol{W}_{<i}^{k-1}, \boldsymbol{W}_i, \boldsymbol{W}_{>i}^k, \boldsymbol{V}_{<i}^{k-1}, \boldsymbol{V}_i^k, \boldsymbol{V}_{>i}^k, \boldsymbol{U}_{<i}^{k-1}, \boldsymbol{U}_i^k, \boldsymbol{U}_{>i}^k) + \frac{\alpha}{2}\|\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1}\|_F^2,$$

where $\boldsymbol{W}_{<i} := (\boldsymbol{W}_1, \boldsymbol{W}_2, \ldots, \boldsymbol{W}_{i-1})$, $\boldsymbol{W}_{>i} := (\boldsymbol{W}_{i+1}, \boldsymbol{W}_{i+1}, \ldots, \boldsymbol{W}_N)$, and $\boldsymbol{V}_{<i}, \boldsymbol{V}_{>i}, \boldsymbol{U}_{<i}, \boldsymbol{U}_{>i}$ are defined similarly. Thus, by (C.7), we establish the descent part along the $\boldsymbol{W}_i$-update ($i = 1, \ldots, N-1$), i.e.,

$$\begin{aligned}&\overline{\mathcal{L}}(\boldsymbol{W}_{<i}^{k-1}, \boldsymbol{W}_i^{k-1}, \boldsymbol{W}_{>i}^k, \boldsymbol{V}_{<i}^{k-1}, \boldsymbol{V}_i^k, \boldsymbol{V}_{>i}^k, \boldsymbol{U}_{<i}^{k-1}, \boldsymbol{U}_i^k, \boldsymbol{U}_{>i}^k)\\ &\qquad \geq \overline{\mathcal{L}}(\boldsymbol{W}_{<i}^{k-1}, \boldsymbol{W}_i^k, \boldsymbol{W}_{>i}^k, \boldsymbol{V}_{<i}^{k-1}, \boldsymbol{V}_i^k, \boldsymbol{V}_{>i}^k, \boldsymbol{U}_{<i}^{k-1}, \boldsymbol{U}_i^k, \boldsymbol{U}_{>i}^k) + \frac{\alpha}{2}\|\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1}\|_F^2. \quad \text{(C.8)}\end{aligned}$$

Similarly, we can establish the similar descent estimates of (C.8) for the other blocks using the proximal updates including $\boldsymbol{V}_N^k, \{\boldsymbol{U}_i^k\}_{i=1}^{N-1}$ and $\boldsymbol{W}_N^k$ blocks.

Specifically, for the $\boldsymbol{V}_N^k$-block, the following holds

$$\overline{\mathcal{L}}(\{\boldsymbol{W}_i^{k-1}\}_{i=1}^N, \boldsymbol{V}_{i<N}^{k-1}, \boldsymbol{V}_N^{k-1}, \{\boldsymbol{U}_i^{k-1}\}_{i=1}^N) \geq \overline{\mathcal{L}}(\{\boldsymbol{W}_i^{k-1}\}_{i=1}^N, \boldsymbol{V}_{i<N}^{k-1}, \boldsymbol{V}_N^k, \{\boldsymbol{U}_i^{k-1}\}_{i=1}^N) + \frac{\alpha}{2}\|\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}\|_F^2. \quad \text{(C.9)}$$

For the $\{\boldsymbol{U}_i^k\}$-block, $i = 1, \ldots, N-1$, the following holds

$$\begin{aligned}&\overline{\mathcal{L}}(\boldsymbol{W}_{<i}^{k-1}, \boldsymbol{W}_i^{k-1}, \boldsymbol{W}_{>i}^k, \boldsymbol{V}_{<i}^{k-1}, \boldsymbol{V}_i^k, \boldsymbol{V}_{>i}^k, \boldsymbol{U}_{<i}^{k-1}, \boldsymbol{U}_i^{k-1}, \boldsymbol{U}_{>i}^k)\\ &\qquad \geq \overline{\mathcal{L}}(\boldsymbol{W}_{<i}^{k-1}, \boldsymbol{W}_i^{k-1}, \boldsymbol{W}_{>i}^k, \boldsymbol{V}_{<i}^{k-1}, \boldsymbol{V}_i^k, \boldsymbol{V}_{>i}^k, \boldsymbol{U}_{<i}^{k-1}, \boldsymbol{U}_i^k, \boldsymbol{U}_{>i}^k) + \frac{\alpha+\gamma}{2}\|\boldsymbol{U}_i^k - \boldsymbol{U}_i^{k-1}\|_F^2. \quad \text{(C.10)}\end{aligned}$$

For the $\boldsymbol{W}_N^k$-block, the following holds

$$\overline{\mathcal{L}}(\boldsymbol{W}_{i<N}^{k-1}, \boldsymbol{W}_N^{k-1}, \boldsymbol{V}_{i<N}^{k-1}, \boldsymbol{V}_N^k, \boldsymbol{U}_{i<N}^{k-1}, \boldsymbol{U}_N^k) \geq \overline{\mathcal{L}}(\boldsymbol{W}_{i<N}^{k-1}, \boldsymbol{W}_N^k, \boldsymbol{V}_{i<N}^{k-1}, \boldsymbol{V}_N^k, \boldsymbol{U}_{i<N}^{k-1}, \boldsymbol{U}_N^k) + \frac{\alpha}{2}\|\boldsymbol{W}_N^k - \boldsymbol{W}_N^{k-1}\|_F^2. \quad \text{(C.11)}$$

(b) Minimization of a strongly convex case: In this case, we take $\boldsymbol{V}_i^k$-update case for example. From Algorithm 2, $\boldsymbol{V}_i^k$ is updated according to the following

$$\boldsymbol{V}_i^k \leftarrow \underset{\boldsymbol{V}_i}{\operatorname{argmin}} \left\{ s_i(\boldsymbol{V}_i) + \frac{\gamma}{2}\|\boldsymbol{V}_i - \sigma_i(\boldsymbol{U}_i^{k-1})\|_F^2 + \frac{\gamma}{2}\|\boldsymbol{U}_{i+1}^k - \boldsymbol{W}_{i+1}^k \boldsymbol{V}_i\|_F^2 \right\}. \quad \text{(C.12)}$$

Let $h^k(\boldsymbol{V}_i) = s_i(\boldsymbol{V}_i) + \frac{\gamma}{2}\|\boldsymbol{V}_i - \sigma_i(\boldsymbol{U}_i^{k-1})\|_F^2 + \frac{\gamma}{2}\|\boldsymbol{U}_{i+1}^k - \boldsymbol{W}_{i+1}^k \boldsymbol{V}_i\|_F^2$. By the convexity of s_i, the function $h^k(\boldsymbol{V}_i)$ is a strongly convex function with modulus no less than γ. By the optimality of $\boldsymbol{V}_i^k$, the following holds

$$h^k(\boldsymbol{V}_i^{k-1}) \geq h^k(\boldsymbol{V}_i^k) + \frac{\gamma}{2}\|\boldsymbol{V}_i^k - \boldsymbol{V}_i^{k-1}\|_F^2. \quad \text{(C.13)}$$

Noting the relation between $h^k(\boldsymbol{V}_i)$ and $\overline{\mathcal{L}}(\boldsymbol{W}_{<i}^{k-1}, \boldsymbol{W}_i^{k-1}, \boldsymbol{W}_{>i}^k, \boldsymbol{V}_{<i}^{k-1}, \boldsymbol{V}_i, \boldsymbol{V}_{>i}^k, \boldsymbol{U}_{<i}^{k-1}, \boldsymbol{U}_i^{k-1}, \boldsymbol{U}_{>i}^k)$, and by (C.13), it yields for $i = 1, \ldots, N-1$,

$$\begin{aligned}&\overline{\mathcal{L}}(\boldsymbol{W}_{<i}^{k-1}, \boldsymbol{W}_i^{k-1}, \boldsymbol{W}_{>i}^k, \boldsymbol{V}_{<i}^{k-1}, \boldsymbol{V}_i^{k-1}, \boldsymbol{V}_{>i}^k, \boldsymbol{U}_{<i}^{k-1}, \boldsymbol{U}_i^{k-1}, \boldsymbol{U}_{>i}^k)\\ &\qquad \geq \overline{\mathcal{L}}(\boldsymbol{W}_{<i}^{k-1}, \boldsymbol{W}_i^{k-1}, \boldsymbol{W}_{>i}^k, \boldsymbol{V}_{<i}^{k-1}, \boldsymbol{V}_i^k, \boldsymbol{V}_{>i}^k, \boldsymbol{U}_{<i}^{k-1}, \boldsymbol{U}_i^{k-1}, \boldsymbol{U}_{>i}^k) + \frac{\gamma}{2}\|\boldsymbol{V}_i^k - \boldsymbol{V}_i^{k-1}\|_F^2. \quad \text{(C.14)}\end{aligned}$$

Similarly, we can establish the similar descent estimates for the $\boldsymbol{U}_N^k$-block, i.e.,

$$\overline{\mathcal{L}}(\boldsymbol{W}_{<N}^{k-1}, \boldsymbol{W}_N^{k-1}, \boldsymbol{V}_{<N}^{k-1}, \boldsymbol{V}_N^k, \boldsymbol{U}_{<N}^{k-1}, \boldsymbol{U}_N^{k-1}) \geq \overline{\mathcal{L}}(\boldsymbol{W}_{<N}^{k-1}, \boldsymbol{W}_N^{k-1}, \boldsymbol{V}_{<N}^{k-1}, \boldsymbol{V}_N^k, \boldsymbol{U}_{<N}^{k-1}, \boldsymbol{U}_N^k) + \gamma\|\boldsymbol{U}_N^k - \boldsymbol{U}_N^{k-1}\|_F^2. \quad \text{(C.15)}$$

Summing (C.8)–(C.11) and (C.14)–(C.15) yields the descent inequality (C.3).

