Global Convergence in Deep Learning with Variable Splitting via the Kurdyka-Łojasiewicz Property

Jinshan Zeng & Shikang Ouyang  
School of Computer and Information Engineering  
Jiangxi Normal University  
Nanchang, China 330022  
jsh.zeng@gmail.com, oyskang@163.com

Tim Tsz-Kit Lau  
Northwestern University  
timlautk@gmail.com

Shaobo Lin  
Department of Stastics  
Wenzhou University  
Wenzhou, China, 325035  
sblin1983@gmail.com

Yuan Yao  
Department of Mathematics  
The Hong Kong University of Science and Technology  
Clear Water Bay, Kowloon, Hong Kong  
yuany@ust.hk

Abstract

Deep learning has recently attracted a significant amount of attention due to its great empirical success. However, the effectiveness in training deep neural networks (DNNs) remains a mystery in the associated nonconvex optimizations. In this paper, we aim to provide some theoretical understanding on such optimization problems. In particular, the Kurdyka-Łojasiewicz (KL) property is established for DNN training with variable splitting schemes, which leads to the global convergence of block coordinate descent (BCD) type algorithms to a critical point of objective functions under natural conditions of DNNs. Some existing BCD algorithms can be viewed as special cases in this framework.

Keywords: Deep learning, Kurdyka-Łojasiewicz inequality, Block coordinate descent, Global convergence

1 Introduction

Deep learning has recently triggered tremendous research activities due to its great success in some real-world applications such as image classification in computer vision [Krizhevsky et al. (2012)], speech recognition [Hinton et al. (2012); Sainath et al. (2013)], natural language processing, statistical machine translation [Devlin et al. (2014)], and especially outperforming human in Go games [Silver et al. (2016)].

Despite the great success of deep learning, the training of DNNs remains a mystery to understand, since it generally concerns a highly nonconvex optimization problem involving the ill-conditioning of the Hessian, the existence of many local minima, saddle points, plateaus and even some flat regions [Goodfellow et al. (2016)]. The concrete training strategies for deep learning can be mainly divided into three categories in terms of the types of information used, namely, gradient-type methods, approximate second-order methods and gradient-free methods.

The predominant methods in training DNNs are gradient-type methods since they can be easily implemented and are usually very robust. Such methods include the popular back propagation (BP) algorithm [Rumelhart et al. (1986)], as a renaissance of stochastic gradient descent (SGD) method originally proposed by [Robbins and Monro (1951)], enhanced by Polyak momentum [Polyak (1964)].
Nesterov momentum \cite{Sutskever2013}, and SGD with adaptive learning rates and momentum schemes like AdaGrad \cite{Duchi2011}, RMSProp \cite{Tieleman2012}, Adam \cite{Kingma2015} and AMSGrad \cite{Reddi2018}. However, one major flaw of gradient-type methods is the vanishing gradient issue \cite{Rognvaldsson1994, Hochreiter2001}, where gradients in early layers of a multi-layer network decrease exponentially with the number of layers. This causes ill-conditioning of the objective function and leads to tiny steps, slowly zigzagging down a curved valley, and a very slow convergence. Moreover, although there are many existing convergence results of SGD for the traditional optimization problems (say, \cite{Ghadimi2013}), unfortunately, these theoretical results cannot cover the DNN training case, which is highly nonconvex and nonsmooth. Typical (approximate) second-order methods in DNN training include the Newton’s method \cite{LeCun1998}, limited memory BFGS and conjugate gradient \cite{Le2011}. The second-order methods can generally achieve faster convergence rates for some optimization problems with good regularity like invertible Hessian, but often require much more memory storage than gradient-type methods, which limits their applications to DNN training with large-scale models. In order to alleviate the vanishing gradient issue, recently some gradient-free methods were studied in DNN training, including (but not limited to) the alternating direction method of multiplier (ADMM) \cite{Taylor2016, Zhang2016, Zhang2017, Lau2018}. The main idea of ADMM and BCD is to decompose the highly coupled and composite DNN training problem into several loosely coupled and nearly separable simple subproblems. The efficiency of both ADMM and BCD has been illustrated empirically in \cite{Taylor2016, Zhang2016, Zhang2017}. However, the existing convergence results for nonconvex optimization problems such as the convergence of ADMM type methods \cite{Hong2016, Wang2018}, and the convergence of BCD type methods \cite{Xu2013, Xu2017}, cannot be directly applied to the general DNN training problem for its nonconvex and nonsmooth nature. Therefore, establishing the convergence results of these DNN training algorithms is in demand and full of challenge.

One major tool for establishing the global convergence of a nonconvex algorithm is the powerful Kurdyka–Łojasiewicz (KL) framework, which was summarized in \cite{Attouch2013} for a class of descent algorithms (some other pioneer work on this can be also found in \cite{Attouch2009, Attouch2010}). The KL property of Lyapunov functions, along which the algorithm is descending, plays a central role in global convergence, i.e., from any initial choice the whole output sequences converge to a critical point of the Lyapunov functions. Despite the popularity of deep learning, very few deep learning algorithms have been proved to have global convergence. Therefore, to establish the global convergence of nonconvex algorithms in deep learning within this framework, the KL property of the DNN training problem is the crucial step. However, since the objective function of the DNN training problem is highly nonconvex, where the variables are highly coupled in a nested network architecture, the KL property in deep learning remains unexplored.

In this paper, we aim to fill in this gap. Our main contributions are as follows.

(a) To overcome the hurdle of variable coupling in original DNN training, we exploit a simple alternative formulation of DNN training using variable splitting, largely decoupling the variables. The new formulation enables us to establish the KL property under very general assumptions, and to adapt the BCD methods efficiently solving the suggested splitting formulation with global convergence guarantee. Existing BCD methods above for DNN training can be viewed as special cases in this framework. To the best of our knowledge, it is the first time on establishing the KL properties in deep learning.

(b) Our global convergence relies on natural and weak conditions in deep learning by relaxing the Lipschitz differentiability and block multiconvexity assumptions in \cite{Xu2013, Xu2017} and the block-wise Lipschitz differentiability assumption in \cite{Bolte2014} that are not satisfied in DNN training. Our analysis exploits the splitting structure of DNN training and the Lipschitz continuity of the activation function. Specifically, using the proximal update for each split variable block, we establish the value convergence of the generated sequence. Then, based on the Lipschitz continuity of the activation function and leveraging the established KL property, one can boost the subsequence convergence to the global convergence of the whole sequence to a critical point of Lyapunov functions. Furthermore, if the initialization is sufficiently close to some global minimum, then the developed BCDs converge to this global minimum. Moreover, some rate of convergence results can be established if the KL exponent is investigated explicitly.
(c) Some preliminary experiments are conducted with MNIST and CIFAR-10 datasets, showing that such BCD algorithms may find network parameters of approximately zero training loss (error) with over-parameterized models, and can achieve moderate accuracies at the very early stage.

The rest of this paper is organized as follows. Section 2 establishes the KL properties of deep learning models with variable splitting. Section 3 adapts the BCD methods to the splitting formulations and then establishes their global convergence under natural assumptions. Section 4 provides several numerical experiments to verify the developed theoretical results.

2 The Kurdyka-Łojasiewicz Property in Deep Learning

2.1 DNN Training with Variable Splitting

Consider $N$-layer feedforward neural networks including $N-1$ hidden layers of the neural networks. Particularly, let $d_i \in \mathbb{N}$ be the number of hidden neurons at the $i$-th hidden layer for $i = 1, \ldots, N-1$. Let $d_0$ and $d_N$ be the number of input and output layers, respectively. Let $W_i \in \mathbb{R}^{d_{i-} \times d_i}$ be the weight between the $(i-1)$-th layer and the $i$-th layer for any $i = 1, \ldots, N$.

Let $Z := \{(x_j, y_j)\}^n_{j=1} \subset \mathbb{R}^{d_{0} \times d_N}$ be $n$ samples. Denote $X := (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{d_0 \times n}$ and $Y := (y_1, y_2, \ldots, y_n) \in \mathbb{R}^{d_N \times n}$. With the help of these notations, the deep neural network (DNN) training problem can be formulated as the following empirical risk minimization:

$$\min_{\{W_i\}_{i=1}^{N}} \mathcal{R}_n(\Phi(X; \{W_i\}_{i=1}^{N}), Y) := \frac{1}{n} \sum_{j=1}^{n} \ell(\Phi(x_j; \{W_i\}_{i=1}^{N}), y_j),$$

where $\ell : \mathbb{R}^{d_{0} \times d_N} \rightarrow \mathbb{R}_+ \cup \{0\}$ is some nonnegative, continuous loss function (often convex but possibly nonconvex), $\Phi(x_j; \{W_i\}_{i=1}^{N})$ is the neural network model with $N$ layers and weights $\{W_i\}_{i=1}^{N}$, and $\mathcal{R}_n$ is called empirical risk (also known as the training loss).

Note that the DNN training model (1) is highly nonconvex as the variables are coupled via the deep neural network architecture, which brings many challenges for the design of efficient training algorithms and also its theoretical analysis. Particularly, the Kurdyka-Łojasiewicz property of Problem (1) is still unclear, and there is also a lack of provable convergent DNN training algorithms.

In order to make Problem (1) more computationally tractable, variable splitting is one of the most commonly used ways. The main idea of variable splitting is to transform a complicated problem (where the variables are coupled highly nonlinearly) into a relatively simpler one (where the variables are coupled much looser) via introducing some additional variables. Considering general deep neural network architectures, the DNN training problem can be naturally formulated as the following model (called 2-splitting formulation henceforth):

$$\min_{\{W_i, V_i\}_{i=1}^{N}} \mathcal{L}_0(\{W_i\}_{i=1}^{N}; \{V_i\}_{i=1}^{N}) := \mathcal{R}_n(V_N; Y) + \sum_{i=1}^{N} r_i(W_i) + \sum_{i=1}^{N} s_i(V_i)$$

subject to $V_i = \sigma_i(W_iV_{i-1})$, $i = 1, \ldots, N$,

where $\mathcal{R}_n(V_N; Y) := \frac{1}{n} \sum_{j=1}^{n} \ell((V_N)_j, y_j)$ denotes the empirical risk, $(V_N)_j$ is the $j$-th column of $V_N$, $r_i$ and $s_i$ ($i = 1, \ldots, N$) are extended-real-valued, nonnegative convex functions revealing the priors of the weight variable $W_i$ and the state variable $V_i$ (or the constraints on $W_i$ and $V_i$), $\sigma_i$ is the activation function of the $i$-th layer (generally, $\sigma_N \equiv 1d$, i.e., the identity function), and let $V_0 := X$. The 2-splitting formulation of DNN training problem has been widely used in deep learning literature (say, Carreira-Perpinan and Wang (2014)). In order to solve the 2-splitting formulation (2), the following alternative minimization problem was suggested in Carreira-Perpinan and Wang (2014):

$$\min_{\{W_i, V_i\}_{i=1}^{N}} \mathcal{L}(\{W_i\}_{i=1}^{N}; \{V_i\}_{i=1}^{N}) := \mathcal{L}_0(\{W_i\}_{i=1}^{N}; \{V_i\}_{i=1}^{N}) + \frac{\gamma}{2} \sum_{i=1}^{N} \|V_i - \sigma_i(W_iV_{i-1})\|_F^2,$$

\footnote{For the simplicity of notations, we regard the input and output layers as the 0-th and $N$-th layers, respectively, and we absorb the bias of each layer into $W_i$.}

\footnote{Here, we consider the more general regularized DNN training model. If there is no regularization, then the model reduces to the original DNN training model (1).}
where $\gamma > 0$ is a hyperparameter.

