frontiers in Applied Mathematics and Statistics

ORIGINAL RESEARCH
published: 06 May 2020
doi: 10.3389/fams.2020.00013

Convergence of a Relaxed Variable Splitting Coarse Gradient Descent Method for Learning Sparse Weight Binarized Activation Neural Network

Thu Dinh and Jack Xin*

Department of Mathematics, University of California, Irvine, Irvine, CA, United States

Sparsification of neural networks is one of the effective complexity reduction methods to improve efficiency and generalizability. Binarized activation offers an additional computational saving for inference. Due to vanishing gradient issue in training networks with binarized activation, coarse gradient (a.k.a. straight through estimator) is adopted in practice. In this paper, we study the problem of coarse gradient descent (CGD) learning of a one hidden layer convolutional neural network (CNN) with binarized activation function and sparse weights. It is known that when the input data is Gaussian distributed, no-overlap one hidden layer CNN with ReLU activation and general weight can be learned by GD in polynomial time at high probability in regression problems with ground truth. We propose a relaxed variable splitting method integrating thresholding and coarse gradient descent. The sparsity in network weight is realized through thresholding during the CGD training process. We prove that under thresholding of ℓ_1, ℓ_0, and transformed-ℓ_1 penalties, no-overlap binary activation CNN can be learned with high probability, and the iterative weights converge to a global limit which is a transformation of the true weight under a novel sparsifying operation. We found explicit error estimates of sparse weights from the true weights.

Keywords: sparsification, 1-bit activation, regularization, convergence, coarse gradient descent

OPEN ACCESS

Edited by:
Lucia Tabacu,
Old Dominion University, United States

Reviewed by:
Jianjun Wang,
Southwest University, China
Yuguang Wang,
University of New South Wales, Australia

****Correspondence:***
Thu Dinh
thud2@uci.edu

Specialty section:
This article was submitted to Mathematics of Computation and Data Science, a section of the journal Frontiers in Applied Mathematics and Statistics

Received: *24 January 2020*
Accepted: *14 April 2020*
Published: *06 May 2020*

Citation:
Dinh T and Xin J (2020) Convergence of a Relaxed Variable Splitting Coarse Gradient Descent Method for Learning Sparse Weight Binarized Activation Neural Network. Front. Appl. Math. Stat. 6:13. doi: 10.3389/fams.2020.00013

1. INTRODUCTION

Deep neural networks (DNN) have achieved state-of-the-art performance on many machine learning tasks such as speech recognition [1], computer vision [2], and natural language processing [3]. Training such networks is a problem of minimizing a high-dimensional non-convex and non-smooth objective function, and is often solved by first-order methods such as stochastic gradient descent (SGD). Nevertheless, the success of neural network training remains to be understood from a theoretical perspective. Progress has been made in simplified model problems. Blum and Rivest [4] showed that even training a three-node neural network is NP-hard, and Shamir [5] showed learning a simple one-layer fully connected neural network is hard for some specific input distributions. Recently, several works [6, 7] focused on the geometric properties of loss functions, which is made possible by assuming that the input data distribution is Gaussian. They showed that SGD with random or zero initialization is able to train a no-overlap neural network in polynomial time.

Another prominent issue is that DNNs contain millions of parameters and lots of redundancies, potentially causing over-fitting and poor generalization [8] besides spending unnecessary computational resources. One way to reduce complexity is to sparsify the network weights using an empirical technique called pruning [9] so that the non-essential ones are zeroed out with minimal loss of performance [10–12]. Recently a surrogate ℓ_0 regularization approach based on a continuous relaxation of Bernoulli random variables in the distribution sense is introduced with encouraging results on small size image data sets [13]. This motivated our work here to study deterministic regularization of ℓ_0 via its Moreau envelope and related ℓ_1 penalties in a one hidden layer convolutional neural network model [7]. Moreover, we consider binarized activation which further reduces computational costs [14].

The architecture of the network is illustrated in **Figure 1** similar to Brutzkus and Globerson [7]. We consider the convolutional setting in which a sparse filter $\boldsymbol{w} \in \mathbb{R}^d$ is shared among different hidden nodes. The input sample is $\boldsymbol{Z} \in \mathbb{R}^{k\times d}$. Note that this is identical to the one layer non-overlapping case where the input is $x \in \mathbb{R}^{k\times d}$ with k non-overlapping patches, each of size d. We also assume that the vectors of $\boldsymbol{Z}$ are i.i.d. Gaussian random vectors with zero mean and unit variance. Let $\mathcal{G}$ denote this distribution. Finally, let σ denote the binarized ReLU activation function, $\sigma(z) := \chi_{\{z>0\}}$ which equals 1 if $z > 0$, and 0 otherwise. The output of the network in **Figure 1** is given by:

$$h(\boldsymbol{w}, \boldsymbol{Z}) = \mathbf{1}^T \sigma(\boldsymbol{Z}\boldsymbol{w}). \tag{1}$$

We address the realizable case, where the response training data is mapped from the input training data $\boldsymbol{Z}$ by Equation (1) with a ground truth unit weight vector $\boldsymbol{w}^*$. The input training data is generated by sampling m training points $\boldsymbol{Z}^1, .., \boldsymbol{Z}^m$ from a Gaussian distribution. The learning problem seeks $\boldsymbol{w}$ to minimize the empirical risk function:

$$l(\boldsymbol{w}, \boldsymbol{Z}) := \frac{1}{m}\sum_{j=1}^{m} (h(\boldsymbol{w}, \boldsymbol{Z}^j) - h(\boldsymbol{w}^*, \boldsymbol{Z}^j))^2 \tag{2}$$

Due to binarized activation, the gradient of l in w is almost everywhere zero, hence in-effective for descent. Instead, an approximate gradient on the coarse scale, the so called coarse gradient (denoted as $\tilde{\nabla}_{\boldsymbol{w}} l$) is adopted as proxy and is proved to drive the iterations to global minimum [14].