(c) Prox-linear case for $\boldsymbol{V}_N$, i.e., (3.2): From (3.2), similarly, we let $h^k(\boldsymbol{V}_N) := s_N(\boldsymbol{V}_N) + \mathcal{R}_n(\boldsymbol{V}_N; \boldsymbol{Y}) + \frac{\gamma}{2}\|\boldsymbol{V}_N - \boldsymbol{U}_N^{k-1}\|_F^2$ and $\bar{h}^k(\boldsymbol{V}_N) = s_N(\boldsymbol{V}_N) + \mathcal{R}_n(\boldsymbol{V}_N^{k-1}; \boldsymbol{Y}) + \langle \nabla \mathcal{R}_n(\boldsymbol{V}_N^{k-1}; \boldsymbol{Y}), \boldsymbol{V}_N - \boldsymbol{V}_N^{k-1} \rangle + \frac{\alpha}{2}\|\boldsymbol{V}_N - \boldsymbol{V}_N^{k-1}\|_F^2 + \frac{\gamma}{2}\|\boldsymbol{V}_N - \boldsymbol{U}_N^{k-1}\|_F^2$. By the optimality of $\boldsymbol{V}_N^k$ and the strong convexity of $\bar{h}^k(\boldsymbol{V}_N)$ with modulus at least $\alpha + \gamma$, the following holds

$$\bar{h}^k(\boldsymbol{V}_N^{k-1}) \geq \bar{h}^k(\boldsymbol{V}_N^k) + \frac{\alpha+\gamma}{2}\|\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}\|_F^2.$$

After some simplifications and noting the relation between $h^k(\boldsymbol{V}_N)$ and $\bar{h}^k(\boldsymbol{V}_N)$, we have

$$h^k(\boldsymbol{V}_N^{k-1}) \geq h^k(\boldsymbol{V}_N^k) - \left(\mathcal{R}_n(\boldsymbol{V}_N^k;\boldsymbol{Y}) - \mathcal{R}_n(\boldsymbol{V}_N^{k-1};\boldsymbol{Y}) - \langle \nabla\mathcal{R}_n(\boldsymbol{V}_N^{k-1};\boldsymbol{Y}), \boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}\rangle\right) + \left(\alpha_N + \frac{\gamma}{2}\right)\|\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}\|_F^2 \geq h^k(\boldsymbol{V}_N^k) + \left(\alpha + \frac{\gamma - L_R}{2}\right)\|\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}\|_F^2, \quad \text{(C.16)}$$

where the last inequality holds for the L_R-Lipschitz continuity of $\nabla\mathcal{R}_n$, i.e., the following inequality by Nesterov (2018),

$$\mathcal{R}_n(\boldsymbol{V}_N^k;\boldsymbol{Y}) \leq \mathcal{R}_n(\boldsymbol{V}_N^{k-1};\boldsymbol{Y}) + \langle \nabla\mathcal{R}_n(\boldsymbol{V}_N^{k-1};\boldsymbol{Y}), \boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}\rangle + \frac{L_R}{2}\|\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}\|_F^2.$$

Summing (C.8)–(C.11), (C.14)–(C.15) and (C.9) yields the descent inequality (C.3). □

PROOF (PROOF OF THEOREM 4) By (C.3), $\overline{\mathcal{L}}(\mathcal{P}^k)$ is monotonically nonincreasing and lower bounded by 0 since each term of $\overline{\mathcal{L}}$ is nonnegative, thus, $\overline{\mathcal{L}}(\mathcal{P}^k)$ converges to some nonnegative, finite $\overline{\mathcal{L}}^*$. Similarly, we can show the claims in Theorem 4 holds for Algorithm 1. □

Based on Lemma 7, we can obtain the following corollary.

Corollary 1 (Square summable) *The following hold:*

(a) $\sum_{k=1}^{\infty}\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F^2 < \infty$,

(b) $\frac{1}{K}\sum_{k=1}^{K}\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F^2 \to 0$ *at the rate of* $\mathcal{O}(1/K)$, *and*

(c) $\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F \to 0$ *as* $k \to \infty$,

PROOF Summing (C.3) over k from 1 to ∞ yields

$$\sum_{k=1}^{\infty}\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F^2 \leq \overline{\mathcal{L}}(\mathcal{P}^0) < \infty,$$

which directly implies $\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F \to 0$ as $k \to \infty$. Similarly, summing (C.3) over k from 1 to K yields

$$\frac{1}{K}\sum_{k=1}^{K}\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F^2 \leq \frac{1}{K}\overline{\mathcal{L}}(\mathcal{P}^0),$$

which implies claim (b) of this corollary. □

C.3. Global convergence of BCD

Theorem 4 implies that the quality of the generated sequence is gradually improving during the iterative procedure in the sense of the descent of the objective value, and eventually achieves some level of objective value, then keeps stable. However, the convergence of the generated sequence $\{\mathcal{Q}^k\}_{k\in\mathbb{N}}$ (resp. $\{\mathcal{P}^k\}_{k\in\mathbb{N}}$) itself is still unclear. In the following, we will show that under some natural conditions, the whole sequence converges to some critical point of the objective, and further if the initial point is sufficiently close to some global minimum, then the generated sequence can converge to this global minimum.

Theorem 5 (Global convergence and rate) *Under the assumptions of Theorem 1, the following hold*

(a) $\{\mathcal{Q}^k\}_{k\in\mathbb{N}}$ *(resp.* $\{\mathcal{P}^k\}_{k\in\mathbb{N}}$*) converges to a critical point of* $\mathcal{L}$ *(resp.* $\overline{\mathcal{L}}$*).*

(b) If further the initialization $\mathcal{Q}^0$ *(resp.* $\mathcal{P}^0$*) is sufficiently close to some global minimum* $\mathcal{Q}^*$ *of* $\mathcal{L}$ *(resp.* $\mathcal{P}^*$ *of* $\overline{\mathcal{L}}$*), then* $\mathcal{Q}^k$ *(resp.* $\mathcal{P}^k$*) converges to* $\mathcal{Q}^*$ *(resp.* $\mathcal{P}^*$*).*

(c) Let θ *be the KŁ exponent of* $\mathcal{L}$ *(resp.* $\overline{\mathcal{L}}$*) at* $\mathcal{Q}^*$ *(resp.* $\mathcal{P}^*$*). There hold:* ***(a)*** *if* $\theta = 0$, *then* $\{\mathcal{Q}^k\}_{k\in\mathbb{N}}$ *converges in a finite number of steps;* ***(b)*** *if* $\theta \in \left(0, \frac{1}{2}\right]$, *then* $\|\mathcal{Q}^k - \mathcal{Q}^*\|_F \leq C\tau^k$ *for all* $k \geq k_0$, *for certain* $k_0 > 0, C > 0, \tau \in (0,1)$*; and* ***(c)*** *if* $\theta \in (\frac{1}{2}, 1)$, *then* $\|\mathcal{Q}^k - \mathcal{Q}^*\|_F \leq Ck^{-\frac{1-\theta}{2\theta-1}}$ *for* $k \geq k_0$, *for certain* $k_0 > 0, C > 0$. *The same claims hold for the sequence* $\{\mathcal{P}^k\}$.

(d) $\frac{1}{K}\sum_{k=1}^{K}\|\boldsymbol{g}^k\|_F^2 \to 0$ *at the rate* $\mathcal{O}(1/K)$ *where* $\boldsymbol{g}^k \in \partial\mathcal{L}(\mathcal{Q}^k)$. *Similarly,* $\frac{1}{K}\sum_{k=1}^{K}\|\bar{\boldsymbol{g}}^k\|_F^2 \to 0$ *at the rate* $\mathcal{O}(1/K)$ *where* $\bar{\boldsymbol{g}}^k \in \partial\overline{\mathcal{L}}(\mathcal{Q}^k)$.

Our proof is mainly based on the *Kurdyka-Łojasiewicz* framework established in Attouch et al. (2013) (some other pioneer work can be also found in Attouch et al., 2010). According to Attouch et al. (2013), three key conditions including the *sufficient decrease condition*, the *relative error condition* and the *continuity condition*, together with the KŁ property at some limiting point are required to establish the global convergence of a descent algorithm from the subsequence convergence, where the sufficient decrease condition and the KŁ property restricted to any closed set have been established in Lemma 7 and Proposition 2, respectively. The relative error condition is developed in Lemma 8, while the continuity condition holds naturally due to the continuity assumption. Particularly, the closed set assumption in Proposition 2 can be satisfied naturally by the boundedness of the sequence as well as its limiting points, which is yielded by Lemma 7 and the coerciveness assumption. By exploiting the boundedness, we only need to consider the KŁ property of the objective function restricted to a large bounded, closed set including all limiting points, instead of the total real-valued set. In the following, we first prove Theorem 5 under the subsequence convergence assumption, i.e., condition (a) of this theorem, and then show that both condition (b) and condition (c) can imply the boundedness of the sequence (see Lemma 10), and thus the subsequence convergence as required in condition (a). The rate of convergence results follow the same argument as in the proof of Attouch & Bolte (2009, Theorem 2).

C.3.1. Establishing relative error condition

We restate Lemma 2 precisely as follows.