The DNN training model (3) can be very general, where: (a) $\ell$ can be the squared, logistic, hinge, cross-entropy or other commonly used loss functions; (b) $\sigma$ can be ReLU, sigmoid, linear link, or other commonly used activation functions; (c) $r_i$ can be the squared $\ell_2$ norm, $\ell_1$ norm, elastic net [Zou and Hastie (2005), the indicator function of some nonempty closed convex set $\mathbb{C}$ (such as the nonnegative closed half space or an interval set $[0, 1]$), or others; (d) $s_i$ can be the indicator function of some convex set with simple projection [Zhang and Brand (2017), or others]. Particularly, if there is no regularizer or constraint on $W_i$ (or $V_i$), then $r_i$ (or $s_i$) can be zero.

Note that the variables $W_i$ and $V_{i-1}$ are coupled by the nonlinear activation function in the $i$-th constraint of the 2-splitting formulation (2), which may bring some difficulties and challenges for solving problem (2) efficiently, particularly, when the activation function is ReLU. Instead, we follow the variable splitting idea from [Taylor et al., 2016], which reformulates the 2-splitting formulation (2) into the following equivalent form called 3-splitting formulation:

$$
\min_{\{W_i, V_i, U_i\}_{i=1}^N} \mathcal{L}_0(\{W_i\}_{i=1}^N, \{V_i\}_{i=1}^N)
$$

subject to $U_i = W_i V_{i-1}$, $V_i = \sigma_i(U_i)$, $i = 1, \ldots, N$.

From (4), the variables are coupled much more loosely, particularly for variables $W_i$ and $V_{i-1}$. As described latter, such 3-splitting formulation can bring many benefits to designing some more efficient methods, though $N$ auxiliary variables $U_i$’s are introduced. Similarly, the following alternative unconstrained problem can be suggested as an alternative of (4), that is,

$$
\min_{\{W_i, V_i, U_i\}_{i=1}^N} \tilde{\mathcal{L}}(\{W_i\}_{i=1}^N, \{V_i\}_{i=1}^N, \{U_i\}_{i=1}^N)
$$

$$:= \mathcal{L}_0(\{W_i\}_{i=1}^N, \{V_i\}_{i=1}^N) + \gamma \sum_{i=1}^N \left(\|V_i - \sigma_i(U_i)\|_F^2 + \|U_i - W_i V_{i-1}\|_F^2\right).$$

Before presenting the definition of Kurdyka-Łojasiewicz (KL) property, we first introduce some notions and notations from variational analysis, which can be found in Rockafellar and Wets (1998).

Let $h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be a real-extended-valued function (respectively, $h : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ be a point-to-set mapping), its graph is defined by

Graph$(h) := \{(x, y) \in \mathbb{R}^p \times \mathbb{R} : y = h(x)\}$

and its domain by dom$(h) := \{x \in \mathbb{R}^p : h(x) < +\infty\}$ (resp. dom$(h) := \{x \in \mathbb{R}^p : h(x) \neq \emptyset\}$).

When $h$ is a proper function, i.e., when dom$(h) \neq \emptyset$, the set of its global minimizers (possibly empty) is denoted by

$$\arg\min h := \{x \in \mathbb{R}^p : h(x) = \inf h\}.$$

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

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

When $x \notin \text{dom}(h)$, we set $\partial_f h(x) = \emptyset$. The limiting-subdifferential (or simply subdifferential, henceforth) of $h$ introduced in [Mordukhovich (2006)], written $\partial h(x)$ at $x \in \text{dom}(f)$, is defined by

$$\partial h(x) := \{v \in \mathbb{R}^p : \exists x^k \to x, h(x^k) \to h(x), v^k \in \partial_f h(x^k) \to v\}.$$ (7)

A necessary (but not sufficient) condition for $x \in \mathbb{R}^p$ to be a minimizer of $h$ is $0 \in \partial h(x)$. A point that satisfies this inclusion is called limiting-critical or simply critical. The distance between a point $x$ to a subset $S$ of $\mathbb{R}^p$, written $\text{dist}(x, S)$, is defined by

$$\text{dist}(x, S) = \inf\{\|x-s\| : s \in S\},$$

where $\| \cdot \|$ represents the Euclidean norm.

The Kurdyka-Łojasiewicz (KL) property [Łojasiewicz (1963, 1993); Kurdyka (1998); Bolte et al. (2007a,b)] plays a central role in the convergence analysis of nonconvex algorithms (see, Attouch et al. (2013); Xu and Yin (2013); Wang et al. (2018) for instance). The following definition is taken from [Bolte et al. (2007a)].

In (6), we use a uniform hyperparameter $\gamma$ for the sum of all quadratic terms for the simplicity of notation. In practice, $\gamma$ can be different for each quadratic term and our proof still goes through.

The indicator function $\iota_C$ of a nonempty convex set $C$ is defined as $\iota_C(x) = 0$ if $x \in C$ and $\infty$ otherwise.
According to Łojasiewicz (1965); Bochnak et al. (1998) and (Shiota, 1997, I.2.9, p.52), the class of proper lower semi-continuous functions which satisfy the Kurdyka-Łojasiewicz inequality at each point of \( \text{dom}(\partial h) \) are called KL functions.

Proper lower semi-continuous functions which satisfy the Kurdyka-Łojasiewicz inequality at each point of \( \text{dom}(\partial h) \) are called KL functions.

Note that we have adopted in the definition of KL inequality \( (8) \) 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 and Parks (2002) for \( \theta \in [\frac{1}{2}, 1) \), then was extended to functions defined on the \( o \)-minimal structure in Kurdyka (1998), and latter was extended to nonsmooth subanalytic functions in Bolte et al. (2007a).

From the definition of KL 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 \( h(x^*) = 0 \), inequality \( (8) \) can be rewritten

\[
\| \nabla (\varphi \circ h)(x) \| \geq 1,
\]

for each convenient \( x \in \mathbb{R}^p \). This inequality may be interpreted as follows: up to the reparametrization of the values of \( h \) via \( \varphi \), we face a sharp function. Since the function \( \varphi \) 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). KL functions include real analytic functions (see, the latter Definition 5), semi-algebraic functions (see, the latter Definition 3), tame functions defined in some \( o \)-minimal structures Kurdyka (1998), continuous subanalytic functions Bolte et al. (2007a) and locally strongly convex functions Xu and Yin (2013).

In the following, we give the definitions of real-analytic and semi-algebraic functions.

**Definition 1 (KL property).** A function \( h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\} \) is said to have the Kurdyka-Łojasiewicz property at \( x^* \in \text{dom}(\partial h) \) if there exist a neighborhood \( U \) of \( x^* \), a constant \( \eta \), 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

\[
\varphi'(h(x) - h(x^*)) \text{dist}(0, \partial h(x)) \geq 1, \quad \forall x \in U \cap \text{dom}(\partial h) \text{ and } h(x^*) < h(x) < h(x^*) + \eta, \quad (8)
\]

where \( \theta \) is called the KL exponent of \( h \) at \( x^* \).

Proper lower semi-continuous functions which satisfy the Kurdyka-Łojasiewicz inequality at each point of \( \text{dom}(\partial h) \) are called KL functions.

According to Krantz and Parks (2002), some typical real analytic functions include polynomials, exponential functions, and the logarithm, trigonometric and power functions on any open set of their domains. One can verify whether a multivariable real function \( h(x) \) on \( \mathbb{R}^p \) is analytic by checking the analyticity of \( g(t) := h(x + ty) \) for any \( x, y \in \mathbb{R}^p \).

**Definition 2 (Real analytic, Definition 1.1.5 in (Krantz and Parks, 2002)).** A function \( h \) with domain an open set \( U \subset \mathbb{R} \) and range either the real or the complex numbers, is said to be real analytic at \( u \) if the function \( f \) may be represented by a convergent power series on some interval of positive radius centered at \( u \):

\[
h(x) = \sum_{j=0}^{\infty} \alpha_j (x - u)^j,
\]

for some \( \{ \alpha_j \} \subset \mathbb{R} \). The function is said to be real analytic on \( V \subset U \) if it is real analytic at each \( u \in V \). The real analytic function \( f \) over \( \mathbb{R}^p \) for some positive integer \( p > 1 \) can be defined similarly.

According to Krantz and Parks (2002), some typical real analytic functions include polynomials, exponential functions, and the logarithm, trigonometric and power functions on any open set of their domains. One can verify whether a multivariable real function \( h(x) \) on \( \mathbb{R}^p \) is analytic by checking the analyticity of \( g(t) := h(x + ty) \) for any \( x, y \in \mathbb{R}^p \).

**Definition 3 (Semialgebraic).**

(a) A set \( D \subset \mathbb{R}^p \) is called semialgebraic [Bochnak et al. (1998)] if it can be represented as

\[
D = \bigcup_{i=1}^{s} \bigcap_{j=1}^{t} \{ x \in \mathbb{R}^p : P_{ij}(x) = 0, Q_{ij}(x) > 0 \},
\]

where \( P_{ij}, Q_{ij} \) are real polynomial functions for \( 1 \leq i \leq s, 1 \leq j \leq t \).

(b) A function \( h : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\} \) (resp. a point-to-set mapping \( h : \mathbb{R}^p \rightharpoonup \mathbb{R}^q \)) is called semialgebraic if its graph \( \text{Graph}(h) \) is semialgebraic.

According to Łojasiewicz (1965), Bochnak et al. (1998) and Shiota (1997) I.2.9, p.52), the class of semialgebraic sets is stable under the operation of finite union, finite intersection, Cartesian product or complementation. Some typical examples include polynomial functions, the indicator function of a semialgebraic set, and the Euclidean norm (p. 26 in Bochnak et al. (1998)).
In the following, we establish the KL properties\(^3\) of the DNN training models with variable splitting, i.e., the functions \(\mathcal{L}\) defined in (3) and \(\hat{\mathcal{L}}\) defined in (6).

**Theorem 1** (KL properties of deep learning). Suppose that the loss function \(\ell\), the activation functions \(\sigma_i (i = 1, \ldots, N-1)\), and the regularizers \(r_i\) and \(s_i (i = 1, \ldots, N)\) are either real analytic or semialgebraic functions, and are continuous on their domains, then functions \(\mathcal{L}\) defined in (3) and \(\hat{\mathcal{L}}\) defined in (6) are KL functions.

It includes: (a) the loss function \(\ell\) is the squared, logistic, hinge, exponential or cross-entropy losses, (b) the activation function \(\sigma_i (i = 1, \ldots, N-1)\) is ReLU, leaky ReLU, sigmoid, tanh, linear, or polynomial function, and (c) the regularizers \(r_i\) and \(s_i (i = 1, \ldots, N)\) are the squared \(\ell_2\) norm, squared Frobenius norm, the entrywise 1-norm, or the sum of squared Frobenius norm and entrywise 1-norm (say, elastic net in the vector case), or the indicator function of nonnegative closed half space or a closed interval.

The proof of this theorem is presented in Appendix A. This theorem shows that most of the DNN training models with variable splitting have some “nice” geometrical properties, that is, they are amenable to sharpness at each point in their domains. However, the similar geometry property of the original DNN training model (1) is still a mystery.

3 BCD for DNN Training and Global Convergence

3.1 Development of BCD Algorithms

In the following, we describe how to apply the BCD method for Problems (3) and (5). The main idea of the BCD method of Gauss-Seidel type for a minimization problem with multi-block variables is to update all the variables cyclically while fixing the remaining blocks at their last updated values. In this paper, we consider the BCD method with the backward order (but not limited to this as discussed later) for the updates of variables, that is, we update the variables from the output layer to the input layer, and for each layer, we update the variables \(\{V_i, W_i\}\) cyclically for Problem (5) as well as the variables \(\{V_i, U_i, W_i\}\) cyclically for Problem (3). Since \(\sigma_N \equiv 1d\), the output layer is paid special attention. Particularly, for most blocks, we use the proximal update strategies. For each subproblem, we assume that its minimizer can be achieved. The BCD algorithms for Problems (3) and (5) can be summarized in Algorithm 1 and Algorithm 2 respectively.