In the limit $m \uparrow \infty$, the empirical risk l converges to the population risk:

$$f(\boldsymbol{w}) := \mathbb{E}_{\boldsymbol{Z}\sim\mathcal{G}} \left[(h(\boldsymbol{w}, \boldsymbol{Z}) - h(\boldsymbol{w}^*, \boldsymbol{Z}))^2 \right] \tag{3}$$

which is more regular in $\boldsymbol{w}$ than l. However, the "true gradient" $\nabla_{\boldsymbol{w}} f$ is inaccessible in practice. On the other hand, the coarse gradient $\tilde{\nabla}_{\boldsymbol{w}} l$ in the limit $m \uparrow \infty$ forms an acute angle with the true gradient [14]. Hence the expected coarse gradient descent (CGD) essentially minimizes the population risk f as desired.

Our task is to sparsify $\boldsymbol{w}$ in CGD. We note that the iterative thresholding algorithms (IT) are commonly used for retrieving sparse signals [[15–19] and references therein]. In high dimensional setting, IT algorithms provide simplicity and low computational cost, while also promote sparsity of the target vector. We shall investigate the convergence of CGD with simultaneous thresholding for the following objective function

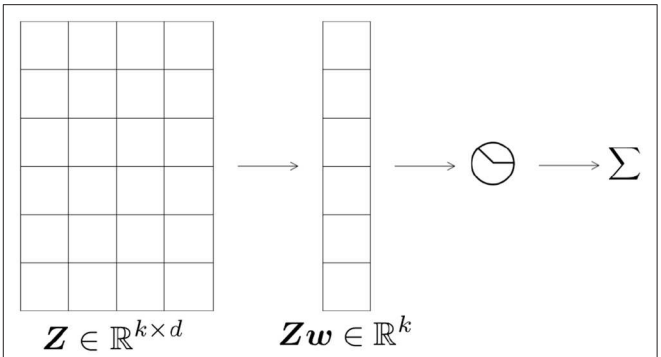

FIGURE 1 | The architecture of a no-overlap neural network.

$$\phi(\boldsymbol{w}) = f(\boldsymbol{w}) + \lambda\, P(\boldsymbol{w}) \tag{4}$$

where $f(\boldsymbol{w})$ is the population loss function of the network, and P is ℓ_0, ℓ_1, or the transformed-ℓ_1 (Tℓ_1) function: a one parameter family of bilinear transformations composed with the absolute value function [20, 21]. When acting on vectors, the Tℓ_1 penalty interpolates ℓ_0 and ℓ_1 with thresholding in closed analytical form for any parameter value [19]. The ℓ_1 thresholding function is known as soft-thresholding [15, 22], and that of ℓ_0 the hard-thresholding [17, 18]. The thresholding part should be properly integrated with CGD to be applicable for learning CNNs. As pointed out in Louizos et al. [13], it is beneficial to attain sparsity during the optimization (training) process.

1.1. Contribution

We propose a Relaxed Variable Splitting (RVS) approach combining thresholding and CGD for minimizing the following augmented objective function

$$\mathcal{L}_\beta(\boldsymbol{u}, \boldsymbol{w}) = f(\boldsymbol{w}) + \lambda\, P(\boldsymbol{u}) + \frac{\beta}{2}\,\|\boldsymbol{w} - \boldsymbol{u}\|^2$$

for a positive parameter β. We note in passing that minimizing $\mathcal{L}_\beta$ in $\boldsymbol{u}$ recovers the original objective (4) with penalty P replaced by its Moreau envelope [23]. We shall prove that our algorithm (RVSCGD), alternately minimizing $\boldsymbol{u}$ and $\boldsymbol{w}$, converges for ℓ_0, ℓ_1, and Tℓ_1 penalties to a global limit $(\bar{\boldsymbol{w}}, \bar{\boldsymbol{u}})$ with high probability. A key estimate is the Lipschitz inequality of the expected coarse gradient (Lemma 4). Then the descent of Lagrangian function (9) and the angles between the iterated $\boldsymbol{w}$ and $\boldsymbol{w}^*$ follows. The $\bar{\boldsymbol{w}}$ is a novel shrinkage of the true weight $\boldsymbol{w}^*$ up to a scalar multiple. The $\bar{\boldsymbol{u}}$ is a sparse approximation of $\boldsymbol{w}^*$. *To our best knowledge, this result is the first to establish the convergence of CGD for sparse weight binarized activation networks*. In numerical experiments, we observed that the $\bar{\boldsymbol{u}}$ limit of RVSCGD with the ℓ_0 penalty recovers sparse $\boldsymbol{w}^*$ accurately.

1.2. Outline

In section 2, we briefly overview related mathematical results in the study of neural networks and complexity reduction. Preliminaries are in section 3. In section 4, we state and discuss the main results. The proofs of the main results are in section 5, and conclusion in section 6.