Lemma 8 (Restatement of Lemma 2) *Under the conditions of Theorem 5, let* $\mathcal{B}$ *be an upper bound of* $\mathcal{P}^{k-1}$ *and* $\mathcal{P}^k$ *for any positive integer* k, $L_{\mathcal{B}}$ *be a uniform Lipschitz constant of* σ_i *on the bounded set* $\{\mathcal{P} : \|\mathcal{P}\|_F \le \mathcal{B}\}$, *and*

$$b := \max\{\gamma, \alpha+\gamma\mathcal{B}, \alpha+\gamma L_{\mathcal{B}}, \gamma\mathcal{B}+2\gamma\mathcal{B}^2, 2\gamma\mathcal{B}+\gamma\mathcal{B}^2\}, \tag{C.17}$$

(or, for the prox-linear case, $b := \max\{\gamma, L_R+\alpha+\gamma\mathcal{B}, \alpha+\gamma L_{\mathcal{B}}, \gamma\mathcal{B}+2\gamma\mathcal{B}^2, 2\gamma\mathcal{B}+\gamma\mathcal{B}^2\}$*), then for any positive integer* k, *there holds,*

$$\operatorname{dist}(\mathbf{0}, \partial\overline{\mathcal{L}}(\mathcal{P}^k)) \le b\sum_{i=1}^{N}[\|\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1}\|_F + \|\boldsymbol{V}_i^k - \boldsymbol{V}_i^{k-1}\|_F + \|\boldsymbol{U}_i^k - \boldsymbol{U}_i^{k-1}\|_F] \le \bar{b}\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F, \tag{C.18}$$

where $\bar{b} := b\sqrt{3N}$, $\operatorname{dist}(\mathbf{0}, \mathcal{S}) := \inf_{\boldsymbol{s}\in\mathcal{S}}\|\boldsymbol{s}\|_F$ *for a set* $\mathcal{S}$, *and*

$$\partial\overline{\mathcal{L}}(\mathcal{P}^k) := (\{\partial_{\boldsymbol{W}_i}\overline{\mathcal{L}}\}_{i=1}^N, \{\partial_{\boldsymbol{V}_i}\overline{\mathcal{L}}\}_{i=1}^N, \{\partial_{\boldsymbol{U}_i}\overline{\mathcal{L}}\}_{i=1}^N)(\mathcal{P}^k).$$

Proof The inequality (C.18) is established via bounding each term of $\partial\overline{\mathcal{L}}(\mathcal{P}^k)$. By the optimality conditions of all updates in Algorithm 2, the following hold

$$\begin{aligned}
&\mathbf{0} \in \partial s_N(\boldsymbol{V}_N^k) + \partial\mathcal{R}_n(\boldsymbol{V}_N^k; \boldsymbol{Y}) + \gamma(\boldsymbol{V}_N^k - \boldsymbol{U}_N^{k-1}) + \alpha(\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}),\\
&\quad(\text{or for prox-linear}, 0 \in \partial s_N(\boldsymbol{V}_N^k) + \nabla\mathcal{R}_n(\boldsymbol{V}_N^{k-1}; \boldsymbol{Y}) + \gamma(\boldsymbol{V}_N^k - \boldsymbol{U}_N^{k-1}) + \alpha(\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}),)\\
&\mathbf{0} = \gamma(\boldsymbol{U}_N^k - \boldsymbol{V}_N^k) + \gamma(\boldsymbol{U}_N^k - \boldsymbol{W}_N^{k-1}\boldsymbol{V}_{N-1}^{k-1}),\\
&\mathbf{0} \in \partial r_N(\boldsymbol{W}_N^k) + \gamma(\boldsymbol{W}_N^k\boldsymbol{V}_{N-1}^{k-1} - \boldsymbol{U}_N^k){\boldsymbol{V}_{N-1}^{k-1}}^\top + \alpha(\boldsymbol{W}_N^k - \boldsymbol{W}_N^{k-1}),
\end{aligned}$$

for $i = N-1, \ldots, 1$,

$$\begin{aligned}
&\mathbf{0} \in \partial s_i(\boldsymbol{V}_i^k) + \gamma(\boldsymbol{V}_i^k - \sigma_i(\boldsymbol{U}_i^{k-1})) + \gamma{\boldsymbol{W}_{i+1}^k}^\top(\boldsymbol{W}_{i+1}^k\boldsymbol{V}_i^k - \boldsymbol{U}_{i+1}^k),\\
&\mathbf{0} \in \gamma[(\sigma_i(\boldsymbol{U}_i^k) - \boldsymbol{V}_i^k) \odot \partial\sigma_i(\boldsymbol{U}_i^k)] + \gamma(\boldsymbol{U}_i^k - \boldsymbol{W}_i^{k-1}\boldsymbol{V}_{i-1}^{k-1}) + \alpha(\boldsymbol{U}_i^k - \boldsymbol{U}_i^{k-1}),\\
&\mathbf{0} \in \partial r_i(\boldsymbol{W}_i^k) + \gamma(\boldsymbol{W}_i^k\boldsymbol{V}_{i-1}^{k-1} - \boldsymbol{U}_i^k){\boldsymbol{V}_{i-1}^{k-1}}^\top + \alpha(\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1}),
\end{aligned}$$

where $\boldsymbol{V}_0^k \equiv \boldsymbol{V}_0 = \boldsymbol{X}$ for all k, and $\odot$ is the Hadamard product. By the above relations, we have

$$
\begin{aligned}
&-\alpha(\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}) - \gamma(\boldsymbol{U}_N^k - \boldsymbol{U}_N^{k-1}) \in \partial s_N(\boldsymbol{V}_N^k) + \partial \mathcal{R}_n(\boldsymbol{V}_N^k; \boldsymbol{Y}) + \gamma(\boldsymbol{V}_N^k - \boldsymbol{U}_N^k) = \partial_{\boldsymbol{V}_N} \overline{\mathcal{L}}(\mathcal{P}^k),\\
&\big(\text{or}, (\nabla \mathcal{R}_n(\boldsymbol{V}_N^k; \boldsymbol{Y}) - \nabla \mathcal{R}_n(\boldsymbol{V}_N^{k-1}; \boldsymbol{Y})) - \alpha(\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}) - \gamma(\boldsymbol{U}_N^k - \boldsymbol{U}_N^{k-1}) \in \partial_{\boldsymbol{V}_N} \overline{\mathcal{L}}(\mathcal{P}^k)\big)\\
&-\gamma(\boldsymbol{W}_N^k - \boldsymbol{W}_N^{k-1})\boldsymbol{V}_{N-1}^k - \gamma \boldsymbol{W}_N^{k-1}(\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1}) = \gamma(\boldsymbol{U}_N^k - \boldsymbol{V}_N^k) + \gamma(\boldsymbol{U}_N^k - \boldsymbol{W}_N^k \boldsymbol{V}_{N-1}^k) = \partial_{\boldsymbol{U}_N} \overline{\mathcal{L}}(\mathcal{P}^k),\\
&\gamma \boldsymbol{W}_N^k \left[\boldsymbol{V}_{N-1}^k (\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1})^\top + (\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1}) \boldsymbol{V}_{N-1}^{k-1\top}\right] - \gamma \boldsymbol{U}_N^k (\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1})^\top - \alpha(\boldsymbol{W}_N^k - \boldsymbol{W}_N^{k-1})\\
&\qquad\qquad \in \partial r_N(\boldsymbol{W}_N^k) + \gamma(\boldsymbol{W}_N^k \boldsymbol{V}_{N-1}^k - \boldsymbol{U}_N^k) \boldsymbol{V}_{N-1}^{k\top} = \partial_{\boldsymbol{W}_N} \overline{\mathcal{L}}(\mathcal{P}^k),
\end{aligned}
$$

for $i = N-1, \ldots, 1$,

$$
\begin{aligned}
&-\gamma(\sigma_i(\boldsymbol{U}_i^k) - \sigma_i(\boldsymbol{U}_i^{k-1})) \in \partial s_i(\boldsymbol{V}_i^k) + \gamma(\boldsymbol{V}_i^k - \sigma_i(\boldsymbol{U}_i^k)) + \gamma \boldsymbol{W}_{i+1}^{k\top}(\boldsymbol{W}_{i+1}^k \boldsymbol{V}_i^k - \boldsymbol{U}_{i+1}^k) = \partial_{\boldsymbol{V}_i} \overline{\mathcal{L}}(\mathcal{P}^k),\\
&-\gamma \boldsymbol{W}_i^{k-1}(\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1}) - \gamma(\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1})\boldsymbol{V}_{i-1}^k - \alpha(\boldsymbol{U}_i^k - \boldsymbol{U}_i^{k-1})\\
&\qquad\qquad \in \gamma[(\sigma_i(\boldsymbol{U}_i^k) - \boldsymbol{V}_i^k) \odot \partial \sigma_i(\boldsymbol{U}_i^k)] + \gamma(\boldsymbol{U}_i^k - \boldsymbol{W}_i^k \boldsymbol{V}_{i-1}^k) = \partial_{\boldsymbol{U}_i} \overline{\mathcal{L}}(\mathcal{P}^k),\\
&\gamma \boldsymbol{W}_i^k \left[\boldsymbol{V}_{i-1}^k (\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1})^\top + (\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1}) \boldsymbol{V}_{i-1}^{k-1\top}\right] - \gamma \boldsymbol{U}_i^k (\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1})^\top - \alpha(\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1})\\
&\qquad\qquad \in \partial r_i(\boldsymbol{W}_i^k) + \gamma(\boldsymbol{W}_i^k \boldsymbol{V}_{i-1}^k - \boldsymbol{U}_i^k) \boldsymbol{V}_{i-1}^{k\top} = \partial_{\boldsymbol{W}_i} \overline{\mathcal{L}}(\mathcal{P}^k).
\end{aligned}
$$