**Algorithm 1** BCD for DNN Training with 2-splitting (3)

Samples: \(X := [x_1, \ldots, x_n] \in \mathbb{R}^{d u \times n}, Y := [y_1, \ldots, y_n] \in \mathbb{R}^{d v \times n}, V_0^k \equiv V_0 := X\)

Initialization: \(\{W_0^k, V_0^k\}_{k=1}^N\)

Parameters: \(\gamma > 0, \alpha > 0\)

for \(k = 1, \ldots, \) do

\[ V_N^k \leftarrow \arg\min_{V_N} \{s_N(V_N) + \mathcal{R}_N(V_N; Y) + \frac{\gamma}{2} \|V_N - W_N^{k-1} V_N^{k-1}\|_F^2 + \frac{\alpha}{2} \|V_N - V_N^{k-1}\|_F^2\} \]

\[ W_N^k \leftarrow \arg\min_{W_N} \{r_N(W_N) + \frac{\gamma}{2} \|W_N - V_N^{k-1}\|_F^2 + \frac{\alpha}{2} \|W_N - W_N^{k-1}\|_F^2\} \]

for \(i = N - 1, \ldots, 1\) do

\[ V_i^k \leftarrow \arg\min_{V_i} \{s_i(V_i) + \frac{\gamma}{2} \|V_i - \sigma_i(W_i^{k-1} V_i^{k-1})\|_F^2 + \frac{\alpha}{2} \|V_i - V_i^{k-1}\|_F^2\} \]

\[ W_i^k \leftarrow \arg\min_{W_i} \{r_i(W_i) + \frac{\gamma}{2} \|W_i - \sigma_i(W_i V_i^{k-1})\|_F^2 + \frac{\alpha}{2} \|W_i - W_i^{k-1}\|_F^2\} \]

end for

end for

The major advantage of Algorithm 2 over Algorithm 1 lies in that in each subproblem, almost all updates are simple proximal updates\(^4\) (or just least squares problems), which usually have closed form solutions for many commonly used DNNs. Some typical examples include: (a) \(r_i, s_i\) are 0 (i.e., no regularization), or the squared \(\ell_2\) norm (a.k.a. weight decay), or the indicator function of\(^5\)

---

\(^3\)It should be pointed out that we need to use the vectorization of the matrix variables involved in \(\mathcal{L}\) and \(\hat{\mathcal{L}}\) in order to adopt the existing definitions of KL property, real analytic functions and semialgebraic functions. We still use the matrix notation for the simplicity of notation.

\(^4\)In practice, for different update blocks, we can use different \(\alpha\) and our proof still goes through.

\(^5\)For \(V_N^k\)-update, we can regard \(s_N(V_N) + \mathcal{R}_N(V_N; y)\) as a new proximal function \(\tilde{s}_N(V_N)\).
Algorithm 2 BCD for DNN Training with 3-splitting (5)

Samples: \( X := [x_1, \ldots, x_n] \in \mathbb{R}^{d_0 \times n} \), \( Y := [y_1, \ldots, y_n] \in \mathbb{R}^{d_1 \times n} \), \( V_k^0 \equiv V_0 := X \)

Initialization: \( \{ W_0, V_0, U_0 \} \), \( i = 1 \)

Parameters: \( \gamma > 0 \), \( \alpha > 0 \)

for \( k = 1 \ldots, \) do

\( V_k^N \leftarrow \arg \min_{V_k} \| s_N(V) + \mathcal{R}_N(V); Y \| + \frac{\gamma}{2}\| V_N - U_k^{-1} \|_F^2 + \frac{\gamma}{2}\| V_N - V_{k-1} \|_F^2 \}

\( U_k^N \leftarrow \arg \min_{U_k} \frac{1}{2}\| U_k - V_N \|_F^2 + \frac{\gamma}{2}\| U_k - W_k^{-1} \|_F^2 \}

\( W_k^N \leftarrow \arg \min_{W_k} \{ r_N(W) + \frac{\gamma}{2}\| U_k - W_k^{-1} \|_F^2 \} \)

for \( i = N - 1 \ldots, 1 \) do

\( V_i^k \leftarrow \arg \min_{V_i} \| s_i(V) + \mathcal{R}_i(V); Y \| + \frac{\gamma}{2}\| V_i - U_i^k \|_F^2 + \frac{\gamma}{2}\| U_i^k - W_i^k \|_F^2 \}

\( U_i^k \leftarrow \arg \min_{U_i} \| V_i - U_i \|_F^2 + \frac{\gamma}{2}\| U_i - W_i^k \|_F^2 + \frac{\gamma}{2}\| U_i - U_i^{-1} \|_F^2 \}

\( W_i^k \leftarrow \arg \min_{W_i} \{ r_i(W) + \frac{\gamma}{2}\| U_i - W_i \|_F^2 \} \)

end for

end for

a nonempty closed convex set with a simple projection like the nonnegative closed half space and the closed interval \([0, 1] \); (b) the loss function \( \ell \) is the squared loss or hinge loss (see, Lemma 10 in Appendix [E]); and (c) \( \sigma_i \) is ReLU (see, Lemma 9 in Appendix [D]) or linear link function. For other cases in which \( r_i \) and \( s_i \) are the \( \ell_1 \) norm, \( \sigma_i \) is the sigmoid function, and the loss \( \ell \) is the logistic function, the associated subproblems can be also solved cheaply via some efficient existing methods.

In the following, we give some remarks on the presented BCD algorithms.

On comparisons of BCD algorithms for DNN training: Two existing BCD algorithms for DNN training [Carreira-Perpinan and Wang (2014); Zhang and Brand (2017)] can be viewed as special cases in 2-splitting formulation (2). In fact, Carreira-Perpinan and Wang (2014) consider a specific DNN training model with squared loss and sigmoid activation function, and proposed the method of auxiliary coordinate (MAC) based on the 2-splitting formulation of DNN training (2), as a two-block BCD method with the weight variables \( W := \{ W_i \}_{i=1}^N \) as one block and the state variables \( Y := \{ V_i \}_{i=1}^N \) as the other block. For each block, a nonlinear least squares problem is solved by some iterative methods. Furthermore, Zhang and Brand (2017) proposed a BCD type method for DNN training with ReLU and squared loss. To avoid the computational hurdle imposed by ReLU, the DNN training model was relaxed to a smooth multi-convex formulation via lifting ReLU into a higher dimensional space Zhang and Brand (2017). Such a relaxed BCD is in fact a special case of 2-splitting formulation (2) with \( \sigma_i \equiv \text{Id}, r_i \equiv 0, s_i(V) = i\chi(V), i = 1, \ldots, N \), where \( \chi \) is the nonnegative closed half-space with the same dimension of \( V_i \). In a contrast, Algorithm 2 exploits a 3-splitting formulation, when applied to such DNN training models, all the updates have closed-form solutions except for \( \{ U_i \}_{i=1}^{N-1} \) updates, which essentially reduce to a one-dimensional minimization problem, i.e., \( \min_{u \in \mathbb{R}} |a - b| + \frac{\beta}{2}|u - b|^2 \) for some given \( a, b \) and \( \beta > 0 \). Such a one-dimensional problem can be easily solved with high accuracy via line search or look-up table strategies. Besides the case of DNN training with squared loss and sigmoid activation function, our BCD method can also deal with many other important DNN training cases.

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 our BCD method, and our proof still goes through.

On the distributed implementation using data parallelism: 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, \( n = \sum_{j=1}^m |S_j| \), where \( \vert S_j \vert \) denotes the cardinality of \( S_j \). For each layer \( i \), the state variable \( V_i \) is divided into \( m \) submatrices by column, that is, \( V_i := (V_i^{1}, \ldots, V_i^{m}) \), where \( V_i^{1} := S_1 \), denotes the submatrix of \( V_i \) including all the columns in the index set \( S_j \). The auxiliary variables \( U_i \)'s are decomposed similarly. From Algorithm 2, the updates of \( \{ V_i \}_{i=1}^N \) and \( \{ U_i \}_{i=1}^N \) do not need any communication and thus, can be computed in a parallel way. The difficult part is the update of weight \( W_i \), which is generally hard to parallelize. To deal with this part, there are some effective strategies suggested in literature like Taylor et al. (2016), Boyd et al. (2011).
On the $V_N$-update: Note that in the $V_N$-update of both Algorithms 1 and 2, the empirical risk is involved in the optimization problems. It is generally hardly to get its closed-form solution except some special case such as the loss is the squared loss. For the other smooth losses such as the logistic, cross-entropy, and exponential losses, we suggest using the following prox-linear update strategies, that is, for some positive parameter $\alpha$,

- $V_N$-update in Algorithm 1
  \[
  V_N^k \leftarrow \arg \min_{V_N} \left\{ s_N(V_N) + \langle \nabla R_n(V_N^{k-1}; Y), V_N - V_N^{k-1} \rangle + \frac{\alpha}{2} \| V_N - V_N^{k-1} \|_F^2 \right. \]
  \[
  + \left. \frac{\gamma}{2} \| V_N - W_N^{k-1}V_N^{k-1} \|_F^2 \right\},
  \]

- $V_N$-update in Algorithm 2
  \[
  V_N^k \leftarrow \arg \min_{V_N} \left\{ s_N(V_N) + \langle \nabla R_n(V_N^{k-1}; Y), V_N - V_N^{k-1} \rangle + \frac{\alpha}{2} \| V_N - V_N^{k-1} \|_F^2 \right. \]
  \[
  + \left. \frac{\gamma}{2} \| V_N - U_N^{k-1} \|_F^2 \right\}.
  \]

3.2 Global Convergence of BCD Algorithms

In this section, we establish the global convergence of both Algorithm 2 for Problem 3, and Algorithm 1 for Problem 3. At first, we show the value convergence of both suggested algorithms as follows.

Theorem 2 (value convergence). Let \( \{Q^k\} := \{(W_i^k)_{i=1}^N; (V_i^k)_{i=1}^N\} \) and \( \{P^k\} := \{(W_i^k)_{i=1}^N; (V_i^k)_{i=1}^N; (U_i^k)_{i=1}^N\} \) be the sequences generated by Algorithms 1 and 2, respectively. Under the assumptions of Theorem 1 and finite initializations \( Q^0 \) and \( P^0 \), then for any positive $\alpha$ and $\gamma$, \( \{L(Q^k)\} \) (resp. \( \{\bar{L}(P^k)\} \)) is nonincreasing and converges to some finite $\bar{L}^\star$ (resp. $L^\star$).

The proof of this theorem is presented in Appendix B. Theorem 2 implies that the quality of the generated sequence is gradually improved during the iterative procedure in the sense of the descent of the objective, and eventually achieves some level of objective value, then keeps stable. However, the convergence of the generated sequence \( \{Q^k\} \) (resp. \( \{P^k\} \)) 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 3 (global convergence and rate). Let assumptions of Theorem 2 hold. If $\sigma_i$ is Lipschitz continuous on any bounded set, $i = 1, \ldots, N - 1$, and also one of the following conditions holds

(a) there exists a convergent subsequence \( \{Q^k_i\}_{i \in \mathbb{N}} \) (resp. \( \{P^k_i\}_{i \in \mathbb{N}} \)),

(b) $r_i$ is coercive\(^8\) for any $i = 1, \ldots, N$,

(c) $\bar{L}$ (resp. $\bar{\bar{L}}$) is coercive,

then \( \{Q^k\} \) (resp. \( \{P^k\} \)) converges to a critical point of $L$ (resp. $\bar{L}$).