2. RELATED WORK

In recent years, significant progress has been made in the study of convergence in neural network training. From a theoretical point of view, optimizing (training) neural network is a non-convex non-smooth optimization problem. Blum and Rivest [4], Livni et al. [24], Shalev-Shwartz et al. and [25] showed that training a neural network is hard in the worst cases. Shamir [5] showed that if either the target function or input distribution is "nice," optimization algorithms used in practice can succeed. Optimization methods in deep neural networks are often categorized into (stochastic) gradient descent methods and others.

Stochastic gradient descent methods were first proposed by Robbins and Monro [26]. The popular back-propagation algorithm was introduced in Rumelhart et al. [27]. Since then, many well-known SGD methods with adaptive learning rates were proposed and applied in practice, such as the Polyak momentum [28], AdaGrad [29], RMSProp [30], Adam [31], and AMSGrad [32].

The behavior of gradient descent methods in neural networks is better understood when the input has *Gaussian* distribution. Tian [6] showed that the population gradient descent can recover the true weight vector with random initialization for one-layer one-neuron model. Brutzkus and Globerson [7] proved that a convolution filter with non-overlapping input can be learned in polynomial time. Du et al. [33] showed (stochastic) gradient descent with random initialization can learn the convolutional filter in polynomial time and the convergence rate depends on the smoothness of the input distribution and the closeness of patches. Du et al. [34] analyzed the polynomial convergence guarantee of randomly initialized gradient descent algorithm for learning a one-hidden-layer convolutional neural network. A hybrid projected SGD (so called BinaryConnect) is widely used for training various weight quantized DNNs [35, 36]. Recently, a Moreau envelope based relaxation method (BinaryRelax) is proposed and analyzed to advance weight quantization in DNN training [37]. Also a blended coarse gradient descent method [14] is introduced to train fully quantized DNNs in weights and activation functions, and overcome vanishing gradients. For earlier work on coarse gradient (a.k.a. straight through estimator) (see [38–40] among others).

Non-SGD methods for deep learning include the Alternating Direction Method of Multipliers (ADMM) to transform a fully-connected neural network into an equality-constrained problem [41]; method of auxiliary coordinates (MAC) to replace a nested neural network with a constrained problem without nesting [42]. Zhang et al. [43] handled deep supervised hashing problem by an ADMM algorithm to overcome vanishing gradients.

For a similar model to (9) and treatment in a general context (see [44]); and in image processing (see [45]).

3. PRELIMINARIES

3.1. The One-Layer Non-overlap Network

Consider the network introduced in **Figure 1**. Let σ denote the binarized ReLU activation function, $\sigma(z) := \chi_{\{z>0\}}$. The training sample loss is

$$l(\boldsymbol{w}, \boldsymbol{Z}) := \frac{1}{2}(\mathbf{1}^T\sigma(\boldsymbol{Z}\boldsymbol{w}) - \mathbf{1}^T\sigma(\boldsymbol{Z}\boldsymbol{w}^*))^2, \tag{5}$$

where $\boldsymbol{w}^* \in \mathbb{R}^d$ is the underlying (non-zero) teaching parameter. Note that (5) is invariant under scaling $\boldsymbol{w} \to \boldsymbol{w}/c$, $\boldsymbol{w}^* \to \boldsymbol{w}^*/c$, for any scalar $c > 0$. Without loss of generality, we assume $\|\boldsymbol{w}^*\| = 1$. Given independent training samples $\{\boldsymbol{Z}^1, ..., \boldsymbol{Z}^N\}$, the associated empirical risk minimization reads

$$\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} l(\boldsymbol{w}, \boldsymbol{Z}^i). \tag{6}$$

The empirical risk function in (6) is piece-wise constant and has i.e., zero partial $\boldsymbol{w}$ gradient. If σ were differentiable, then back-propagation would rely on:

$$\frac{\partial\, l}{\partial \boldsymbol{w}}(\boldsymbol{w}, \boldsymbol{Z}) = \boldsymbol{Z}^T\sigma'(\boldsymbol{Z}\boldsymbol{w})(\sigma(\boldsymbol{Z}\boldsymbol{w}) - \sigma(\boldsymbol{Z}\boldsymbol{w}^*)). \tag{7}$$

However, σ has zero derivative i.e., rendering (7) inapplicable. We study the coarse gradient descent with σ' in (7) replaced by the (sub)derivative μ' of the regular ReLU function $\mu(x) := \max(x, 0)$. More precisely, we use the following surrogate of $\frac{\partial l}{\partial \boldsymbol{w}}(\boldsymbol{w}, \boldsymbol{Z})$:

$$g(\boldsymbol{w}, \boldsymbol{Z}) = \sqrt{\frac{2}{\pi}}\boldsymbol{Z}^T\mu'(\boldsymbol{Z}\boldsymbol{w})(\sigma(\boldsymbol{Z}\boldsymbol{w}) - \sigma(\boldsymbol{Z}\boldsymbol{w}^*)) \tag{8}$$

with $\mu'(x) = \sigma(x)$. The constant $\sqrt{\frac{2}{\pi}}$ represents a ReLU function μ with smaller slope, and will be necessary to give a stronger convergence result for our main findings. To simplify our analysis, we let $N \uparrow \infty$ in (6), so that its coarse gradient approaches $\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}, \boldsymbol{Z})]$. The following lemma asserts that $\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}, \boldsymbol{Z})]$ has positive correlation with the true gradient $\nabla f(\boldsymbol{w})$, and consequently, $-\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}, \boldsymbol{Z})]$ gives a reasonable descent direction.