Based on the above relations, and by the Lipschitz continuity of the activation function on the bounded set $\{\mathcal{P} : \|\mathcal{P}\|_F \le \mathcal{B}\}$ and the bounded assumption of both $\mathcal{P}^{k-1}$ and $\mathcal{P}^k$, we have

$$
\begin{aligned}
&\|\mathcal{G}_{\boldsymbol{V}_N}^k\|_F \le \alpha\|\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}\|_F + \gamma\|\boldsymbol{U}_N^k - \boldsymbol{U}_N^{k-1}\|_F, && \mathcal{G}_{\boldsymbol{V}_N}^k \in \partial_{\boldsymbol{V}_N} \overline{\mathcal{L}}(\mathcal{P}^k),\\
&\qquad (\text{or}, \|\mathcal{G}_{\boldsymbol{V}_N}^k\|_F \le (L_R + \alpha)\|\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}\|_F + \gamma\|\boldsymbol{U}_N^k - \boldsymbol{U}_N^{k-1}\|_F)\\
&\|\mathcal{G}_{\boldsymbol{U}_N}^k\|_F \le \gamma\mathcal{B}\|\boldsymbol{W}_N^k - \boldsymbol{W}_N^{k-1}\|_F + \gamma\mathcal{B}\|\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1}\|_F, && \mathcal{G}_{\boldsymbol{U}_N}^k \in \partial_{\boldsymbol{U}_N} \overline{\mathcal{L}}(\mathcal{P}^k),\\
&\|\mathcal{G}_{\boldsymbol{W}_N}^k\|_F \le 2\gamma\mathcal{B}^2\|\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1}\|_F + \gamma\mathcal{B}\|\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}\|_F + \alpha\|\boldsymbol{W}_N^k - \boldsymbol{W}_N^{k-1}\|_F, && \mathcal{G}_{\boldsymbol{W}_N}^k \in \partial_{\boldsymbol{W}_N} \overline{\mathcal{L}}(\mathcal{P}^k),
\end{aligned}
$$

and for $i = N-1, \ldots, 1$,

$$
\begin{aligned}
&\|\mathcal{G}_{\boldsymbol{V}_i}^k\|_F \le \gamma L_{\mathcal{B}}\|\boldsymbol{U}_i^k - \boldsymbol{U}_i^{k-1}\|_F, && \mathcal{G}_{\boldsymbol{V}_i}^k \in \partial_{\boldsymbol{V}_i} \overline{\mathcal{L}}(\mathcal{P}^k),\\
&\|\mathcal{G}_{\boldsymbol{U}_i}^k\|_F \le \gamma\mathcal{B}\|\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1}\|_F + \gamma\mathcal{B}\|\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1}\|_F + \alpha\|\boldsymbol{U}_i^k - \boldsymbol{U}_i^{k-1}\|_F, && \mathcal{G}_{\boldsymbol{U}_i}^k \in \partial_{\boldsymbol{U}_i} \overline{\mathcal{L}}(\mathcal{P}^k),\\
&\|\mathcal{G}_{\boldsymbol{W}_i}^k\|_F \le (\gamma\mathcal{B}^2 + \gamma\mathcal{B})\|\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1}\|_F + \alpha\|\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1}\|_F, && \mathcal{G}_{\boldsymbol{W}_i}^k \in \partial_{\boldsymbol{W}_i} \overline{\mathcal{L}}(\mathcal{P}^k).
\end{aligned}
$$

Summing the above inequalities and after some simplifications, we obtain (C.18). □

C.3.2. Proof of Theorem 5 under condition (a)

Based on Theorem 4 and under the hypothesis that $\overline{\mathcal{L}}$ is continuous on its domain and there exists a convergent subsequence (i.e., condition (a)), the *continuity condition* required in Attouch et al. (2013) holds naturally, i.e., there exists a subsequence $\{\mathcal{P}^{k_j}\}_{j\in\mathbb{N}}$ and $\mathcal{P}^*$ such that

$$
\mathcal{P}^{k_j} \to \mathcal{P}^* \quad \text{and} \quad \overline{\mathcal{L}}(\mathcal{P}^{k_j}) \to \overline{\mathcal{L}}(\mathcal{P}^*), \text{ as } j \to \infty. \tag{C.19}
$$

Based on Lemmas 7 and 8, and (C.19), we can justify the global convergence of $\mathcal{P}^k$ stated in Theorem 5, following the proof idea of Attouch et al. (2013). For the completeness of the proof, we still present the detailed proof as follows.

Before presenting the main proof, we establish a local convergence result of $\mathcal{P}^k$, i.e., the convergence of $\mathcal{P}^k$ when $\mathcal{P}^0$ is sufficiently close to some point $\mathcal{P}^*$. Specifically, let (φ, η, U) be the associated parameters of the KŁ property of $\overline{\mathcal{L}}$ at $\mathcal{P}^*$, where φ is a continuous concave function, η is a positive constant, and U is a neighborhood of $\mathcal{P}^*$. Let ρ be some constant such that $\mathcal{N}(\mathcal{P}^*, \rho) := \{\mathcal{P} : \|\mathcal{P} - \mathcal{P}^*\|_F \le \rho\} \subset U$, $\mathcal{B} := \rho + \|\mathcal{P}^*\|_F$, and $L_{\mathcal{B}}$ be the uniform Lipschitz constant for σ_i, $i = 1, \ldots, N-1$, within $\mathcal{N}(\mathcal{P}^*, \rho)$. Assume that $\mathcal{P}^0$ satisfies the following condition

$$
\frac{\bar{b}}{a}\varphi(\overline{\mathcal{L}}(\mathcal{P}^0) - \overline{\mathcal{L}}(\mathcal{P}^*)) + 3\sqrt{\frac{\overline{\mathcal{L}}(\mathcal{P}^0)}{a}} + \|\mathcal{P}^0 - \mathcal{P}^*\|_F < \rho, \tag{C.20}
$$

where $\bar{b} = b\sqrt{3N}$, b and a are defined in (C.17) and (C.4), respectively,

Lemma 9 (Local convergence) *Under the conditions of Theorem 5, suppose that $\mathcal{P}^0$ satisfies the condition (C.20), and $\overline{\mathcal{L}}(\mathcal{P}^k) > \overline{\mathcal{L}}(\mathcal{P}^*)$ for $k \in \mathbb{N}$, then*

$$\sum_{i=1}^{k} \|\mathcal{P}^i - \mathcal{P}^{i-1}\|_F \le 2\sqrt{\frac{\overline{\mathcal{L}}(\mathcal{P}^0)}{a}} + \frac{\bar{b}}{a}\varphi(\overline{\mathcal{L}}(\mathcal{P}^0) - \overline{\mathcal{L}}(\mathcal{P}^*)),\ \forall k \ge 1, \tag{C.21}$$

$$\mathcal{P}^k \in \mathcal{N}(\mathcal{P}^*, \rho), \quad \forall k \in \mathbb{N}. \tag{C.22}$$

As k goes to infinity, (C.21) yields

$$\sum_{i=1}^{\infty} \|\mathcal{P}^i - \mathcal{P}^{i-1}\|_F < \infty,$$

which implies the convergence of $\{\mathcal{P}^k\}_{k\in\mathbb{N}}$.

PROOF We will prove $\mathcal{P}^k \in \mathcal{N}(\mathcal{P}^*, \rho)$ by induction on k. It is obvious that $\mathcal{P}^0 \in \mathcal{N}(\mathcal{P}^*, \rho)$. Thus, (C.22) holds for $k = 0$. For $k = 1$, we have from (C.3) and the nonnegativeness of $\{\overline{\mathcal{L}}(\mathcal{P}^k)\}_{k\in\mathbb{N}}$ that

$$\overline{\mathcal{L}}(\mathcal{P}^0) \ge \overline{\mathcal{L}}(\mathcal{P}^0) - \overline{\mathcal{L}}(\mathcal{P}^1) \ge a\|\mathcal{P}^0 - \mathcal{P}^1\|_F^2,$$

which implies $\|\mathcal{P}^0 - \mathcal{P}^1\|_F \le \sqrt{\frac{\overline{\mathcal{L}}(\mathcal{P}^0)}{a}}$. Therefore,

$$\|\mathcal{P}^1 - \mathcal{P}^*\|_F \le \|\mathcal{P}^0 - \mathcal{P}^1\|_F + \|\mathcal{P}^0 - \mathcal{P}^*\|_F \le \sqrt{\frac{\overline{\mathcal{L}}(\mathcal{P}^0)}{a}} + \|\mathcal{P}^0 - \mathcal{P}^*\|_F,$$

which indicates $\mathcal{P}^1 \in \mathcal{N}(\mathcal{P}^*, \rho)$.