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

Let $\theta$ be the KL exponent of $L$ (resp. $\bar{L}$) at $Q^\star$ (resp. $P^\star$). There hold: (a) if $\theta = 0$, then $\{Q^k\}$ converges in a finite number of steps; (b) if $\theta \in (0, \frac{1}{2})$, then $\|Q^k - Q^\star\|_F \leq C\tau^k$ 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 $\|Q^k - Q^\star\|_F \leq CK^{-\frac{\theta}{2\theta-1}}$ for $k \geq k_0$, for certain $k_0 > 0$, $C > 0$. The same claims hold for the sequence $\{P^k\}$.

The proof of this theorem is presented in Appendix C. This theorem establish the global convergence of BCDs under some natural conditions satisfied by DNN training models. It can be easily checked that many commonly used activation functions including the ReLU, leaky ReLU, sigmoid, tanh, linear

---

\(^8\)An extended-valued function $h: \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ is called coercive iff $h(x) \to +\infty$ as $\|x\| \to +\infty$.
and polynomial \cite{Liao_2017} activation functions satisfy the assumption of Theorem 3 and many commonly used regularizers $r_i$ including the squared $\ell_2$ norm, squared Frobenius norm, entrywise 1-norm and elastic net satisfy the associated assumption in Theorem 3. Compared to the existing literature, this theorem establishes the global convergence without the block multiconvex assumption used in literature \cite{Xu_2013, Xu_2017}, and also the block-wise Lipschitz differentiability used in \cite{Bolte_2014}, which are not satisfied in DNN training.

The claims in Theorem 3 still hold for the prox-linear updates adopted for the $V_N$-updates if the loss is smooth with Lipschitz gradient, shown as follows.

**Corollary 1** (global convergence and rate for prox-linear update). Consider adopting the prox-linear updates (2), (10) for the $V_N$-subproblems in Algorithm 1 and Algorithm 2 respectively. Under conditions of Theorem 3 if further $\nabla R_n$ is Lipschitz continuous with a Lipschitz constant $L_R$ and $\alpha > \max\{0, \frac{L_R}{2}\}$, then all claims in Theorem 3 still hold for both algorithms.

**Remark 1.** A popular framework on the global convergence of nonconvex BCD algorithms is in \cite{Xu_2013, Xu_2017}, that studies the following optimization problem

$$\min_{x \in \mathbb{R}^m} h(x_1, \ldots, x_p) := g(x_1, \ldots, x_p) + \sum_{i=1}^p r_i(x_i),$$  \hspace{1cm} (11)

where the variable $x$ is decomposed into $p$ blocks $x_1, \ldots, x_p$ with $x_i \in \mathbb{R}^{m_i}, \sum_{i=1}^p m_i = m$, the function $g$ is assumed to be a differentiable and block multiconvex function, and $r_i, i = 1, \ldots, p,$ are extended-valued convex functions. In order to establish the global convergence of BCD methods, it further needs the assumption that $\nabla g$ is Lipschitz continuous on any bounded set in \cite{Xu_2013}. In our case, the unregularized part of $L$ in (5), i.e., $R_n(V_N; Y) + \frac{\gamma}{2} \sum_{i=1}^N (\|V_i - \sigma_i(U_i)\|_F^2 + \|U_i - W_iV_{i-1}\|_F^2)$ might not be block multiconvex and differentiable, e.g., when $\sigma_i$ is ReLU, the function $\|V_i - \sigma_i(U_i)\|_F^2$ is non-differentiable and not convex with respect to $U_i$-block. Moreover, in \cite{Xu_2013}, the nonconvex BCD with prox-linear update scheme is especially considered, which needs certain smoothness on the unregularized part of the objective function. In contrast, both Algorithms 1 and 2 only use the minimization or proximal update strategies, and Theorem 3 only requires the Lipschitz continuity of the activation functions.

**Remark 2.** Another related work on the convergence of nonconvex BCD is \cite{Bolte_2014}, which studied the global convergence of the so-called proximal alternating linearized minimization (PALM) for solving a class of nonconvex-nonsmooth problem of the form

$$\min_{x \in \mathbb{R}^N, y \in \mathbb{R}^M} \Psi(x, y) := f(x) + g(y) + H(x, y)$$  \hspace{1cm} (12)

where $f, g$ are proper and lower semicontinuous functions with well-defined proximal maps, $H$ is a continuously differentiable function and satisfies so-called block-wise Lipschitz differentiability, that is, $H$ is Lipschitz differentiable\footnote{Lipschitz differentiability means that the function is differentiable and its gradient is Lipschitz continuous.} with respect to one block while fixing the other block. However, $\mathcal{L}$ in (3) and $\mathcal{L}$ in (5) generally do not satisfy the block-wise Lipschitz differentiable assumption, particularly when a non-smooth activation function is used (e.g. ReLU and leaky ReLU).

4 Experimental Studies

In this section, we conduct several experiments to verify our developed theoretical results, as well as to show the feasibility of the suggested BCD algorithms. We mainly show Algorithm 2 for the 3-splitting formulation of DNN training problem, since the efficiency of Algorithm 1 has been verified in literature (say, \cite{Zhang_2017}). We compare a mini-batch version of Algorithm 2 with SGD and Adam, which are two of the most widely used DNN training methods. It should be pointed out that our BCD method is currently implemented using MATLAB without optimizing the code on CPU, while SGD and Adam are implemented using Keras with TensorFlow backend with GPU.

In these experiments, we use a MLP with three hidden layers (i.e., $N = 4$ in DNN training model (2)) with ReLU and squared loss. In this scenario, all the updates of BCD have closed-form solutions. We use a MLP with three hidden layers (i.e., $N = 4$ in DNN training model (2)) with ReLU and squared loss. In this scenario, all the updates of BCD have closed-form solutions. We
implement the experiments on two datasets, i.e., MNIST [LeCun and Cortes (2010)] and CIFAR-10 [Krizhevsky et al. (2009)] datasets. The specific settings are summarized as follows:

(a) For MNIST dataset, we implement a 784-2048-2048-2048 MLP (that is, the input dimension \(d_0 = 28 \times 28 = 784\), the output dimension \(d_4 = 10\), and the numbers of neurons in three hidden layers are all 2048), and set \(\gamma = 1\) and \(\alpha = 0.05\) for BCD. The sizes of training and test samples are 60000 and 10000, respectively.

(b) For CIFAR-10 dataset, we implement a 3072-4000-1000-4000-10 MLP (that is, the input dimension \(d_0 = 3072\), the output dimension \(d_4 = 10\), and the numbers of neurons in three hidden layers are 4000, 1000 and 4000, respectively), and set \(\gamma = 0.75\) and \(\alpha = 5\) for BCD. The sizes of training and test samples are 50000 and 10000, respectively.

(c) The learning rates of SGD and Adam on both datasets are 0.05.

(d) For each experiment, we use the same mini-batch sizes (200) and initializations for all algorithms. Specifically, all the weights \(\{W_i\}_{i=1}^4\) 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 \(\{U_i\}_{i=1}^4\) and state variables \(\{V_i\}_{i=1}^4\) are initialized by a single forward pass.

Under these settings, we plot the curves of training loss defined in (1), training accuracy (Acc.), and testing accuracy (Acc.) of three methods as shown in Figure 1, while the details of these figures at the early stage are presented in Appendix E (Figure 2). The details are better seen by zooming on a computer screen. One can see from Figures 1(a)-(b) and (d)-(e) that all the three algorithms can eventually find some parameters of (approximately) zero training loss and (approximately) 100% training accuracy. Among these three algorithms, Adam is generally the fastest one in the sense of training speed at the expense of stability, followed by BCD, while SGD is generally the slowest one in early epochs. According to Figures 1(c) and (f), the testing accuracy of BCD is comparable to that of SGD; particularly Figure 1(f) shows that both BCD and Adam exhibit some overfitting phenomena for CIFAR-10 dataset, while SGD is more resistant to overfitting.

These experiments suggest that with over-parameterized models, the global convergence of BCD to a critical point of Lyapunov functions may guide us to find approximate global optima of near zero training losses (or errors).
5 Conclusion and Discussions

In this paper, we study the Kurdyka-Łojasiewicz (KL) property of DNN training problem. Such property is very crucial for the establishment of the global convergence of the nonconvex DNN training algorithms. We first show the KL properties of DNN training models with variable splitting, then adapt the popular BCD methods for the splitting formulations of the DNN training problem, and later establish the global convergence of the suggested BCD methods based on the developed KL properties. Some experiments are conducted to verify our theoretical results and also to demonstrate the feasibility of the suggested BCD methods.

Note that our main theorems show the global convergence (only to a stationary point) of BCD. In order to show the convergence to a global minima, more efforts should be made to investigate the landscape of the objective of DNN training problem. Currently, there are quite limited theoretical characterization on the landscape for special networks (say, Mei et al. (2016); Liao and Poggio (2017); Venturi et al. (2018)). Moreover, as shown in Figure 1, BCD can achieve a moderate accuracy at the very early stage (say, achieving 95% above testing accuracy within 5 epochs for MNIST data set). The early stopping criteria are of value to investigate. The convergence of the stochastic and parallel BCD algorithms for DNN training is worth of study in the future.

References


Appendix of On Global Convergence in Deep Learning with Variable Splitting

A Proof of Theorem 1

In order to prove this theorem, we need the following lemmas. The first lemma shows some important properties of real analytic functions.

Lemma 1 (Krantz and 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).

Lemma 2. The following hold

1. Finite union, finite intersection, or complementation of semialgebraic sets is 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 \( h : A \rightarrow B \) and \( g : B \rightarrow C \) is semialgebraic (Bochnak et al., 1998, Proposition 2.2.6).
3. The sum of two semialgebraic functions is a semialgebraic function (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 will involve the sum of real analytic functions and semialgebraic functions, we still need the following lemma, of which the claims can be found in Shiota (1997) or derived directly from Shiota (1997).

Lemma 3. 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 KL function.

Lemma 4 (Theorem 3.1 in Bolte et al. (2007a)). Let \( h : \mathbb{R}^p \rightarrow \mathbb{R} \cup \{+\infty\} \) be a subanalytic function with closed domain, and assume that \( h \) is continuous on its domain, then \( h \) is a KL function.

Proof of Theorem 1. We first verify the KL property of \( \tilde{L} \), then similarly show that of \( L \), and finally check some special cases arisen in deep learning.

From (5),

\[
\tilde{L}(\{W_i\}_{i=1}^N; \{V_i\}_{i=1}^N; \{U_i\}_{i=1}^N) := \mathcal{R}_n(V_N; Y) + \sum_{i=1}^N (r_i(W_i) + s_i(V_i)) + \frac{\gamma}{2} \sum_{i=1}^N (\|V_i - \sigma_i(U_i)\|^2_F + \|U_i - W_iV_i\|^2_F),
\]

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

\[
\mathcal{R}_n(V_N; Y), r_i(W_i), s_i(V_i), \|V_i - \sigma_i(U_i)\|^2_F, \|U_i - W_iV_i\|^2_F.
\]

To verify the KL property of the function \( \tilde{L} \), we consider the above functions one by one under the hypothesis of Theorem 1.
On $\mathcal{R}_n(V_N; Y)$: Note that given the output data $Y$, $\mathcal{R}_n(V_N; Y) := \frac{1}{n} \sum_{j=1}^n \ell((V_N)_j, y_j)$, where $\ell : \mathbb{R}^{dN} \times \mathbb{R}^{dN} \to \mathbb{R}_+ \cup \{0\}$ is some loss function. If $\ell$ is real analytic (resp. semialgebraic), then by Lemma 1 (resp. Lemma 2(3)), $\mathcal{R}_n(V_N; Y)$ is real-analytic (resp. semialgebraic).