Lemma 1. *[14] If $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \in (0, \pi)$, and $\|\boldsymbol{w}\| \neq 0$, then the inner product between the expected coarse and true gradient w.r.t. $\boldsymbol{w}$ is*

$$\left\langle \mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}, \boldsymbol{Z})], \nabla f(\boldsymbol{w}) \right\rangle = \frac{\sin(\theta(\boldsymbol{w}, \boldsymbol{w}^*))}{4\pi^2\|\boldsymbol{w}\|}k^2 \geq 0.$$

3.2. The Relaxed Variable Splitting Coarse Gradient Descent Method

Suppose we want to train the network in a way that $\boldsymbol{w}^t$ converges to a limit $\bar{\boldsymbol{w}}$ in some neighborhood of $\boldsymbol{w}^*$, and we also want to promote sparsity in the limit $\bar{\boldsymbol{w}}$. A classical approach is to

minimize the Lagrangian: $\phi(\boldsymbol{w}) = f(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1$, for some $\lambda > 0$. In practice, the ℓ_1 penalty can also be replaced by ℓ_0 or $T\ell_1$. Our proposed relaxed variable splitting (RVS) proceeds by first extending ϕ into a function of two variables $f(\boldsymbol{w}) + \lambda\|\boldsymbol{u}\|_1$, and consider the augmented Lagrangian:

$$\mathcal{L}_\beta(\boldsymbol{u}, \boldsymbol{w}) = f(\boldsymbol{w}) + \lambda \|\boldsymbol{u}\|_1 + \frac{\beta}{2}\|\boldsymbol{w} - \boldsymbol{u}\|^2 \tag{9}$$

Let S_α be the soft thresholding operator, $S_\alpha(x) = \text{sgn}(x)\max\{|x| - \alpha, 0\}$. The resulting RSVCGD method is described in Algorithm 1:

Algorithm 1: RVSCGD Algorithm

Input: The step size η, parameters λ, β
Initialize: $\boldsymbol{u}^1, \boldsymbol{w}^1$
for $t = 1, 2, \ldots, T$ **do**
  $\boldsymbol{u}^{t+1} \leftarrow \arg\min_{\boldsymbol{u}} \mathcal{L}_\beta(\boldsymbol{w}^t, \boldsymbol{u}) = S_{\lambda/\beta}(\boldsymbol{w}^t)$
  $\hat{\boldsymbol{w}}^{t+1} \leftarrow \boldsymbol{w}^t - \eta\mathbb{E}_{\boldsymbol{Z}}[\, g(\boldsymbol{w}^t, \boldsymbol{Z})] - \eta\beta(\boldsymbol{w}^t - \boldsymbol{u}^{t+1})$
  $\boldsymbol{w}^{t+1} = \frac{\hat{\boldsymbol{w}}^{t+1}}{\|\hat{\boldsymbol{w}}^{t+1}\|}$
Output: $\boldsymbol{u}^t, \boldsymbol{w}^t$

3.3. Comparison With ADMM

A well-known, modern method to solve the minimization problem $\phi(\boldsymbol{w}) = f(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1$ is the Alternating Direction Method of Multipliers (or ADMM). In ADMM, we consider the Lagrangian

$$\mathcal{L}_\beta(\boldsymbol{w}, \boldsymbol{u}, \boldsymbol{z}) = f(\boldsymbol{w}) + \lambda\|\boldsymbol{u}\|_1 + \langle \boldsymbol{z}, \boldsymbol{w} - \boldsymbol{u}\rangle + \frac{\beta}{2}\|\boldsymbol{w} - \boldsymbol{u}\|^2. \tag{10}$$

and apply the updates:

$$\begin{cases} \boldsymbol{w}^{t+1} \leftarrow \arg\min_{\boldsymbol{w}} \mathcal{L}_\beta(\boldsymbol{w}, \boldsymbol{u}^t, \boldsymbol{z}^t) \\ \boldsymbol{u}^{t+1} \leftarrow \arg\min_{\boldsymbol{u}} \mathcal{L}_\beta(\boldsymbol{w}^{t+1}, \boldsymbol{u}, \boldsymbol{z}^t) \\ \boldsymbol{z}^{t+1} \leftarrow \boldsymbol{z}^t + \beta(\boldsymbol{w}^{t+1} - \boldsymbol{u}^{t+1}) \end{cases} \tag{11}$$

Although widely used in practice, the ADMM method has several drawbacks when it comes to regularizing deep neural networks: Firstly, the ℓ_1 penalty is often replaced by ℓ_0 in practice; but $\|\cdot\|_0$ is non-differentiable and non-convex, thus current theory in optimization does not apply [46]. Secondly, the update $\boldsymbol{w}^{t+1} \leftarrow \arg\min_{\boldsymbol{w}} \mathcal{L}_\beta(\boldsymbol{w}^{t+1}, \boldsymbol{u}, \boldsymbol{z}^t)$ is not applicable in practice on DNN, as it requires one to know fully how $f(\boldsymbol{w})$ behaves. In most ADMM adaptations on DNN, this step is replaced by a simple gradient descent. Lastly, the Lagrange multiplier $\boldsymbol{z}^t$ tends to reduce the sparsity of the limit of $\boldsymbol{u}^t$, as it seeks to close the gap between $\boldsymbol{w}^t$ and $\boldsymbol{u}^t$.