Suppose that $\mathcal{P}^k \in \mathcal{N}(\mathcal{P}^*, \rho)$ for $0 \le k \le K$. We proceed to show that $\mathcal{P}^{K+1} \in \mathcal{N}(\mathcal{P}^*, \rho)$. Since $\mathcal{P}^k \in \mathcal{N}(\mathcal{P}^*, \rho)$ for $0 \le k \le K$, it implies that $\|\mathcal{P}^k\|_F \le \mathcal{B} := \rho + \mathcal{P}^*$ for $0 \le k \le K$. Thus, by Lemma 8, for $1 \le k \le K$,

$$\operatorname{dist}(\mathbf{0}, \partial\overline{\mathcal{L}}(\mathcal{P}^k)) \le \bar{b}\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F,$$

which together with the KŁ inequality (C.2) yields

$$\frac{1}{\varphi'(\overline{\mathcal{L}}(\mathcal{P}^k) - \overline{\mathcal{L}}(\mathcal{P}^*))} \le \bar{b}\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F. \tag{C.23}$$

By (C.3), the above inequality and the concavity of φ, for $k \ge 2$, the following holds

$$\begin{aligned} a\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F^2 &\le \overline{\mathcal{L}}(\mathcal{P}^{k-1}) - \overline{\mathcal{L}}(\mathcal{P}^k) = (\overline{\mathcal{L}}(\mathcal{P}^{k-1}) - \overline{\mathcal{L}}(\mathcal{P}^*)) - (\overline{\mathcal{L}}(\mathcal{P}^k) - \overline{\mathcal{L}}(\mathcal{P}^*)) \\ &\le \frac{\varphi(\overline{\mathcal{L}}(\mathcal{P}^{k-1}) - \overline{\mathcal{L}}(\mathcal{P}^*)) - \varphi(\overline{\mathcal{L}}(\mathcal{P}^k) - \overline{\mathcal{L}}(\mathcal{P}^*))}{\varphi'(\overline{\mathcal{L}}(\mathcal{P}^{k-1}) - \overline{\mathcal{L}}(\mathcal{P}^*))} \\ &\le \bar{b}\|\mathcal{P}^{k-1} - \mathcal{P}^{k-2}\|_F \cdot \left[\varphi(\overline{\mathcal{L}}(\mathcal{P}^{k-1}) - \overline{\mathcal{L}}(\mathcal{P}^*)) - \varphi(\overline{\mathcal{L}}(\mathcal{P}^k) - \overline{\mathcal{L}}(\mathcal{P}^*))\right], \end{aligned}$$

which implies

$$\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F^2 \le \|\mathcal{P}^{k-1} - \mathcal{P}^{k-2}\|_F \cdot \frac{\bar{b}}{a}\left[\varphi(\overline{\mathcal{L}}(\mathcal{P}^{k-1}) - \overline{\mathcal{L}}(\mathcal{P}^*)) - \varphi(\overline{\mathcal{L}}(\mathcal{P}^k) - \overline{\mathcal{L}}(\mathcal{P}^*))\right].$$

Taking the square root on both sides and using the inequality $2\sqrt{\alpha\beta} \le \alpha + \beta$, the above inequality implies

$$2\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F \le \|\mathcal{P}^{k-1} - \mathcal{P}^{k-2}\|_F + \frac{\bar{b}}{a}\left[\varphi(\overline{\mathcal{L}}(\mathcal{P}^{k-1}) - \overline{\mathcal{L}}(\mathcal{P}^*)) - \varphi(\overline{\mathcal{L}}(\mathcal{P}^k) - \overline{\mathcal{L}}(\mathcal{P}^*))\right].$$

Summing the above inequality over k from 2 to K and adding $\|\mathcal{P}^1 - \mathcal{P}^0\|_F$ to both sides, it yields

$$\|\mathcal{P}^K - \mathcal{P}^{K-1}\|_F + \sum_{k=1}^{K}\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F \le 2\|\mathcal{P}^1 - \mathcal{P}^0\|_F + \frac{\bar{b}}{a}\left[\varphi(\overline{\mathcal{L}}(\mathcal{P}^0) - \overline{\mathcal{L}}(\mathcal{P}^*)) - \varphi(\overline{\mathcal{L}}(\mathcal{P}^K) - \overline{\mathcal{L}}(\mathcal{P}^*))\right]$$

which implies

$$\sum_{k=1}^{K}\|\mathcal{P}^k-\mathcal{P}^{k-1}\|_F\le 2\sqrt{\frac{\overline{\mathcal{L}}(\mathcal{P}^0)}{a}}+\frac{\bar{b}}{a}\varphi(\overline{\mathcal{L}}(\mathcal{P}^0)-\overline{\mathcal{L}}(\mathcal{P}^*)), \tag{C.24}$$

and further,

$$\begin{aligned}\|\mathcal{P}^{K+1}-\mathcal{P}^*\|_F&\le\|\mathcal{P}^{K+1}-\mathcal{P}^K\|_F+\sum_{k=1}^{K}\|\mathcal{P}^k-\mathcal{P}^{k-1}\|_F+\|\mathcal{P}^0-\mathcal{P}^*\|_F\\&\le\sqrt{\frac{\overline{\mathcal{L}}(\mathcal{P}^K)-\overline{\mathcal{L}}(\mathcal{P}^{K+1})}{a}}+2\sqrt{\frac{\overline{\mathcal{L}}(\mathcal{P}^0)}{a}}+\frac{\bar{b}}{a}\varphi(\overline{\mathcal{L}}(\mathcal{P}^0)-\overline{\mathcal{L}}(\mathcal{P}^*))+\|\mathcal{P}^0-\mathcal{P}^*\|_F\\&\le 3\sqrt{\frac{\overline{\mathcal{L}}(\mathcal{P}^0)}{a}}+\frac{\bar{b}}{a}\varphi(\overline{\mathcal{L}}(\mathcal{P}^0)-\overline{\mathcal{L}}(\mathcal{P}^*))+\|\mathcal{P}^0-\mathcal{P}^*\|_F<\rho,\end{aligned}$$

where the second inequality holds for (C.3) and (C.24), the third inequality holds for $\overline{\mathcal{L}}(\mathcal{P}^K)-\overline{\mathcal{L}}(\mathcal{P}^{K+1})\le\overline{\mathcal{L}}(\mathcal{P}^K)\le\overline{\mathcal{L}}(\mathcal{P}^0)$. Thus, $\mathcal{P}^{K+1}\in\mathcal{N}(\mathcal{P}^*,\rho)$. Therefore, we prove this lemma. □

PROOF (PROOF OF THEOREM 5) We prove the whole sequence convergence stated in Theorem 5 according to the following two cases.

Case 1: $\overline{\mathcal{L}}(\mathcal{P}^{k_0})=\overline{\mathcal{L}}(\mathcal{P}^*)$ at some k_0. In this case, by Lemma 7, $\mathcal{P}^k=\mathcal{P}^{k_0}=\mathcal{P}^*$ holds for all $k\ge k_0$, which implies the convergence of $\mathcal{P}^k$ to a limit point $\mathcal{P}^*$.

Case 2: $\overline{\mathcal{L}}(\mathcal{P}^k)>\overline{\mathcal{L}}(\mathcal{P}^*)$ for all $k\in\mathbb{N}$. In this case, since $\mathcal{P}^*$ is a limit point and $\overline{\mathcal{L}}(\mathcal{P}^k)\to\overline{\mathcal{L}}(\mathcal{P}^*)$, by Theorem 4, there must exist an integer k_0 such that $\mathcal{P}^{k_0}$ is sufficiently close to $\mathcal{P}^*$ as required in Lemma 9 (see the inequality (C.20)). Therefore, the whole sequence $\{\mathcal{P}^k\}_{k\in\mathbb{N}}$ converges according to Lemma 9. Since $\mathcal{P}^*$ is a limit point of $\{\mathcal{P}^k\}_{k\in\mathbb{N}}$, we have $\mathcal{P}^k\to\mathcal{P}^*$.

Next, we show $\mathcal{P}^*$ is a critical point of $\overline{\mathcal{L}}$. By Corollary 1(c), $\lim_{k\to\infty}\|\mathcal{P}^k-\mathcal{P}^{k-1}\|_F=0$. Furthermore, by Lemma 8,

$$\lim_{k\to\infty}\operatorname{dist}(\mathbf{0},\partial\overline{\mathcal{L}}(\mathcal{P}^k))=0,$$

which implies that any limit point is a critical point. Therefore, we prove the global convergence of the sequence generated by Algorithm 2.

The convergence to a global minimum is a straightforward variant of Lemma 9.

The $\mathcal{O}(1/k)$ rate of convergence is a direct claim according to the proof of Lemma 8 and Corollary 1(c).

The proof of the convergence of Algorithm 1 is similar to that of Algorithm 2. We give a brief description about this. Note that in Algorithm 1, all blocks of variables are updated via the proximal strategies (or, prox-linear strategy for $\boldsymbol{V}_N$-block). Thus, it is easy to show the similar descent inequality, i.e.,

$$\mathcal{L}(\mathcal{Q}^{k-1})-\mathcal{L}(\mathcal{Q}^k)\ge a\|\mathcal{Q}^k-\mathcal{Q}^{k-1}\|_F^2, \tag{C.25}$$

for some $a>0$. Then similar to the proof of Lemma 8, we can establish the following inequality via checking the optimality conditions of all subproblems in Algorithm 1, i.e.,

$$\operatorname{dist}(\mathbf{0},\partial\mathcal{L}(\mathcal{Q}^k))\le b\|\mathcal{Q}^k-\mathcal{Q}^{k-1}\|_F, \tag{C.26}$$

for some $b>0$. By (C.25), (C.26) and the KŁ property of $\mathcal{L}$ (by Proposition 2), the global convergence of Algorithm 1 can be proved via a similar proof procedure of Algorithm 2. □

C.3.3. CONDITION (B) OR (C) IMPLIES CONDITION (A)

Lemma 10 *Under condition (b) or condition (c) of Theorem 5, $\mathcal{P}^k$ is bounded for any $k\in\mathbb{N}$, and thus, there exists a convergent subsequence.*

Algorithm 3 BCD for DNN Training with ResNets (3.3)