On $\|V_i - \sigma_i(U_i)\|^2_2$: Note that $\|V_i - \sigma_i(U_i)\|^2_2$ is a finite sum of simple functions of the form $|v - \sigma_i(u)|^2$ for any $v, u \in \mathbb{R}$. If $\sigma_i$ is real analytic (resp. semialgebraic), then $v - \sigma_i(u)$ is real analytic (resp. semialgebraic), and further by Lemma 1 (resp. Lemma 2(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 $\|U_i - W_i V_{i-1}\|^2_2$: Note that the function $\|U_i - W_i V_{i-1}\|^2_2$ is a polynomial function with the variables $U_i$, $W_i$, and $V_{i-1}$, and thus according to Krantz and Parks (2002) and Bochnak et al. (1998), it is both real analytic and semialgebraic.

On $r_i(W_i), s_i(V_i)$: All $r_i$'s and $s_i$'s are real analytic or semialgebraic by the hypothesis of Theorem 1.

Since each part of the function $\mathcal{L}$ is either real analytic or semialgebraic, then by Lemma 3 $\mathcal{L}$ is a subanalytic function. Furthermore, by the continuous hypothesis of Theorem 1 $\mathcal{L}$ is continuous in its domain. Therefore, $\mathcal{L}$ is a KL function according to Lemma 4.

Similarly, we can verify the KL property of $\mathcal{L}$ by checking each part is either real analytic or semialgebraic. The major task is to check the KL property of functions $\|V_i - \sigma_i(W_i V_{i-1})\|^2_2$ $(i = 1, \ldots, N)$. This reduces to check the function $h : \mathbb{R} \times \mathbb{R}^{d_{i-1}} \times \mathbb{R}^{d_{i-1}} \to \mathbb{R}$, $h(u, v, w) := |u - \sigma_i(wv)|^2$. Similar to the case $|v - \sigma_i(u)|^2$ for any $u, v \in \mathbb{R}$, $h$ is real analytic (resp. semialgebraic) if $\sigma_i$ is real analytic (resp. semialgebraic) by Lemma 1 (resp. Lemma 2(2)). As a consequence, each part of function $\mathcal{L}$ is either real analytic or semialgebraic, then $\mathcal{L}$ is a subanalytic function, and further by the continuous hypothesis of Theorem 1 $\mathcal{L}$ is a KL function according to Lemma 4.

In the following, we verify some special cases.

On loss function $\ell$:

(a1) If $\ell$ is the squared ($t^2$), exponential ($e^t$), or cross-entropy loss ($\log t$), then according to Krantz and Parks (2002), $\ell((V_N)_j, y_j)$ is real-analytic for each sample $j$.

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

(a3) If $\ell$ is the hinge loss, that is, given $y \in \mathbb{R}^{dN}$, $\ell(u, y) := \max\{0, 1 - u^T y\}$ for any $u \in \mathbb{R}^{dN}$, by Lemma 1, it is semialgebraic, because its graph is $\text{cl}(D)$, a closure of the set $D$, where $D = \{(u, z) : 1 - u^T y - z = 0, 1 - u > 0\} \cup \{(u, z) : z = 0, u^T y - 1 > 0\}$.

On activation function $\sigma_i$:

(b1) If $\sigma_i$ is linear or polynomial function, then according to Krantz and Parks (2002), $\sigma_i$ is real analytic.

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

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

15
Similarly. To prove Theorem 2, we first show the following lemma.

According to Algorithm 2, the decreasing property of the sequence \( \bar{D} \), a closure of the set \( D \), where

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

(b5) If \( \sigma_i \) is polynomial as used in [Liao and Poggio, 2017], then it is obvious real analytic.

**Proof.**

The descent quantity in (13) 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 \( \alpha \) (say, updates of \( V_k^N\), \( \{U_i^{k+N-1}_{i=1}\}, \{W_i^{k+N}_{i=1}\}^N_{i=1} \)-blocks in Algorithm 2).
or by minimizing a strongly convex function\(^{11}\) with parameter \(\gamma > 0\) (say, updates of \(\{V_{i}^{k}\}_{i=1}^{N-1}, U_{N}^{k}\)-blocks in Algorithm[2]), we will consider both cases one by one.

(a) Proximal update case: In this case, we take the \(W_{i}^{k}\)-update case for example. By Algorithm[2] \(W_{i}^{k}\) is updated according to the following

\[
W_{i}^{k} \leftarrow \arg \min_{W_{i}} \left\{ r_{i}(W_{i}) + \frac{\gamma}{2} \|U_{i}^{k} - W_{i}V_{i-1}^{k-1}\|_{F}^{2} + \frac{\alpha}{2} \|W_{i} - W_{i}^{k-1}\|_{F}^{2} \right\}. \tag{16}
\]

Let \(h^{k}(W_{i}) = r_{i}(W_{i}) + \frac{\gamma}{2} \|U_{i}^{k} - W_{i}V_{i-1}^{k-1}\|_{F}^{2}\) and \(\tilde{h}^{k}(W_{i}) = r_{i}(W_{i}) + \frac{\gamma}{2} \|U_{i}^{k} - W_{i}V_{i-1}^{k-1}\|_{F}^{2} + \frac{\gamma}{2} \|W_{i} - W_{i}^{k-1}\|_{F}^{2}\). By the optimality of \(W_{i}^{k}\), it holds

\[
\tilde{h}^{k}(W_{i}^{k-1}) \geq \tilde{h}^{k}(W_{i}^{k}),
\]

which implies

\[
h^{k}(W_{i}^{k-1}) \geq h^{k}(W_{i}^{k}) + \frac{\alpha}{2} \|W_{i} - W_{i}^{k-1}\|_{F}^{2}. \tag{17}
\]

Note that the \(W_{i}^{k}\)-update (16) is equivalent to the following original proximal BCD update, i.e.,

\[
W_{i}^{k} \leftarrow \arg \min_{W_{i}} \mathcal{L}(W_{i}^{k-1}, W_{i}, V_{i}^{k-1}, V_{i}^{k}, V_{i}^{k}, V_{i}^{k}, U_{i}^{k-1}, U_{i}^{k}, U_{i}^{k}) + \frac{\alpha}{2} \|W_{i} - W_{i}^{k-1}\|_{F}^{2},
\]

where \(W_{i} := (W_{1}, W_{2}, \ldots, W_{i-1}), W_{i+1} := (W_{i+1}, W_{i+2}, \ldots, W_{N})\), and \(V_{i}, V_{i+1}, U_{i}, U_{i+1}\) are defined similarly. Thus, by (17), we establish the descent part along the \(W_{i}\)-update \((i = 1, \ldots, N-1)\), that is,

\[
\tilde{\mathcal{L}}(W_{i}^{k-1}, W_{i}^{k-1}, W_{i+1}, V_{i}^{k-1}, V_{i}^{k}, V_{i}^{k}, V_{i}^{k}, U_{i}^{k-1}, U_{i}^{k}, U_{i}^{k}) \\
\geq \tilde{\mathcal{L}}(W_{i}^{k-1}, W_{i}^{k-1}, W_{i+1}, V_{i}^{k-1}, V_{i}^{k}, V_{i}^{k}, V_{i}^{k}, U_{i}^{k-1}, U_{i}^{k}, U_{i}^{k}) + \frac{\alpha}{2} \|W_{i} - W_{i}^{k-1}\|_{F}^{2}. \tag{18}
\]

Similarly, we can establish the similar descent estimates of (18) for the other blocks using the proximal updates including \(V_{N}^{k}, (U_{i}^{k})_{i=1}^{N-1}\) and \(W_{N}^{k}\) blocks.

Specifically, for the \(V_{N}^{k}\)-block, the following holds

\[
\tilde{\mathcal{L}}(\{W_{i}^{k-1}\}_{i=1}^{N}, V_{i}^{k-1}, (U_{i}^{k-1})_{i=1}^{N}) \\
\geq \tilde{\mathcal{L}}(\{W_{i}^{k-1}\}_{i=1}^{N}, V_{i}^{k-1}, (U_{i}^{k-1})_{i=1}^{N}) + \frac{\alpha}{2} \|V_{N} - V_{N}^{k-1}\|_{F}^{2}. \tag{19}
\]

For the \(\{U_{i}^{k}\}\)-block, \(i = 1, \ldots, N-1\), the following holds

\[
\tilde{\mathcal{L}}(W_{i}^{k-1}, W_{i}^{k-1}, V_{i}^{k-1}, V_{i}^{k}, V_{i}^{k}, V_{i}^{k}, U_{i}^{k-1}, U_{i}^{k}, U_{i}^{k}) \\
\geq \tilde{\mathcal{L}}(W_{i}^{k-1}, W_{i}^{k-1}, V_{i}^{k-1}, V_{i}^{k}, V_{i}^{k}, V_{i}^{k}, U_{i}^{k-1}, U_{i}^{k}, U_{i}^{k}) + \frac{\alpha + \gamma}{2} \|U_{i} - U_{i}^{k-1}\|_{F}^{2}. \tag{20}
\]

For the \(W_{N}^{k}\)-block, the following holds

\[
\tilde{\mathcal{L}}(W_{i}^{k-1}, W_{i}^{k-1}, V_{i}^{k-1}, V_{i}^{k}, V_{i}^{k}, V_{i}^{k}, U_{i}^{k-1}, U_{i}^{k}) \\
\geq \tilde{\mathcal{L}}(W_{i}^{k-1}, W_{i}^{k-1}, V_{i}^{k-1}, V_{i}^{k}, V_{i}^{k}, V_{i}^{k}, U_{i}^{k-1}, U_{i}^{k}) + \frac{\alpha}{2} \|W_{N} - W_{N}^{k-1}\|_{F}^{2}. \tag{21}
\]

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

\[
V_{i}^{k} \leftarrow \arg \min_{V_{i}} \left\{ s_{i}(V_{i}) + \frac{\gamma}{2} \|V_{i} - \sigma_{i}(U_{i}^{k-1})\|_{F}^{2} + \frac{\alpha}{2} \|U_{i+1}^{k-1} - W_{i+1}^{k-1}V_{i}\|_{F}^{2} \right\}. \tag{22}
\]

Let \(h^{k}(V_{i}) = s_{i}(V_{i}) + \frac{\gamma}{2} \|V_{i} - \sigma_{i}(U_{i}^{k-1})\|_{F}^{2} + \frac{\alpha}{2} \|U_{i+1}^{k-1} - W_{i+1}^{k-1}V_{i}\|_{F}^{2}\). By the convexity of \(s_{i}\), the function \(h^{k}(V_{i})\) is obviously a strongly convex function with parameter no less than \(\gamma\). By the optimality of \(V_{i}^{k}\), it holds that

\[
h^{k}(V_{i}^{k-1}) \geq h^{k}(V_{i}^{k}) + \frac{\gamma}{2} \|V_{i}^{k-1} - V_{i}^{k-1}\|_{F}^{2}. \tag{23}
\]

\(^{11}\)The function \(h\) is called a strongly convex function with parameter \(\gamma > 0\) if \(h(u) \geq h(v) + \langle \nabla h(v), u - v \rangle + \frac{\gamma}{2} \|u - v\|^{2}\).
Noting the relation between $h^k(V_i)$ and $\bar{\mathcal{L}}(W_{k}^{<i}, \bar{W}_{k}^{<i}, V_{k}^{<i}, V_k, V_{>i}^{<i}, U_{>i}^{<i}, U_{>i}^{<i}, U_{>i}^{<i}, U_{>i}^{<i})$, and by (23), it yields for $i = 1, \ldots, N - 1$,

$$
\bar{\mathcal{L}}(W_{k}^{<i}, \bar{W}_{k}^{<i}, V_{k}^{<i}, V_k, V_{>i}^{<i}, U_{>i}^{<i}, U_{>i}^{<i}, U_{>i}^{<i}) \geq \bar{\mathcal{L}}(W_{k}^{<i}, \bar{W}_{k}^{<i}, V_{k}^{<i}, V_k, V_{>i}^{<i}, U_{>i}^{<i}, U_{>i}^{<i}, U_{>i}^{<i}) + \frac{\gamma}{2} \|V_i^k - V_{i}^{k-1}\|_F^2.
$$

Similarly, we can establish the similar descent estimates for the $U_{N}^{k}$-block, that is,

$$
\bar{\mathcal{L}}(W_{<N}^{k-1}, W_{<N}^{k-1}, V_{<N}^{k-1}, V_N^{k-1}, U_{<N}^{k-1}, U_{N}^{k-1}) \geq \bar{\mathcal{L}}(W_{<N}^{k-1}, W_{<N}^{k-1}, V_{<N}^{k-1}, V_N^{k-1}, U_{<N}^{k-1}, U_{N}^{k-1}) + \gamma \|U_{N}^{k-1} - U_{N}^{k-1}\|_F^2.
$$

Summing up (18)-(21) and (24)-(25) yields the descent inequality (13).