In contrast, the RVSCGD method resolves all these difficulties presented by ADMM. Firstly, without the linear term $\langle \boldsymbol{z}, \boldsymbol{w} - \boldsymbol{u}\rangle$, one has an explicit formula for the update of $\boldsymbol{u}$, which can be easily implemented. Secondly, the update of $\boldsymbol{w}^t$ is not an arg min update, but rather a gradient descent iteration itself, so our theory does not deviate from practice. Lastly, without the Lagrange multiplier term $\boldsymbol{z}^t$, there will be a gap between $\boldsymbol{w}^t$ and $\boldsymbol{u}^t$ at the limit. The $\boldsymbol{u}^t$ is much more sparse than in the case of ADMM, and numerical results showed that $f(\boldsymbol{w}^t)$ and $f(\boldsymbol{u}^t)$ behave very similarly on deep networks. An intuitive explanation for this is that when the dimension of $\boldsymbol{w}^t$ is high, most of its components that will be pruned off to get $\boldsymbol{u}^t$ have very small magnitudes, and are often the redundant weights.

In short, the RVSCGD method is easier to implement (no need to keep track of the variable $\boldsymbol{z}^t$), can greatly increase sparsity in the weight variable $\boldsymbol{u}^t$, while also maintaining the same performance as the ADMM method. Moreover, RVSCGD has convergence guarantee and limit characterization as stated below.

4. MAIN RESULTS

Theorem 1. *Suppose that the initialization and penalty parameters of the RVSCGD algorithm satisfy:*
(i) $\theta(\boldsymbol{w}^0, \boldsymbol{w}^*) \le \pi - \delta$, *for some* $\delta > 0$;
(ii) $\beta \le \frac{k\sin\delta}{2\sqrt{2\pi}}$, *and* $\lambda < \frac{k}{2\sqrt{2\pi d}}$;
(iii) η *is small such* $\eta \le \min\left\{\frac{1}{\beta+L}, \frac{2\sqrt{2\pi}}{k}\right\}$, *where* L *is the Lipschitz constant in Lemma 4; and for all* t, $\eta\,\|\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}^t, \boldsymbol{Z})] + \beta\,(\boldsymbol{w}^t - \boldsymbol{u}^{t+1})\| \le \frac{1}{2}$.
Then the Lagrangian $\mathcal{L}_\beta(\boldsymbol{u}^t, \boldsymbol{w}^t)$ *decreases monotonically; and* $(\boldsymbol{u}^t, \boldsymbol{w}^t)$ *converges sub-sequentially to a limit point* $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{w}})$, *with* $\bar{\boldsymbol{u}} = S_{\lambda/\beta}(\bar{\boldsymbol{w}})$, *such that:*
(i) Let $\theta := \theta(\bar{\boldsymbol{w}}, \boldsymbol{w}^*)$ *and* $\gamma := \theta(\bar{\boldsymbol{u}}, \bar{\boldsymbol{w}})$, *then* $\theta < \delta$;
(ii) The limit point $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{w}})$ *satisfies* $\bar{\boldsymbol{u}} = S_{\lambda/\beta}(\bar{\boldsymbol{w}})$ *and*

$$\boldsymbol{w}^* = \frac{2\sqrt{2\pi}}{k}\beta(\bar{\boldsymbol{w}} - S_{\lambda/\beta}(\bar{\boldsymbol{w}})) + C\bar{\boldsymbol{w}} \tag{12}$$

where $S_{\lambda/\beta}(\cdot)$ *is the soft-thresholding operator of* ℓ_1, *for some constant* $C \ge \frac{k - 2\lambda\sqrt{2\pi d}}{k}$;
(iii) The limit point $\bar{\boldsymbol{w}}$ *is close to the ground truth* $\boldsymbol{w}^*$ *such that*

$$\|\boldsymbol{w}^* - \bar{\boldsymbol{w}}\| \le \frac{4\sqrt{2\pi}\beta\sin\gamma}{k}. \tag{13}$$

Remark 1. *As the sign of* $(\bar{\boldsymbol{w}} - S_{\lambda/\beta}(\bar{\boldsymbol{w}}))$ *agrees with* $\bar{\boldsymbol{w}}$, *Equation (12) implies that* $\boldsymbol{w}^*$ *equals an expansion of* $C\,\bar{\boldsymbol{w}}$ *or equivalently* $\bar{\boldsymbol{w}}$ *is (up to a scalar multiple) a shrinkage of* w^*, *which explains the source of sparsity in* $\bar{\boldsymbol{w}}$. *The assumption on* η *is reasonable, as will be shown below:* $\|\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}^t, \boldsymbol{Z})]\|$ *is bounded away from zero, and thus* $\|\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}^t, \boldsymbol{Z})] + \beta(\boldsymbol{w}^t - \boldsymbol{u}^{t+1})\|$ *is also bounded.*

The proof is provided in details in section 5. Here we provide an overview of the key steps. First, we show that there exists a constant L_f such that

$$\|\nabla f(\boldsymbol{w}^{t+1}) - \nabla f(\boldsymbol{w}^t)\| \le L_f\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|$$

then we show that the Lipschitz gradient property still holds when replaced by the coarse gradient:

$$\|\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}^{t+1}, \boldsymbol{Z})] - \mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}^t, \boldsymbol{Z})]\| \le K\|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|$$

and subsequently show

$$f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1) \leq \langle \mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_1, \boldsymbol{Z})], \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + \frac{L}{2}\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2.$$

These inequalities hold when $\|\boldsymbol{w}^t\| \geq M$, for some $M > 0$. It can be shown that with bad initialization, one may have $\|\boldsymbol{w}^t\| \to 0$ as $t \to \infty$. We circumvent this problem by normalizing $\boldsymbol{w}^t$ at each iteration.