Samples: $\boldsymbol{X} \in \mathbb{R}^{d_0 \times n}$, $\boldsymbol{Y} \in \mathbb{R}^{d_N \times n}$, $\boldsymbol{V}_0^k \equiv \boldsymbol{V}_0 := \boldsymbol{X}$
Initialization: $\{\boldsymbol{W}_i^0, \boldsymbol{V}_i^0, \boldsymbol{U}_i^0\}_{i=1}^N$
Parameters: $\gamma > 0$, $\alpha > 0$
for $k = 1, \ldots$ **do**
  $\boldsymbol{V}_N^k = \operatorname{argmin}_{\boldsymbol{V}_N} \{s_N(\boldsymbol{V}_N) + \mathcal{R}_n(\boldsymbol{V}_N; \boldsymbol{Y}) + \frac{\gamma}{2}\|\boldsymbol{V}_N - \boldsymbol{V}_{N-1}^{k-1} - \boldsymbol{U}_N^{k-1}\|_F^2 + \frac{\alpha}{2}\|\boldsymbol{V}_N - \boldsymbol{V}_N^{k-1}\|_F^2\}$
  $\boldsymbol{U}_N^k = \operatorname{argmin}_{\boldsymbol{U}_N} \{\frac{\gamma}{2}(\|\boldsymbol{V}_N^k - \boldsymbol{V}_{N-1}^{k-1} - \boldsymbol{U}_N\|_F^2 + \|\boldsymbol{U}_N - \boldsymbol{W}_N^{k-1}\boldsymbol{V}_{N-1}^{k-1}\|_F^2)\}$
  $\boldsymbol{W}_N^k = \operatorname{argmin}_{\boldsymbol{W}_N} \{r_N(\boldsymbol{W}_N) + \frac{\gamma}{2}\|\boldsymbol{U}_N^k - \boldsymbol{W}_N\boldsymbol{V}_{N-1}^{k-1}\|_F^2 + \frac{\alpha}{2}\|\boldsymbol{W}_N - \boldsymbol{W}_N^{k-1}\|_F^2\}$
  for $i = N-1, \ldots, 1$ **do**
    $\boldsymbol{V}_i^k = \operatorname{argmin}_{\boldsymbol{V}_i} \{s_i(\boldsymbol{V}_i) + \frac{\gamma}{2}(\|\boldsymbol{V}_i - \boldsymbol{V}_{i-1}^{k-1} - \sigma_i(\boldsymbol{U}_i^{k-1})\|_F^2 + \|\boldsymbol{V}_{i+1}^k - \boldsymbol{V}_i - \sigma_{i+1}(\boldsymbol{U}_{i+1}^k)\|_F^2 + \|\boldsymbol{U}_{i+1}^k - \boldsymbol{W}_{i+1}^k\boldsymbol{V}_i\|_F^2)\}$
    $\boldsymbol{U}_i^k = \operatorname{argmin}_{\boldsymbol{U}_i} \{\frac{\gamma}{2}(\|\boldsymbol{V}_i^k - \boldsymbol{V}_{i-1}^{k-1} - \sigma_i(\boldsymbol{U}_i)\|_F^2 + \|\boldsymbol{U}_i - \boldsymbol{W}_i^{k-1}\boldsymbol{V}_{i-1}^{k-1}\|_F^2) + \frac{\alpha}{2}\|\boldsymbol{U}_i - \boldsymbol{U}_i^{k-1}\|_F^2\}$
    $\boldsymbol{W}_i^k = \operatorname{argmin}_{\boldsymbol{W}_i}\{r_i(\boldsymbol{W}_i) + \frac{\gamma}{2}\|\boldsymbol{U}_i^k - \boldsymbol{W}_i V_{i-1}^{k-1}\|_F^2 + \frac{\alpha}{2}\|\boldsymbol{W}_i - \boldsymbol{W}_i^{k-1}\|_F^2\}$
  end for
end for

PROOF We first show the boundedness of the sequence as well as the subsequence convergence under **condition (b)** of Theorem 5, then under **condition (c)** of Theorem 5.

1. **Condition (b) implies condition (a):** We first establish the boundedness of $\boldsymbol{W}_i^k$, $i = 1, \ldots, N$. Then, recursively, we establish the boundedness of $\boldsymbol{U}_i^k$ via the boudedness of $\boldsymbol{W}_i^k$ and $\boldsymbol{V}_{i-1}^k$ (noting that $\boldsymbol{V}_0^k \equiv \boldsymbol{X}$), followed by that of $\boldsymbol{V}_i^k$ via the boundedness of $\boldsymbol{U}_i^k$, $i = 1, \ldots, N$.

   (1) Boundedness of $\boldsymbol{W}_i^k$ $(i = 1, \ldots, N)$: By Lemma 7, $\overline{\mathcal{L}}(\mathcal{P}^k) < \infty$ for all $k \in \mathbb{N}$. Noting that each term of $\overline{\mathcal{L}}$ is nonnegative, thus, $0 \le r_i(\boldsymbol{W}_i^k) < \infty$ for any $k \in \mathbb{N}$ and $i = 1, \ldots, N$. By the coercivity of r_i, $\boldsymbol{W}_i^k$ is boundedness for any $k \in \mathbb{N}$ and $i = 1, \ldots, N$.
   In the following, we establish the boundedness of $\boldsymbol{U}_i^k$ for any $k \in \mathbb{N}$ and $i = 1, \ldots, N$.

   (2) $i = 1$: Since $\overline{\mathcal{L}}(\mathcal{P}^k) < \infty$, then $\|\boldsymbol{U}_1^k - \boldsymbol{W}_1^k\boldsymbol{X}\|_F^2 < \infty$ for any $k \in \mathbb{N}$. By the boundedness of $\boldsymbol{W}_1^k$ and the coercivity of the function $\|\cdot\|_F^2$, we have the boundedness of $\boldsymbol{U}_1^k$ for any $k \in \mathbb{N}$. Then we show the boundedness of $\boldsymbol{V}_1^k$ by the boundedness of $\boldsymbol{U}_1^k$. Since $\overline{\mathcal{L}}(\mathcal{P}^k) < \infty$, then $\|\boldsymbol{V}_1^k - \sigma_1(\boldsymbol{U}_1^k)\|_F^2 < \infty$ for any $k \in \mathbb{N}$. By the Lipschitz continuity of σ_1 and the boundedness of $\boldsymbol{U}_1^k$, $\sigma_1(\boldsymbol{U}_1^k)$ is uniformly bounded for any $k \in \mathbb{N}$. Thus, by the coercivity of $\|\cdot\|_F^2$, $\boldsymbol{V}_1^k$ is bounded for any $k \in \mathbb{N}$.

   (3) $i > 1$: Recursively, we show that the boundedness of $\boldsymbol{W}_i^k$ and $\boldsymbol{V}_{i-1}^k$ implies the boundedness of $\boldsymbol{U}_i^k$, and then the boundedness of $\boldsymbol{V}_i^k$ from $i = 2$ to N.
   Now, we assume that the boundedness of $\boldsymbol{V}_{i-1}^k$ has been established. Similar to **(2)**, the boundedness of $\boldsymbol{U}_i^k$ is guaranteed by $\|\boldsymbol{U}_i^k - \boldsymbol{W}_i^k\boldsymbol{V}_{i-1}^k\|_F^2 < \infty$ and the boundedness of $\boldsymbol{W}_i^k$ and $\boldsymbol{V}_{i-1}^k$. The boundedness of $\boldsymbol{V}_i^k$ is guaranteed by $\|\boldsymbol{V}_i^k - \sigma_i(\boldsymbol{U}_i^k)\|_F^2 < \infty$ and the boundedness of $\boldsymbol{U}_i^k$, as well as the Lipschitz continuity of σ_i.
   As a consequence, we prove the boundedness of $\{\mathcal{P}^k\}_{k\in\mathbb{N}}$ under condition (b), which implies the subsequence convergent.

2. **Condition (c) implies condition (a):** By Lemma 7 and the finite initialization assumption, we have

$$\overline{\mathcal{L}}(\mathcal{P}^k) \le \overline{\mathcal{L}}(\mathcal{P}^0) < \infty,$$

   which implies the boundedness of $\mathcal{P}^k$ due to the coercivity of $\overline{\mathcal{L}}$ (i.e., condition (c)), and thus, there exists a convergent subsequence.

This completes the proof of this lemma. □

D. Proof of Theorem 3

The proof of Theorem 3 is very similar to those of Lemma 7 and Theorem 5, by noting that the updates are slightly different. In the following, we present the proof of Theorem 3.

Lemma 7 still holds for Algorithm 3 via replacing $\overline{\mathcal{L}}$ with $\overline{\mathcal{L}}_{\text{res}}$, which is stated as the following lemma.

Lemma 11 *Let $\{\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N\}_{k\in\mathbb{N}}$ be a sequence generated by the BCD method (Algorithm 3) for the DNN training with ResNets, then for any $\gamma > 0, \alpha > 0$,*

$$\overline{\mathcal{L}}_{\rm res}\left(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N\right) \le \overline{\mathcal{L}}_{\rm res}\left(\{\boldsymbol{W}_i^{k-1}, \boldsymbol{V}_i^{k-1}, \boldsymbol{U}_i^{k-1}\}_{i=1}^N\right) - a\sum_{i=1}^N \left[\|\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1}\|_F^2 + \|\boldsymbol{V}_i^k - \boldsymbol{V}_i^{k-1}\|_F^2 + \|\boldsymbol{U}_i^k - \boldsymbol{U}_i^{k-1}\|_F^2\right], \quad \text{(D.1)}$$

and $\left\{\overline{\mathcal{L}}_{\rm res}\left(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N\right)\right\}_{k\in\mathbb{N}}$ converges to some $\overline{\mathcal{L}}_{\rm res}^*$, where $a := \min\left\{\frac{\alpha}{2}, \frac{\gamma}{2}\right\}$.

PROOF The proof of this lemma is the same as that of Lemma 7. □

In the ResNets case, Lemma 8 should be revised as the following lemma.