(c) Prox-linear case for $V_N$, i.e., (10): From (10), similarly, we let $h^k(V_N) := s_N(V_N) + \mathcal{R}_n(V_N; Y) + \frac{\alpha}{2} \|V_N - V_N^{k-1}\|_F^2$ and $\bar{h}^k(V_N) = s_N(V_N) + \mathcal{R}_n(V_N^{k-1}; Y) + \langle \nabla \mathcal{R}_n(V_N^{k-1}; Y), V_N - V_N^{k-1}\rangle + \frac{\gamma}{2} \|V_N - V_N^{k-1}\|_F^2$. By the optimality of $V_N^k$ and the strong convexity of $\bar{h}(V_N)$ with the modular at least $\alpha + \gamma$, it holds

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

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

$$
h^k(V_N^{k-1}) \geq h^k(V_N^k) - (\mathcal{R}_n(V_N^k; Y) - \mathcal{R}_n(V_N^{k-1}; Y) - \langle \nabla \mathcal{R}_n(V_N^{k-1}; Y), V_N^k - V_N^{k-1}\rangle) + \frac{\alpha + \gamma}{2} \|V_N^k - V_N^{k-1}\|_F^2,
$$

where the last inequality holds for the $L_H$-Lipschitz continuity of $\nabla \mathcal{R}_n$, that is, the following inequality by Nesterov (2004),

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

Summing up (18)-(21), (24)-(25) and (26) yields the descent inequality (13). □

Proof of Theorem 2. By (13), $\bar{\mathcal{L}}(P^k)$ is monotonically nonincreasing and lower bounded by 0 since each term of $\bar{\mathcal{L}}$ is nonnegative, thus, $\bar{\mathcal{L}}(P^k)$ converges to some nonnegative, finite $\bar{\mathcal{L}}^\ast$. □

Based on Lemma 5, we can obtain the following corollary.

**Corollary 2** (square summable). The following hold:

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

(b) $\|P^k - P^{k-1}\|_F \to \infty$ as $k \to \infty$.

**Proof.** Summing up (13) over $k$ from 1 to $\infty$ yields

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

which directly implies $\|P^k - P^{k-1}\|_F \to \infty$ as $k \to \infty$. □
C Proof of Theorem 3

Our proof is mainly based on the well-known 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, relative error condition and continuity condition, together with the KL property are required to establish the global convergence of a descent algorithm from the subsequence convergence, where the sufficient decrease condition and KL property have been established in Lemma 6 and Theorem 1 respectively, and the relative error condition is developed in the latter Lemma 6 while the continuity condition holds naturally due to the continuous assumption.

In the following, we first prove Theorem 3 under the subsequence convergence assumption, i.e., condition (a) of this theorem, then show that both condition (b) and condition (c) imply the boundedness of the sequence (see, Lemma 8), 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 and Bolte, 2009) Theorem 2.

C.1 Establishing relative error condition

Lemma 6 (relative error). Under conditions of Theorem 3, let $\mathcal{B}$ be an upper bound of $\mathcal{P}^{k-1}$ and $\mathcal{P}^k$ for any positive integer $k$, $L_B$ be a uniform Lipschitz constant of $\sigma$, on the bounded set $\{\mathcal{P} : \|\mathcal{P}\|_F \leq \mathcal{B}\}$, and

$$b := \max\{\gamma, \alpha + \gamma L_B, \alpha + \gamma B, \alpha + 2\gamma B^2, 2\gamma B + \gamma B^2\},$$

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

$$\text{dist}(0, \partial \mathcal{L}(\mathcal{P}^k)) \leq b \sum_{k=1}^N \|W_i^k - W_i^{k-1}\|_F + \|V_i^k - V_i^{k-1}\|_F + \|U_i^k - U_i^{k-1}\|_F$$

$$\leq b\|\mathcal{P}^k - \mathcal{P}^{k-1}\|_F,$$

(27)

where $\tilde{b} := b\sqrt{3N}$, $\text{dist}(0, S) := \inf_{s \in S} \|s\|_F$ for a set $S$, and

$$\partial \mathcal{L}(\mathcal{P}^k) := \{\partial W_i \mathcal{L}^N_{i=1}, \{\partial V_i \mathcal{L}^N_{i=1}, \{\partial U_i \mathcal{L}^N_{i=1}\}(\mathcal{P}^k)\}.$$

Proof. The inequality (28) is established via bounding each term of $\partial \mathcal{L}(\mathcal{P}^k)$.

By the optimality conditions of all updates in Algorithm 2, the following hold

$$0 \in \partial s_N(V_N^k) + \partial R_n(V_N^k; Y) + \gamma (V_N^k - U_N^{k-1}) + \alpha (V_N^k - V_N^{k-1}),$$

(or for prox-linear, $0 \in \partial s_N(V_N^k; Y) + \nabla R_n(V_N^{k-1}; Y) + \gamma (V_N^k - U_N^{k-1}) + \alpha (V_N^k - V_N^{k-1}),$)

$$0 = \gamma (U_N^k - V_N^k) + \gamma (U_N^k - W_N^{k-1}V_N^{-1}),$$

$$0 = \partial r_N(W_N^k) + \gamma (W_N^k V_N^{-1} - U_N^k) V_N^{-1} T + \alpha (W_N^k - W_N^{-1}),$$

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

$$0 \in \partial s_i(V_i^k) + \gamma (V_i^k - \sigma_i(U_i^{k-1})) + \gamma W_i^{k+1} T (W_i^{k+1} V_i^k - U_i^{k+1}),$$

$$0 \in \gamma [(\sigma_i(U_i^k) - V_i^k) \circ \partial \sigma_i(U_i^k)] + \gamma (U_i^k - W_i^{k-1} V_i^{k-1}) + \alpha (U_i^k - U_i^{k-1}),$$

$$0 \in \partial r_i(W_i^k) + \gamma (W_i^k V_i^{-1} - U_i^k) V_i^{-1} T + \alpha (W_i^k - W_i^{-1}),$$

where $V_i^k \equiv V_0 = X, \forall k$, and $\circ$ is the Hadamard product, i.e., the componentwisewise product. By the above relations, we have

$$- \alpha (V_N^k - V_N^{k-1}) \in \partial s_N(V_N^k) + \partial R_n(V_N^k; Y) + \gamma (V_N^k - U_N^{k-1}) \in \partial v_N \mathcal{L}(\mathcal{P}^k),$$

(or, $\nabla R_n(V_N^{k-1}; Y) - \gamma (V_N^k - U_N^{k-1}) \in \partial v_N \mathcal{L}(\mathcal{P}^k)$)

$$- \gamma (W_N^k - W_N^{k-1}) V_N^{-1} - \gamma W_N^{k} V_N^{-1} (V_N^{-1} - V_N^{k-1}) = \gamma (U_N^k - V_N^k) + \gamma (U_N^k - W_N^k V_N^{-1}) = \partial u_N \mathcal{L}(\mathcal{P}^k),$$

$$\gamma W_N^k \left[ W_N^{k-1} (V_N^{-1} - V_N^{k-1}) T + (V_N^{-1} - V_N^{k-1}) V_N^{-1} T \right] - \gamma U_N^k (V_N^k - V_N^{k-1}) T - \alpha (W_N^k - W_N^{k-1})$$

$$\in \partial r_N(W_N^k) + \gamma (W_N^k V_N^{-1} - U_N^k) V_N^{-1} T = \partial w_N \mathcal{L}(\mathcal{P}^k),$$

$$\in \partial s_i(V_i^k) + \gamma (V_i^k - \sigma_i(U_i^{k-1})) + \gamma W_i^{k+1} T (W_i^{k+1} V_i^k - U_i^{k+1}),$$

$$0 \in \gamma [(\sigma_i(U_i^k) - V_i^k) \circ \partial \sigma_i(U_i^k)] + \gamma (U_i^k - W_i^{k-1} V_i^{k-1}) + \alpha (U_i^k - U_i^{k-1}),$$

$$0 \in \partial r_i(W_i^k) + \gamma (W_i^k V_i^{-1} - U_i^k) V_i^{-1} T + \alpha (W_i^k - W_i^{-1}),$$

where $V_i^k \equiv V_0 = X, \forall k$, and $\circ$ is the Hadamard product, i.e., the componentwisewise product. By the above relations, we have

$$- \alpha (V_N^k - V_N^{k-1}) \in \partial s_N(V_N^k) + \partial R_n(V_N^k; Y) + \gamma (V_N^k - U_N^{k-1}) \in \partial v_N \mathcal{L}(\mathcal{P}^k),$$

(or, $\nabla R_n(V_N^{k-1}; Y) - \gamma (V_N^k - U_N^{k-1}) \in \partial v_N \mathcal{L}(\mathcal{P}^k)$)

$$- \gamma (W_N^k - W_N^{k-1}) V_N^{-1} - \gamma W_N^{k} V_N^{-1} (V_N^{-1} - V_N^{k-1}) = \gamma (U_N^k - V_N^k) + \gamma (U_N^k - W_N^k V_N^{-1}) = \partial u_N \mathcal{L}(\mathcal{P}^k),$$

$$\gamma W_N^k \left[ W_N^{k-1} (V_N^{-1} - V_N^{k-1}) T + (V_N^{-1} - V_N^{k-1}) V_N^{-1} T \right] - \gamma U_N^k (V_N^k - V_N^{k-1}) T - \alpha (W_N^k - W_N^{k-1})$$

$$\in \partial r_N(W_N^k) + \gamma (W_N^k V_N^{-1} - U_N^k) V_N^{-1} T = \partial w_N \mathcal{L}(\mathcal{P}^k),$$

$$\in \partial s_i(V_i^k) + \gamma (V_i^k - \sigma_i(U_i^{k-1})) + \gamma W_i^{k+1} T (W_i^{k+1} V_i^k - U_i^{k+1}),$$

$$0 \in \gamma [(\sigma_i(U_i^k) - V_i^k) \circ \partial \sigma_i(U_i^k)] + \gamma (U_i^k - W_i^{k-1} V_i^{k-1}) + \alpha (U_i^k - U_i^{k-1}),$$

$$0 \in \partial r_i(W_i^k) + \gamma (W_i^k V_i^{-1} - U_i^k) V_i^{-1} T + \alpha (W_i^k - W_i^{-1}),$$
for $i = N - 1, \ldots, 1$,
\[
- \gamma (\sigma_i(U^k_{i+1}) - \sigma_i(U^{k-1}_{i+1})) \in \partial s_i(V^k_{i} + \gamma (V_{i}^k - \sigma_i(U^k_{i}))) + \gamma W^k_{i+1} (W^k_{i+1} V^k_{i} - U^k_{i+1}) = \partial \gamma \bar{L}(P^k),
\]
\[
- \gamma W^{k-1}_{i} (V^k_{i+1} - V^{k-1}_{i+1}) - \gamma (W^k_{i} - W^{k-1}_{i}) V^k_{i+1} - \alpha (U^k_{i+1} - U^{k-1}_{i+1})
\]
\[
\in [\sigma_i(U^k_{i}) - \sigma_i(U^{k-1}_{i})] + \gamma (U^k_{i} - U^{k-1}_{i}) V^k_{i+1} = \partial \gamma \bar{L}(P^k),
\]
\[
\gamma W^k_{i} \left[ V^k_{i+1} (V^k_{i+1} - V^{k-1}_{i+1}) + (V^k_{i+1} - V^{k-1}_{i+1}) V^k_{i+1} - V^{k-1}_{i+1} \right] - \gamma U^k_{i+1} (V^k_{i+1} - V^{k-1}_{i+1}) - \alpha (W^k_{i} - W^{k-1}_{i})
\]
\[
\in \partial \gamma \bar{L}(P^k) + \gamma (W^k_{i+1} V^k_{i+1} - U^k_{i+1}) V^k_{i+1} = \partial \gamma \bar{L}(P^k).
\]