Next, we show the iterations satisfy $\theta^{t+1} \leq \theta^t$, and $\mathcal{L}_\beta(\boldsymbol{u}^{t+1}, \boldsymbol{w}^{t+1}) \leq \mathcal{L}_\beta(\boldsymbol{u}^t, \boldsymbol{w}^t)$. Finally, an analysis of the stationary point yields the desired bound.

In none of these steps do we use convexity of the ℓ_1 penalty term. Here we extend our result to ℓ_0 and transformed ℓ_1 (Tℓ_1) regularization [21].

Corollary 1.1. *Suppose that the initialization of the RVSCGD algorithm satisfies the conditions in Theorem 1, and that the ℓ_1 penalty is replaced by ℓ_0 or Tℓ_1. Then the RVSCGD iterations converge to a limit point $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{w}})$ satisfying Equation (12) with ℓ_0's hard thresholding operator [18] or Tℓ_1 thresholding [19] replacing $S_{\lambda/\beta}$, and similar bound (13) holds.*

5. PROOF OF MAIN RESULTS

The following Lemmas give an outline for the proof of Theorem 1.

Lemma 2. *If every entry of $\boldsymbol{Z}$ is i.i.d. sampled from $\mathcal{N}(0,1)$, $\|\boldsymbol{w}^*\| = 1$, and $\|\boldsymbol{w}\| \neq 0$, then the true gradient of the population loss $f(\boldsymbol{w})$ is*

$$\nabla f(\boldsymbol{w}) = \frac{-k}{2\pi\|\boldsymbol{w}\|} \frac{\left(\mathbf{I} - \frac{\boldsymbol{w}\boldsymbol{w}^T}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\mathbf{I} - \frac{\boldsymbol{w}\boldsymbol{w}^T}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|}, \tag{14}$$

for $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \in (0, \pi)$; and the expected coarse gradient w.r.t. $\boldsymbol{w}$ is

$$\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}, \boldsymbol{Z})] = \frac{k}{\pi}\left[\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - \cos\left(\frac{\theta(\boldsymbol{w}, \boldsymbol{w}^*)}{2}\right)\frac{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \boldsymbol{w}^*}{\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \boldsymbol{w}^*\right\|}\right] \tag{15}$$

Lemma 3. *(Properties of true gradient)*
Given $\boldsymbol{w}_1, \boldsymbol{w}_2$ with $\min\{\|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\|\} = c > 0$ and $\max\{\|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\|\} = C$, there exists a constant $L_f > 0$ depends on c and C such that

$$\|\nabla f(\boldsymbol{w}_1) - \nabla f(\boldsymbol{w}_2)\| \leq L_f\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|$$

Moreover, we have

$$f(\boldsymbol{w}_2) \leq f(\boldsymbol{w}_1) + \langle \nabla f(\boldsymbol{w}_1), \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + \frac{L_f}{2}\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2.$$

Lemma 4. *(Properties of expected coarse gradient)*
If $\boldsymbol{w}_1, \boldsymbol{w}_2$ satisfy $\frac{1}{2} \leq \|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\| \leq \frac{3}{2}$, and $\theta(\boldsymbol{w}_1, \boldsymbol{w}^*), \theta(\boldsymbol{w}_2, \boldsymbol{w}^*) \in (0, \pi)$, then there exists a constant K such that

$$\|\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_1, \boldsymbol{Z})] - \mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_2, \boldsymbol{Z})]\| \leq K\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \tag{16}$$

Moreover, there exists a constant L such that

$$f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1) \leq \langle \mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_1, \boldsymbol{Z})], \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + \frac{L}{2}\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2. \tag{17}$$

Remark 2. *The condition $\frac{1}{2} \leq \|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\| \leq \frac{3}{2}$ in Lemma 4 is to match the RVSCGD algorithm and to give an explicit value for K. The result still holds in general when $0 < c \leq \|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\| \leq C$. Compared to Lemma 3, when $c = \frac{1}{2}$ and $C = \frac{1}{2}$, one has $L_f = \frac{4\sqrt{k}}{\pi}$, which is a sharper bound than $K = \frac{k}{\sqrt{2\pi}}$ in the coarse gradient case.*

Lemma 5. *(Angle Descent)*
Let $\theta^t := \theta(\boldsymbol{w}^t, \boldsymbol{w}^*)$. Suppose the initialization of the RVSCGD algorithm satisfies $\theta^0 \leq \pi - \delta$ and $\beta \leq \frac{k\sin\delta}{2\sqrt{2\pi}}$, then $\theta^{t+1} \leq \theta^t$.

Lemma 6. *(Lagrangian Descent)*
Suppose the initialization of the RVSCGD algorithm satisfies $\eta \leq \frac{1}{\beta + L}$, where L is the Lipschitz constant in Lemma 4, then $\mathcal{L}_\beta(\boldsymbol{u}^{t+1}, \boldsymbol{w}^{t+1}) \leq \mathcal{L}_\beta(\boldsymbol{u}^t, \boldsymbol{w}^t)$.