Lemma 12 *Let $\{\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 3. Under Assumptions of Theorem 3, let $\bar{b} := \max\{\alpha + \gamma L, \alpha + \gamma\mathcal{B}, 2\gamma(1 + \mathcal{B} + \mathcal{B}^2), \gamma(1 + L\mathcal{B} + 2\mathcal{B} + 2\mathcal{B}^2)\}$, then*

$$\operatorname{dist}(\mathbf{0}, \partial\overline{\mathcal{L}}_{\rm res}(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N)) \le \bar{b}\sum_{i=1}^N \left[\|\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1}\|_F + \|\boldsymbol{V}_i^k - \boldsymbol{V}_i^{k-1}\|_F + \|\boldsymbol{U}_i^k - \boldsymbol{U}_i^{k-1}\|_F\right], \quad \text{(D.2)}$$

where

$$\partial\overline{\mathcal{L}}_{\rm res}(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N) := (\{\partial_{\boldsymbol{W}_i}\overline{\mathcal{L}}_{\rm res}, \partial_{\boldsymbol{V}_i}\overline{\mathcal{L}}_{\rm res}, \partial_{\boldsymbol{U}_i}\overline{\mathcal{L}}_{\rm res}\}_{i=1}^N)(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N).$$

PROOF From updates of Algorithm 3,

$$\begin{aligned}
&\mathbf{0} \in \partial s_N(\boldsymbol{V}_N^k) + \partial\mathcal{R}_n(\boldsymbol{V}_N^k; \boldsymbol{Y}) + \gamma(\boldsymbol{V}_N^k - \boldsymbol{V}_{N-1}^{k-1} - \boldsymbol{U}_N^{k-1}) + \alpha(\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}),\\
&\mathbf{0} = \gamma(\boldsymbol{U}_N^k + \boldsymbol{V}_{N-1}^{k-1} - \boldsymbol{V}_N^k) + \gamma(\boldsymbol{U}_N^k - \boldsymbol{W}_N^{k-1}\boldsymbol{V}_{N-1}^{k-1}),\\
&\mathbf{0} \in \partial r_N(\boldsymbol{W}_N^k) + \gamma(\boldsymbol{W}_N^k\boldsymbol{V}_{N-1}^{k-1} - \boldsymbol{U}_N^k){\boldsymbol{V}_{N-1}^{k-1}}^\top + \alpha(\boldsymbol{W}_N^k - \boldsymbol{W}_N^{k-1}),\\
&\text{for } i = N-1, \ldots, 1,\\
&\mathbf{0} \in \partial s_i(\boldsymbol{V}_i^k) + \gamma(\boldsymbol{V}_i^k - \boldsymbol{V}_{i-1}^{k-1} - \sigma_i(\boldsymbol{U}_i^{k-1})) - \gamma(\boldsymbol{V}_{i+1}^k - \boldsymbol{V}_i^k - \sigma_{i+1}(\boldsymbol{U}_{i+1}^k)) + \gamma{\boldsymbol{W}_{i+1}^k}^\top(\boldsymbol{W}_{i+1}^k\boldsymbol{V}_i^k - \boldsymbol{U}_{i+1}^k),\\
&\mathbf{0} \in \gamma[(\sigma_i(\boldsymbol{U}_i^k) + \boldsymbol{V}_{i-1}^{k-1} - \boldsymbol{V}_i^k) \odot \partial\sigma_i(\boldsymbol{U}_i^k)] + \gamma(\boldsymbol{U}_i^k - \boldsymbol{W}_i^{k-1}\boldsymbol{V}_{i-1}^{k-1}) + \alpha(\boldsymbol{U}_i^k - \boldsymbol{U}_i^{k-1}),\\
&\mathbf{0} \in \partial r_i(\boldsymbol{W}_i^k) + \gamma(\boldsymbol{W}_i^k\boldsymbol{V}_{i-1}^{k-1} - \boldsymbol{U}_i^k){\boldsymbol{V}_{i-1}^{k-1}}^\top + \alpha(\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1}),
\end{aligned}$$

where $\boldsymbol{V}_0^k \equiv \boldsymbol{V}_0 = \boldsymbol{X}$, for all k, and $\odot$ is the Hadamard product. By the above relations, we have

$$\begin{aligned}
&-\alpha(\boldsymbol{V}_N^k - \boldsymbol{V}_N^{k-1}) - \gamma(\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1}) - \gamma(\boldsymbol{U}_N^k - \boldsymbol{U}_N^{k-1})\\
&\in \partial s_N(\boldsymbol{V}_N^k) + \partial\mathcal{R}_n(\boldsymbol{V}_N^k; \boldsymbol{Y}) + \gamma(\boldsymbol{V}_N^k - \boldsymbol{V}_{N-1}^k - \boldsymbol{U}_N^k) = \partial_{\boldsymbol{V}_N}\overline{\mathcal{L}}_{\rm res}(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N),\\
&-\gamma(\boldsymbol{W}_N^k - \boldsymbol{W}_N^{k-1})\boldsymbol{V}_{N-1}^k - \gamma\boldsymbol{W}_N^{k-1}(\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1}) + \gamma(\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1})\\
&= \gamma(\boldsymbol{U}_N^k + \boldsymbol{V}_{N-1}^k - \boldsymbol{V}_N^k) + \gamma(\boldsymbol{U}_N^k - \boldsymbol{W}_N^k\boldsymbol{V}_{N-1}^k) = \partial_{\boldsymbol{U}_N}\overline{\mathcal{L}}_{\rm res}(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N),\\
&\gamma\boldsymbol{W}_N^k\left[\boldsymbol{V}_{N-1}^k(\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1})^\top + (\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1}){\boldsymbol{V}_{N-1}^{k-1}}^\top\right] - \gamma\boldsymbol{U}_N^k(\boldsymbol{V}_{N-1}^k - \boldsymbol{V}_{N-1}^{k-1})^\top - \alpha(\boldsymbol{W}_N^k - \boldsymbol{W}_N^{k-1})\\
&\qquad\qquad \in \partial r_N(\boldsymbol{W}_N^k) + \gamma(\boldsymbol{W}_N^k\boldsymbol{V}_{N-1}^k - \boldsymbol{U}_N^k){\boldsymbol{V}_{N-1}^k}^\top = \partial_{\boldsymbol{W}_N}\overline{\mathcal{L}}_{\rm res}(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N),
\end{aligned}$$

For $i = N-1, \ldots, 1$, and $\boldsymbol{\xi}_i^k \in \partial\sigma_i(\boldsymbol{U}_i^k)$,

$$\begin{aligned}
&-\gamma(\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1}) - \gamma(\sigma_i(\boldsymbol{U}_i^k) - \sigma_i(\boldsymbol{U}_i^{k-1}))\\
&\qquad \in \partial s_i(\boldsymbol{V}_i^k) + \gamma(\boldsymbol{V}_i^k - \boldsymbol{V}_{i-1}^k - \sigma_i(\boldsymbol{U}_i^k)) - \gamma(\boldsymbol{V}_{i+1}^k - \boldsymbol{V}_i^k - \sigma_{i+1}(\boldsymbol{U}_{i+1}^k)) + \gamma{\boldsymbol{W}_{i+1}^k}^\top(\boldsymbol{W}_{i+1}^k\boldsymbol{V}_i^k - \boldsymbol{U}_{i+1}^k)\\
&\qquad = \partial_{\boldsymbol{V}_i}\overline{\mathcal{L}}_{\rm res}(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N),\\
&-\gamma\boldsymbol{W}_i^{k-1}(\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1}) - \gamma(\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1})\boldsymbol{V}_{i-1}^k - \alpha(\boldsymbol{U}_i^k - \boldsymbol{U}_i^{k-1}) + \gamma(\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1}) \odot \boldsymbol{\xi}_i^k\\
&\qquad\qquad \in \gamma[(\sigma_i(\boldsymbol{U}_i^k) + \boldsymbol{V}_{i-1}^k - \boldsymbol{V}_i^k) \odot \partial\sigma_i(\boldsymbol{U}_i^k)] + \gamma(\boldsymbol{U}_i^k - \boldsymbol{W}_i^k\boldsymbol{V}_{i-1}^k) = \partial_{\boldsymbol{U}_i}\overline{\mathcal{L}}_{\rm res}(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N),
\end{aligned}$$

$$\gamma \boldsymbol{W}_i^k \left[\boldsymbol{V}_{i-1}^k (\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1})^\top + (\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1}) \boldsymbol{V}_{i-1}^{k-1}{}^\top \right] - \gamma \boldsymbol{U}_i^k (\boldsymbol{V}_{i-1}^k - \boldsymbol{V}_{i-1}^{k-1})^\top - \alpha (\boldsymbol{W}_i^k - \boldsymbol{W}_i^{k-1})$$
$$\in \partial r_i(\boldsymbol{W}_i^k) + \gamma (\boldsymbol{W}_i^k \boldsymbol{V}_{i-1}^k - \boldsymbol{U}_i^k) \boldsymbol{V}_{i-1}^k{}^\top$$
$$= \partial_{\boldsymbol{W}_i} \overline{\mathcal{L}}_{\mathrm{res}}(\{\boldsymbol{W}_i^k, \boldsymbol{V}_i^k, \boldsymbol{U}_i^k\}_{i=1}^N).$$

From the above relations, the uniform boundedness of the generated sequence (whose bound is $\mathcal{B}$) and the Lipschitz continuity of the activation function by the hypothesis of this lemma, we have $\|\boldsymbol{\xi}_i^k\| \leq L\mathcal{B}$, and further we get (D.2). □

PROOF (PROOF OF THEOREM 3) The proof of this theorem is very similar to that of Theorem 5. First, similar to Proposition 2, it is easy to show that $\overline{\mathcal{L}}_{\mathrm{res}}$ is also a KŁ function. Then, based on Lemma 11 and Lemma 12, and the KŁ property of $\overline{\mathcal{L}}_{\mathrm{res}}$, we can prove this corollary by Attouch et al. (2013, Theorem 2.9). The other claims of this theorem follow from the same proof of Theorem 5. When the prox-linear strategy is adopted for the $\boldsymbol{V}_N$-update, the claims of Theorem 3 can be proved via following the same proof of Theorem 2. □

E. Closed form solutions of some subproblems

In this section, we provide the closed form solutions to the ReLU involved subproblem and the hinge loss involved subproblem.