Based on the above relations, and by the Lipschitz continuity of the activation function on the bounded set $\{P : \|P\|_F \leq B\}$ and the bounded assumption of both $P^{k-1}$ and $P^k$, we have
\[
\|
\mathcal{G}_{V^k} \|_F \leq \alpha \|V^k - V^{k-1}\|_F + \gamma \|U^k - U^{k-1}\|_F, \quad \mathcal{G}_{V^k} \in \partial \gamma \bar{L}(P^k),
\]
\[
(\text{or,} \quad \|
\mathcal{G}_{V^k} \|_F \leq (L_R + \alpha) \|V^k - V^{k-1}\|_F + \gamma \|U^k - U^{k-1}\|_F)
\]
\[
\|
\mathcal{G}_{U^k} \|_F \leq \gamma B \|W^k - W^{k-1}\|_F + \gamma B \|V^k - V^{k-1}\|_F, \quad \mathcal{G}_{U^k} \in \partial \gamma \bar{L}(P^k),
\]
\[
\|
\mathcal{G}_{U^k} \|_F \leq 2 \gamma B \|V^k - V^{k-1}\|_F + \gamma B \|V^k - V^{k-1}\|_F + \alpha \|W^k - W^{k-1}\|_F, \quad \mathcal{G}_{U^k} \in \partial \gamma \bar{L}(P^k),
\]
\[
\text{and for } i = N - 1, \ldots, 1,
\]
\[
\|
\mathcal{G}_{V^k} \|_F \leq \gamma L_B \|U^k - U^{k-1}\|_F, \quad \mathcal{G}_{V^k} \in \partial \gamma \bar{L}(P^k),
\]
\[
\|
\mathcal{G}_{U^k} \|_F \leq \gamma B \|V^k - V^{k-1}\|_F + \gamma B \|W^k - W^{k-1}\|_F + \alpha \|U^k - U^{k-1}\|_F, \quad \mathcal{G}_{U^k} \in \partial \gamma \bar{L}(P^k),
\]
\[
\|
\mathcal{G}_{U^k} \|_F \leq (\gamma B^2 + \gamma B) \|V^k - V^{k-1}\|_F + \alpha \|W^k - W^{k-1}\|_F, \quad \mathcal{G}_{U^k} \in \partial \gamma \bar{L}(P^k).
\]

Summing up the above inequalities and after some simplifications, we can obtain (30).

\section{Proof of Theorem 3 under condition (a)}

Based on Theorem 2 and under the hypothesis that $\bar{L}$ 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, that is, there exists a subsequence $\{P^{k_j}\}_{j \in \mathbb{N}}$ and $P^*$ such that
\[
P^{k_j} \rightarrow P^* \text{ and } \bar{L}(P^{k_j}) \rightarrow \bar{L}(P^*), \text{ as } j \rightarrow \infty.
\]

Based on Lemma 5, Lemma 6, and (29), we can justify the global convergence of $P^k$ as stated in Theorem 3 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 $P^k$, that is, the convergence of $P^k$ when $P^0$ is sufficiently close to some point $P^*$. Specifically, let $\varphi, \eta, U$ be the associated parameters of the KL property of $\bar{L}$ at $P^*$, where $\varphi$ is a continuous concave function, $\eta$ is a positive constant, and $U$ is a neighborhood of $P^*$. Let $\rho$ be some constant such that $N(P^*, \rho) := \{P : \|P - P^*\|_F \leq \rho\} \subset U$, $B := \rho + \|P^*\|_F$, and $L_B$ be the uniform Lipschitz constant for $\sigma_i, i = 1, \ldots, N - 1$, within $N(P^*, \rho)$. Assume that $P^0$ satisfies the following condition
\[
\frac{\tilde{b}}{\alpha} \varphi(\bar{L}(P^0) - \bar{L}(P^*)) + 3 \sqrt{\frac{\bar{L}(P^0)}{\alpha}} + \|P^0 - P^*\|_F < \rho,
\]
where $\tilde{b} = b/\sqrt{3N}$, $b$ and $a$ are defined in (27) and (14), respectively.

**Lemma 7 (Local convergence).** Under conditions of Theorem 3, suppose that $P^0$ satisfies the condition (30), and $\bar{L}(P^k) > \bar{L}(P^*)$ for $k \in \mathbb{N}$, then
\[
\sum_{i=1}^{k} \|P^i - P^{i-1}\|_F \leq 2 \sqrt{\frac{\bar{L}(P^0)}{\alpha}} + \frac{\tilde{b}}{\alpha} \varphi(\bar{L}(P^0) - \bar{L}(P^*)), \quad \forall k \geq 1,
\]
\[
P^k \in N(P^*, \rho), \quad \forall k \in \mathbb{N}.
\]

Letting $k$ tend to infinity, (31) yields
\[
\sum_{i=1}^{\infty} \|P^i - P^{i-1}\|_F < \infty,
\]
which implies the convergence of $\{P^k\}$.

20
\textbf{Proof.} We will prove $P^k \in \mathcal{N}(P^*, \rho)$ by induction on $k$. It is obvious that $P^0 \in \mathcal{N}(P^*, \rho)$. Thus, (32) holds for $k = 0$.

For $k = 1$, we have from (13) and the nonnegativity of $\{\bar{\mathcal{L}}(P^k)\}$ that
\[
\mathcal{L}(P^0) \geq \bar{\mathcal{L}}(P^0) - \bar{\mathcal{L}}(P^1) \geq \frac{a}{2} \|P^0 - P^1\|_F^2,
\]
which implies $\|P^0 - P^1\|_F \leq \sqrt{\frac{\mathcal{L}(P^0)}{a}}$. Therefore,
\[
\|P^1 - P^*\|_F \leq \|P^0 - P^1\|_F + \|P^0 - P^*\|_F \leq \sqrt{\frac{\mathcal{L}(P^0)}{a}} + \|P^0 - P^*\|_F,
\]
which indicates $P^1 \in \mathcal{N}(P^*, \rho)$.

Suppose that $P^k \in \mathcal{N}(P^*, \rho)$ for $0 \leq k \leq K$. We proceed to show that $P^k + 1 \in \mathcal{N}(P^*, \rho)$. Since $P^k \in \mathcal{N}(P^*, \rho)$ for $0 \leq k \leq K$, it implies that $\|P^k\|_F \leq B := \rho + P^*$ for $0 \leq k \leq K$. Thus, by Lemma 6 for $1 \leq k \leq K$,
\[
\text{dist}(0, \partial \bar{\mathcal{L}}(P^k)) \leq \bar{b}\|P^k - P^{k-1}\|_F,
\]
which together with the KL inequality (3) yields
\[
\frac{1}{\varphi'(\bar{\mathcal{L}}(P^k) - \bar{\mathcal{L}}(P^*)))} \leq \frac{\bar{b}}{a}\|P^k - P^{k-1}\|_F. \tag{33}
\]
By (13), the above inequality and the concavity of $\varphi$, for $k \geq 2$, it holds
\[
a\|P^k - P^{k-1}\|_F^2 \leq \bar{\mathcal{L}}(P^{k-1}) - \bar{\mathcal{L}}(P^k) = (\bar{\mathcal{L}}(P^{k-1}) - \bar{\mathcal{L}}(P^*)) - (\bar{\mathcal{L}}(P^k) - \bar{\mathcal{L}}(P^*))
\leq \frac{\varphi'(\bar{\mathcal{L}}(P^{k-1}) - \bar{\mathcal{L}}(P^*))) - \varphi'(\bar{\mathcal{L}}(P^k) - \bar{\mathcal{L}}(P^*)))}{\varphi'(\bar{\mathcal{L}}(P^{k-1}) - \bar{\mathcal{L}}(P^*))) - \varphi'(\bar{\mathcal{L}}(P^k) - \bar{\mathcal{L}}(P^*)))}
\leq \bar{b}\|P^{k-1} - P^{k-2}\|_F \cdot \left[\varphi'(\bar{\mathcal{L}}(P^{k-1}) - \bar{\mathcal{L}}(P^*))) - \varphi'(\bar{\mathcal{L}}(P^k) - \bar{\mathcal{L}}(P^*)))\right],
\]
which implies
\[
\|P^k - P^{k-1}\|_F^2 \leq \|P^{k-1} - P^{k-2}\|_F \cdot \frac{\bar{b}}{a} \left[\varphi'(\bar{\mathcal{L}}(P^{k-1}) - \bar{\mathcal{L}}(P^*))) - \varphi'(\bar{\mathcal{L}}(P^k) - \bar{\mathcal{L}}(P^*)))\right].
\]

Taking the square root on both sides and using the inequality $2\sqrt{\alpha\beta} \leq \alpha + \beta$, the above inequality implies
\[
2\|P^k - P^{k-1}\|_F \leq \|P^{k-1} - P^{k-2}\|_F + \frac{\bar{b}}{a} \left[\varphi'(\bar{\mathcal{L}}(P^{k-1}) - \bar{\mathcal{L}}(P^*))) - \varphi'(\bar{\mathcal{L}}(P^k) - \bar{\mathcal{L}}(P^*)))\right].
\]

Summing the above inequality over $k$ from 2 to $K$ and adding $\|P^1 - P^0\|_F$ to both sides, it yields
\[
\|P^K - P^{K-1}\|_F + \sum_{k=1}^{K-1} \|P^k - P^{k-1}\|_F
\leq 2\|P^1 - P^0\|_F + \frac{\bar{b}}{a} \left[\varphi'(\bar{\mathcal{L}}(P^{k-1}) - \bar{\mathcal{L}}(P^*))) - \varphi'(\bar{\mathcal{L}}(P^k) - \bar{\mathcal{L}}(P^*)))\right],
\]
which implies
\[
\sum_{k=1}^{K} \|P^k - P^{k-1}\|_F \leq 2\sqrt{\frac{\mathcal{L}(P^0)}{a}} + \frac{\bar{b}}{a} \varphi'(\bar{\mathcal{L}}(P^0) - \bar{\mathcal{L}}(P^*))), \tag{34}
\]
and further,
\[
\|P^{K+1} - P^*\|_F \leq \|P^{K+1} - P^K\|_F + \sum_{k=1}^{K} \|P^k - P^{k-1}\|_F + \|P^0 - P^*\|_F
\leq \sqrt{\frac{\mathcal{L}(P^K) - \mathcal{L}(P^{K+1})}{a}} + 2\sqrt{\frac{\mathcal{L}(P^0)}{a}} + \frac{\bar{b}}{a} \varphi'(\bar{\mathcal{L}}(P^0) - \bar{\mathcal{L}}(P^*)) + \|P^0 - P^*\|_F
\leq 3\sqrt{\frac{\mathcal{L}(P^0)}{a}} + \frac{\bar{b}}{a} \varphi'(\bar{\mathcal{L}}(P^0) - \bar{\mathcal{L}}(P^*)) + \|P^0 - P^*\|_F < \rho,
where the second inequality holds for (13) and (34), the third inequality holds for \( \tilde{L}(P^k) - L(P^k) \leq L(P^0) \). Thus, \( P^k = N(P^*, \rho) \). Therefore, we prove this lemma.