Lemma 7. *(Properties of limit point)*
Suppose the initialization of the RVSCGD algorithm satisfies: $\theta(\boldsymbol{w}^0, \boldsymbol{w}^*) \leq \pi - \delta$, for some $\delta > 0$, λ is small such that $\frac{2\sqrt{2\pi}}{k}\lambda\sqrt{d} < 1$, and η is small such that $\eta\frac{k}{2\sqrt{2\pi}} < 1$. Let $\theta := \theta(\bar{\boldsymbol{w}}, \boldsymbol{w}^*)$ and $\gamma := \theta(\bar{\boldsymbol{u}}, \bar{\boldsymbol{w}})$, then $(\boldsymbol{u}^t, \boldsymbol{w}^t)$ converges to a limit point $(\bar{\boldsymbol{u}}, \bar{\boldsymbol{w}})$ such that

$$\theta < \delta \text{ and } \|\boldsymbol{w}^* - \bar{\boldsymbol{w}}\| \leq \frac{4\sqrt{2\pi}\beta\sin\gamma}{k}.$$

Lemmas 2, 3 follow directly from Yin et al. [14]. The proof of Lemmas 4, 5, 6, 7 are provided below.

5.1. Proof of Lemma 4

First suppose $\|\boldsymbol{w}_1\| = \|\boldsymbol{w}_2\| = 1$. By Lemma 2, we have

$$\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_j, \boldsymbol{Z})] = \frac{k}{\pi}\left[\boldsymbol{w}_j - \cos\left(\frac{\theta(\boldsymbol{w}_j, \boldsymbol{w}^*)}{2}\right)\frac{\boldsymbol{w}_j + \boldsymbol{w}^*}{\|\boldsymbol{w}_j + \boldsymbol{w}^*\|}\right]$$

for $j = 1, 2$. Consider the plane formed by $\boldsymbol{w}_j$ and $\boldsymbol{w}^*$, since $\|\boldsymbol{w}^*\| = 1$, we have an equilateral triangle formed by $\boldsymbol{w}_j$ and $\boldsymbol{w}^*$ (see **Figure 2**).

Simple geometry shows

$$\cos\left(\frac{\theta(\boldsymbol{w}_j, \boldsymbol{w}^*)}{2}\right) = \frac{\frac{1}{2}\|\boldsymbol{w}_j + \boldsymbol{w}^*\|}{\|\boldsymbol{w}^*\|} = \frac{1}{2}\|\boldsymbol{w}_j + \boldsymbol{w}^*\|$$

Thus the expected coarse gradient simplifies to

$$\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_j, \boldsymbol{Z})] = \frac{k}{\pi}\left[\boldsymbol{w}_j - \frac{\boldsymbol{w}_j + \boldsymbol{w}^*}{2}\right] = \frac{k}{2\pi}\boldsymbol{w}_j - \frac{k}{2\pi}\boldsymbol{w}^* \tag{18}$$

which implies

$$\|\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_1, \boldsymbol{Z})] - \mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_2, \boldsymbol{Z})]\| \leq K\|\boldsymbol{w}_1 - \boldsymbol{w}_2\| \tag{19}$$

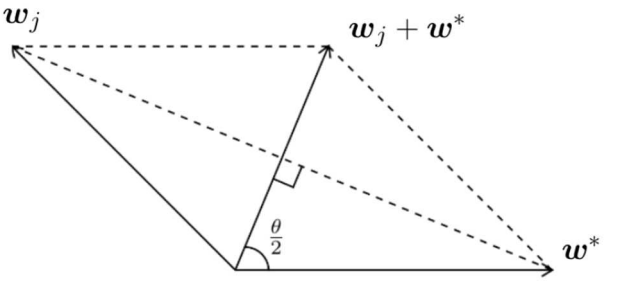

FIGURE 2 | Geometry of $\boldsymbol{w}^t$ and $\boldsymbol{w}^*$ when $\|\boldsymbol{w}^t\| = \|\boldsymbol{w}^*\| = 1$.

with $K = \frac{k}{2\pi}$.

Now suppose $\frac{1}{2} \le \|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\| \le \frac{3}{2}$. By Equation (15), we have $\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}, \boldsymbol{Z})] = \mathbb{E}_{\boldsymbol{Z}}[g(\frac{\boldsymbol{w}}{C}, \boldsymbol{Z})]$, for all $C > 0$. Then,

$$\begin{aligned}
&\|\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_1, \boldsymbol{Z})] - \mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_2, \boldsymbol{Z})]\| \\
=&\left\| \mathbb{E}_{\boldsymbol{Z}}\left[g\left(\frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|}, \boldsymbol{Z} \right) \right] - \mathbb{E}_{\boldsymbol{Z}}\left[g\left(\frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|}, \boldsymbol{Z} \right) \right] \right\| \\
\le& K' \left\| \frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|} - \frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|} \right\| \\
\le& 2K' \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|
\end{aligned}$$

where the first inequality follows from (19), and the second inequality is from the constraint $\frac{1}{2} \le \|\boldsymbol{w}_1\|, \|\boldsymbol{w}_2\| \le \frac{3}{2}$, with equality when $\|\boldsymbol{w}_1\| = \|\boldsymbol{w}_2\| = \frac{1}{2}$. Letting $K = 2K' = \frac{k}{\pi}$, the first claim is proved.