E.1. Closed form solution to ReLU-subproblem

From Algorithm 2, when σ_i is ReLU, then the $\boldsymbol{U}_i^k$-update actually reduces to the following one-dimensional minimization problem,

$$u^* = \operatorname*{argmin}_u f(u) := \frac{1}{2}(\sigma(u) - a)^2 + \frac{\gamma}{2}(u - b)^2, \tag{E.1}$$

where $\sigma(u) = \max\{0, u\}$ and $\gamma > 0$. The solution to the above one-dimensional minimization problem can be presented in the following lemma.

Lemma 13 *The optimal solution to Problem* (E.1) *is shown as follows*

$$\mathrm{prox}_{\frac{1}{2\gamma}(\sigma(\cdot)-a)^2}(b) = \begin{cases} \dfrac{a+\gamma b}{1+\gamma}, & \text{if } a + \gamma b \geq 0,\ b \geq 0, \\ \dfrac{a+\gamma b}{1+\gamma}, & \text{if } -(\sqrt{\gamma(\gamma+1)} - \gamma)a \leq \gamma b < 0, \\ b, & \text{if } -a \leq \gamma b \leq -(\sqrt{\gamma(\gamma+1)} - \gamma)a < 0, \\ \min\{b, 0\}, & \text{if } a + \gamma b < 0. \end{cases}$$

PROOF In the following, we divide this into two cases.

(a) $u \geq 0$: In this case,

$$f(u) = \frac{1}{2}(u - a)^2 + \frac{\gamma}{2}(u - b)^2.$$

It is easy to check that

$$u^* = \begin{cases} \dfrac{a+\gamma b}{1+\gamma}, & \text{if } a + \gamma b \geq 0 \\ 0, & \text{if } a + \gamma b < 0 \end{cases}, \tag{E.2}$$

and

$$f\left(\frac{a+\gamma b}{1+\gamma}\right) = \frac{\gamma}{2(1+\gamma)}(b - a)^2, \quad f(0) = \frac{1}{2}a^2 + \frac{\gamma}{2}b^2.$$

(b) $u < 0$: In this case,

$$f(u) = \frac{1}{2}a^2 + \frac{\gamma}{2}(u - b)^2.$$

It is easy to check that

$$u^* = \begin{cases} 0, & \text{if } b \geq 0 \\ b, & \text{if } b < 0 \end{cases}, \tag{E.3}$$

and

$$f(b) = \frac{1}{2}a^2, \quad f(0) = \frac{1}{2}a^2 + \frac{\gamma}{2}b^2.$$

Based on (E.2) and (E.3), we obtain the solution to Problem (E.1) by considering the following four cases.

1. $a + \gamma b \geq 0, b \geq 0$: In this case, we need to compare the values $f\left(\frac{a+\gamma b}{1+\gamma}\right) = \frac{\gamma}{2(1+\gamma)}(b-a)^2$ and $f(0) = \frac{1}{2}a^2 + \frac{\gamma}{2}b^2$. It is obvious that
$$u^* = \frac{a+\gamma b}{1+\gamma}.$$

2. $a + \gamma b \geq 0, b < 0$: In this case, we need to compare the values $f\left(\frac{a+\gamma b}{1+\gamma}\right) = \frac{\gamma}{2(1+\gamma)}(b-a)^2$ and $f(b) = \frac{1}{2}a^2$. By the hypothesis of this case, it is obvious that $a > 0$. We can easily check that
$$u^* = \begin{cases} \dfrac{a+\gamma b}{1+\gamma}, & \text{if } -(\sqrt{\gamma(\gamma+1)} - \gamma)a \leq \gamma b < 0, \\ b, & \text{if } -a \leq \gamma b \leq -(\sqrt{\gamma(\gamma+1)} - \gamma)a < 0. \end{cases}$$

3. $a + \gamma b < 0, b \geq 0$: It is obvious that
$$u^* = 0.$$

4. $a + \gamma b < 0, b < 0$: It is obvious that
$$u^* = b.$$

Thus, the solution to Problem (E.1) is

$$\text{prox}_{\frac{1}{2\gamma}(\sigma(\cdot)-a)^2}(b) = \begin{cases} \dfrac{a+\gamma b}{1+\gamma}, & \text{if } a + \gamma b \geq 0,\ b \geq 0, \\ \dfrac{a+\gamma b}{1+\gamma}, & \text{if } -(\sqrt{\gamma(\gamma+1)} - \gamma)a \leq \gamma b < 0, \\ b, & \text{if } -a \leq \gamma b \leq -(\sqrt{\gamma(\gamma+1)} - \gamma)a < 0, \\ \min\{b, 0\}, & \text{if } a + \gamma b < 0. \end{cases}$$

□

E.2. The closed form of the proximal operator of hinge loss

Consider the following optimization problem

$$u^* = \underset{u}{\text{argmin}}\, g(u) := \max\{0, 1 - a \cdot u\} + \frac{\gamma}{2}(u-b)^2, \tag{E.4}$$

where $\gamma > 0$.

Lemma 14 *The optimal solution to Problem* (E.4) *is shown as follows*

$$\text{hinge}_\gamma(a, b) = \begin{cases} b, & \textit{if } a = 0, \\ b + \gamma^{-1}a, & \textit{if } a \neq 0 \textit{ and } ab \leq 1 - \gamma^{-1}a^2, \\ a^{-1}, & \textit{if } a \neq 0 \textit{ and } 1 - \gamma^{-1}a^2 < ab < 1, \\ b, & \textit{if } a \neq 0 \textit{ and } ab \geq 1. \end{cases}$$

PROOF We consider the problem in the following three different cases: (1) $a > 0$, (2) $a = 0$ and (3) $a < 0$.

(1) $a > 0$: In this case,

$$g(u) = \begin{cases} 1 - au + \dfrac{\gamma}{2}(u-b)^2, & \text{if } u < a^{-1}, \\ \dfrac{\gamma}{2}(u-b)^2, & \text{if } u \geq a^{-1}. \end{cases}$$

It is easy to show that the solution to the problem is

$$u^* = \begin{cases} b + \gamma^{-1}a, & \text{if } a > 0 \text{ and } b \leq a^{-1} - \gamma^{-1}a, \\ a^{-1}, & \text{if } a > 0 \text{ and } a^{-1} - \gamma^{-1}a < b < a^{-1}, \\ b, & \text{if } a > 0 \text{ and } b \geq a^{-1}. \end{cases} \tag{E.5}$$

(2) $a = 0$: It is obvious that

$$u^* = b. \tag{E.6}$$

(3) $a < 0$: Similar to (1),

$$g(u) = \begin{cases} 1 - au + \dfrac{\gamma}{2}(u-b)^2, & u \geq a^{-1}, \\ \dfrac{\gamma}{2}(u-b)^2, & u < a^{-1}. \end{cases}$$

Similarly, it is easy to show that the solution to the problem is

$$u^* = \begin{cases} b + \gamma^{-1}a, & \text{if } a < 0 \text{ and } b \geq a^{-1} - \gamma^{-1}a, \\ a^{-1}, & \text{if } a < 0 \text{ and } a^{-1} < b < a^{-1} - \gamma^{-1}a, \\ b, & \text{if } a < 0 \text{ and } b \leq a^{-1}. \end{cases} \tag{E.7}$$

Thus, we finish the proof of this lemma. □

F. BCD vs. SGD for training ten-hidden-layer MLPs

In this experiment, we attempt to verify the capability of BCD for training MLPs with many layers. Reproducible PyTorch codes can be found at: https://github.com/timlautk/BCD-for-DNNs-PyTorch or https://github.com/yao-lab/BCD-for-DNNs-PyTorch.

Specifically, we consider the DNN training model (2.2) with ReLU activation, the squared loss, and the network architecture being an MLPs with ten hidden layers, on the MNIST data set. The specific settings were summarized as follows:

(a) For the MNIST data set, we implemented a 784-(600×10)-10 MLPs (i.e., the input dimension $d_0 = 28 \times 28 = 784$, the output dimension $d_{11} = 10$, and the numbers of hidden units are all 600), and set $\gamma = \alpha = 1$ for BCD. The sizes of training and test samples are 60000 and 10000, respectively.

(b) The learning rate of SGD is 0.001 (a very conservative learning rate to see if SGD can train the DNNs). More greedy learning rates such as 0.01 and 0.05 have also been used, and similar failure of training is also observed.

(c) For each experiment, we used the same mini-batch sizes (512) and initializations for all algorithms. Specifically, all the weights $\{\boldsymbol{W}_i\}_{i=1}^N$ are initialized from a Gaussian distribution with a standard deviation of 0.01 and the bias vectors are initialized as vectors of all 0.1, while the auxiliary variables $\{\boldsymbol{U}_i\}_{i=1}^N$ and state variables $\{\boldsymbol{V}_i\}_{i=1}^N$ are initialized by a single forward pass.

Under these settings, we plot the curves of training accuracy (acc.) and test accuracy (acc.) of BCD and SGD as shown in Figure 1. According to Figure 1, vanilla SGD usually fails to train such deeper MLPs since it suffers from the vanishing gradient issue (Goodfellow et al., 2016), whereas BCD still works and achieves a moderate accuracy within a few epochs.