**Proof of Theorem 3** We prove the whole sequence convergence stated in Theorem 3 according to the following two cases.

**Case 1:** \( \tilde{L}(P^{k_0}) = \tilde{L}(P^*) \) at some \( k_0 \). In this case, by Lemma 5, it holds \( P^k = P^{k_0} = P^* \) for all \( k \geq k_0 \), which implies the convergence of \( P^k \) to a limit point \( P^* \).

**Case 2:** \( \tilde{L}(P^k) > \tilde{L}(P^*) \) for all \( k \in \mathbb{N} \). In this case, since \( P^* \) is a limit point and \( \tilde{L}(P^k) \to \tilde{L}(P^*) \), by Theorem 2, there must exist an integer \( k_0 \) such that \( P^{k_0} \) is sufficiently close to \( P^* \) as required in Lemma 7 (see, the inequality (30)). Therefore, the whole sequence \( \{P^k\} \) converges according to Lemma 7. Since \( P^* \) is a limit point of \( \{P^k\} \), we have \( P^k \to P^* \).

In the next, we show \( P^* \) is a critical point of \( \tilde{L} \). By Corollary 2(b), \( \lim_{k \to \infty} \|P^k - P^{k-1}\|_F = 0 \). Furthermore, by Lemma 6,

\[
\lim_{k \to \infty} \text{dist}(0, \partial \tilde{L}(P^k)) = 0,
\]

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 7. 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 \( V_N \)-block). Thus, it is easy to show the similar descent inequality, i.e.,

\[
\mathcal{L}(Q^{k-1}) - \mathcal{L}(Q^k) \geq a\|Q^k - Q^{k-1}\|_F^2,
\]

for some \( a > 0 \). Then similar to the proof of Lemma 6, we can establish the following inequality via checking the optimality conditions of all subproblems in Algorithm 1, that is,

\[
\text{dist}(0, \partial \mathcal{L}(Q^k)) \leq b\|Q^k - Q^{k-1}\|_F,
\]

for some \( b > 0 \). By (35), (36) and the KL property of \( \mathcal{L} \) (by Theorem 1), the global convergence of Algorithm 1 can be proved via a similar proof procedure of Algorithm 2.

**C.3 Condition (b) or (c) implies condition (a)**

**Lemma 8.** Under condition (b) or (c) of Theorem 3, \( P^k \) is bounded for any \( k \in \mathbb{N} \), and thus, there exists a convergent subsequence.

**Proof.** We first show the boundedness of the sequence as well as the subsequence convergence under condition (b) of Theorem 3, then under condition (c) of Theorem 3.

1. **condition (b) implies condition (a):** We first establish the boundedness of \( W^k_i \), \( i = 1, \ldots, N \), then recursively, we establish the boundedness of \( U^k_i \) via the boundedness of \( W^k_i \) and \( V^k_{i-1} \) (noting that \( V^k_0 \equiv X \)) and latter that of \( V^k_i \) via the boundedness of \( U^k_i \) from \( i = 1 \) to \( N \).

   **(1) Boundedness of \( W^k_i \) (i = 1, ..., N):** By Lemma 5, \( \tilde{L}(P^k) < \infty \), \( \forall k \in \mathbb{N} \). Noting that each term of \( \mathcal{L} \) is nonnegative, thus, \( 0 \leq r_i(W^k_i) < \infty \) for any \( k \in \mathbb{N} \) and \( i = 1, \ldots, N \). By the coercivity of \( r_i \), \( W^k_i \) is boundedness for any \( k \in \mathbb{N} \) and \( i = 1, \ldots, N \).

   In the following, we establish the boundedness of \( U^k_i \) for any \( k \in \mathbb{N} \) and \( i = 1, \ldots, N \).

   **(2) i = 1:** Since \( \tilde{L}(P^k) < \infty \), then \( \|U^k_1 - W^k_1 X\|_F^2 < \infty \) for any \( k \in \mathbb{N} \). By the boundedness of \( W^k_1 \) and the coercivity of the function \( \|\cdot\|_F^2 \), we have the boundedness of \( U^k_1 \) for any \( k \in \mathbb{N} \). Then we show the boundedness of \( V^k_1 \) by the boundedness of \( U^k_1 \). Due to \( \mathcal{L}(P^k) < \infty \), then \( \|U^k_1 - \sigma_1(U^k_1)\|_F^2 < \infty \) for any \( k \in \mathbb{N} \). By the Lipschitz continuity of \( \sigma_1 \) and the boundedness of \( U^k_1 \), \( \sigma_1(U^k_1) \) is uniformly bounded for any \( k \in \mathbb{N} \). Thus, by the coercivity of \( \|\cdot\|_F^2 \), \( V^k_1 \) is bounded for any \( k \in \mathbb{N} \).
(3) \( i > 1 \): Recursively, we show that the boundedness of \( W^k_i \) and \( V^k_{i-1} \) implies the boundedness of \( U^k_i \), and then the boundedness of \( V^k_i \) from \( i = 2 \) to \( N \). Now, we assume that the boundedness of \( V^k_{i-1} \) has been established. Similar to (2), the boundedness of \( U^k_i \) is guaranteed by \( \|U^k_i - W^k_i V_{i-1}^k\|_F^2 < \infty \) and the boundedness of \( W^k_i \) and \( V^k_{i-1} \), and the boundedness of \( V^k_i \) is guaranteed by \( \|V^k_i - \sigma_i(U^k_i)\|_F < \infty \) and the boundedness of \( U^k_i \), as well as the Lipschitz continuity of \( \sigma_i \).

As a consequence, we prove the boundedness of \( \{P^k_i\} \) under condition (b), which implies the subsequence convergent.

2. condition (c) implies condition (a): By Lemma 5 and the finite initialization assumption, we have
\[
\bar{L}(P^k_i) \leq \bar{L}(P^0) < \infty,
\]
which implies the boundedness of \( P^k_i \) due to the coercivity of \( \bar{L} \) (i.e. condition (c)), and thus, there exists a convergent subsequence.

As a consequence, we finish this lemma.

D Closed form solution to ReLU-subproblem

From Algorithm 2, when \( \sigma_i \) is ReLU, then the \( U^k_i \)-update actually reduces to the following one-dimensional minimization problem,
\[
u^* = \arg \min_u f(u) = \frac{1}{2}(\sigma(u) - a)^2 + \frac{\gamma}{2}(u - b)^2, \tag{37}
\]
where \( \sigma(u) = \max\{0, u\} \) and \( \gamma > 0 \). The solution to the above one-dimensional minimization problem can be represented in the following lemma.

Lemma 9. The optimal solution to Problem (37) is shown as follows
\[
\text{prox}_{\frac{1}{2}(\sigma(\cdot) - a)^2}(b) = \begin{cases} 
\frac{a + \gamma b}{1 + \gamma}, & \text{if } a + \gamma b \geq 0, \ b \geq 0, \\
\frac{a + \gamma b}{1 + \gamma}, & \text{if } - (\sqrt{2\gamma(1 + \gamma)} - \gamma)a \leq \gamma b < 0, \\
b, & \text{if } - a \leq \gamma b \leq -(\sqrt{2\gamma(1 + \gamma)} - \gamma)a < 0, \\
\min\{b, 0\}, & \text{if } a + \gamma b < 0.
\end{cases}
\]

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

Case 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
\[
u^* = \begin{cases} 
\frac{a + \gamma b}{1 + \gamma}, & \text{if } a + \gamma b \geq 0, \\
0, & \text{if } a + \gamma b < 0,
\end{cases}
\tag{38}
\]
and
\[
f\left(\frac{a + \gamma b}{1 + \gamma}\right) = \frac{\gamma}{2(1 + \gamma)}(b - a)^2, \quad f(0) = \frac{1}{2}a^2 + \frac{\gamma}{2}b^2.
\]

Case b) \( u \leq 0 \): In this case,
\[
f(u) = \frac{1}{2}u^2 + \frac{\gamma}{2}(u - b)^2.
\]
It is easy to check that
\[
u^* = \begin{cases} 
0, & \text{if } b \geq 0, \\
b, & \text{if } b < 0,
\end{cases}
\tag{39}
\]
and
\[
f(b) = \frac{1}{2}a^2, \quad f(0) = \frac{1}{2}a^2 + \frac{\gamma}{2}b^2.
\]
Based on (38) and (39), we obtain the solution to Problem (37) by considering the following four cases.

**Case 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 $f(0) = \frac{1}{2} a^2 + \frac{1}{2} b^2$. It is obvious that

$$u^* = \frac{a + \gamma b}{1 + \gamma}.$$

**Case 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

$$u^* = \begin{cases} \frac{a + \gamma b}{1 + \gamma}, & \text{if } - (\sqrt{\gamma(\gamma + 1)} - \gamma)a \leq \gamma b < 0, \\ b, & \text{if } - a \leq \gamma b \leq -(\sqrt{\gamma(\gamma + 1)} - \gamma)a < 0. \end{cases}$$

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

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

Thus, the solution to Problem (37) is shown as follows

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

### E Closed form solution to Proximal Operator of Hinge Loss

Consider the following optimization problem

$$u^* = \arg \min_u g(u) := \max\{0, 1 - a \cdot u\} + \frac{\gamma}{2} (u - b)^2, \quad (40)$$

where $\gamma > 0$.

**Lemma 10.** The optimal solution to Problem (40) is shown as follows

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

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

**Case 1.** $a > 0$: In this case,

$$g(u) = \begin{cases} 1 - au + \frac{\gamma}{2} (u - b)^2, & u < a^{-1}, \\ \frac{\gamma}{2} (u - b)^2, & u \geq a^{-1}. \end{cases}$$
It is easy to show that the solution to the problem is

$$u^* = \begin{cases} 
  b + \gamma^{-1}a, & \text{if } a > 0 \text{ and } b \leq a^{-1} - \gamma^{-1}a, \\
  a^{-1}, & \text{if } a > 0 \text{ and } a^{-1} - \gamma^{-1}a < b < a^{-1}, \\
  b, & \text{if } a > 0 \text{ and } b \geq a^{-1}.
\end{cases} \quad (41)$$

Case 2. $a = 0$: It is obvious that

$$u^* = b. \quad (42)$$

Case 3. $a < 0$: Similar to Case 1,

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

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

$$u^* = \begin{cases} 
  b + \gamma^{-1}a, & \text{if } a < 0 \text{ and } b \geq a^{-1} - \gamma^{-1}a, \\
  a^{-1}, & \text{if } a < 0 \text{ and } a^{-1} < b < a^{-1} - \gamma^{-1}a, \\
  b, & \text{if } a < 0 \text{ and } b \leq a^{-1}.
\end{cases} \quad (43)$$

Thus, we finish the proof of this lemma.

F Detailed comparisons of three algorithms at early stage

![Graphs showing comparisons of performance](images)

Figure 2: Comparisons of the performance of three methods at the early stage.