It remains to show the gradient descent inequality. By Yin et al. [14], we have

$$\begin{aligned}
f(\boldsymbol{w}) = \frac{1}{8}\Bigg[&\mathbf{1}^T(I + \mathbf{1}\mathbf{1}^T)\mathbf{1} - 2\mathbf{1}^T\left(\left(1 - \frac{2}{\pi}\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right)I + \mathbf{1}\mathbf{1}^T\right)\mathbf{1} \\
&+ \mathbf{1}^T(I + \mathbf{1}\mathbf{1}^T)\mathbf{1}\Bigg]
\end{aligned}$$

Let $\theta_1 = \theta(\boldsymbol{w}_1, \boldsymbol{w}^*), \theta_2 = \theta(\boldsymbol{w}_2, \boldsymbol{w}^*)$. Then

$$f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1) = \frac{1}{4}\left[\mathbf{1}^T \left(\left(\frac{2}{\pi}\theta_2 - \frac{2}{\pi}\theta_1 \right) I \right) \mathbf{1} \right] = \frac{k}{2\pi}(\theta_2 - \theta_1)$$

We will show

$$f(\boldsymbol{w}_2) - f(\boldsymbol{w}_1) \le \langle \mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_1, \boldsymbol{Z})], \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + L\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2$$

for $\|\boldsymbol{w}_1\| = \|\boldsymbol{w}_2\| = 1$ and $\theta_2 \le \theta_1$. By Equation (18),

$$\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}_1, \boldsymbol{Z})] = \frac{k}{2\pi}\left(\boldsymbol{w}_1 - \boldsymbol{w}^*\right)$$

It remains to show

$$\frac{k}{2\pi}(\theta_2 - \theta_1) \le \left\langle \frac{k}{2\pi}\left(\boldsymbol{w}_1 - \boldsymbol{w}^*\right), \boldsymbol{w}_2 - \boldsymbol{w}_1 \right\rangle + L\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2$$

or there exists a constant K_1 such that

$$\theta_2 - \theta_1 \le \langle \boldsymbol{w}_1 - \boldsymbol{w}^*, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + K_1\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2$$

Notice that by writing $K_1 = \frac{1}{2} + K_2$, we have

$$\begin{aligned}
&\langle \boldsymbol{w}_1 - \boldsymbol{w}^*, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + K_1\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2 \\
=&\langle \boldsymbol{w}_1 - \boldsymbol{w}^*, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + K_1\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle \\
=&\langle \boldsymbol{w}_1 - \boldsymbol{w}^*, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + \frac{1}{2}\langle \boldsymbol{w}_2 - \boldsymbol{w}_1, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + K_2\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2 \\
=&\langle \frac{1}{2}\boldsymbol{w}_1 + \frac{1}{2}\boldsymbol{w}_2 - \boldsymbol{w}^*, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + K_2\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2 \\
=&\langle -\boldsymbol{w}^*, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + \frac{1}{2}\langle \boldsymbol{w}_1 + \boldsymbol{w}_2, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + K_2\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2 \\
=&\langle -\boldsymbol{w}^*, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle + K_2\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2
\end{aligned}$$

where the last equality follows since $\|\boldsymbol{w}_1\| = \|\boldsymbol{w}_2\| = 1$ implies $\langle \boldsymbol{w}_1 + \boldsymbol{w}_2, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle = 0$. On the other hand,

$$\begin{aligned}
&\langle -\boldsymbol{w}^*, \boldsymbol{w}_2 - \boldsymbol{w}_1 \rangle \\
&= -\|\boldsymbol{w}^*\|\|\boldsymbol{w}_2\|\cos\theta_2 + \|\boldsymbol{w}^*\|\|\boldsymbol{w}_1\|\cos\theta_1 = \cos\theta_1 - \cos\theta_2
\end{aligned}$$

so it suffices to show there exists a constant K_2 such that

$$\theta_2 + \cos\theta_2 - \theta_1 - \cos\theta_1 \le K_2\|\boldsymbol{w}_2 - \boldsymbol{w}_1\|^2$$

Notice the function $\theta \mapsto \theta + \cos\theta$ is monotonically increasing on $[0, \pi]$. For $\theta_1, \theta_2 \in [0, \pi]$ with $\theta_2 \le \theta_1$, the LHS is non-positive, and the inequality holds. Thus, one can take $K_2 = 0, K_1 = \frac{1}{2}$, and $L = \frac{k}{4\pi}$.

5.2. Proof of Lemma 5

Due to normalization in the RVSCGD algorithm, $\|\boldsymbol{w}^t\| = 1$ for all t. By Equation (18), we have

$$\boldsymbol{w}^t - \eta\mathbb{E}_{\boldsymbol{Z}}[g(\boldsymbol{w}^t, \boldsymbol{Z})] = \left(1 - \eta\frac{k}{2\sqrt{2\pi}}\right)\boldsymbol{w}^t + \eta\frac{k}{2\sqrt{2\pi}}\boldsymbol{w}^*$$

and the update of $\boldsymbol{u}$ is the well-known soft-thresholding of $\boldsymbol{w}$ [15, 22]:

$$\boldsymbol{u}^{t+1} = \arg\min_{\boldsymbol{u}} \mathcal{L}_\beta(\boldsymbol{u}, \boldsymbol{w}^t) = S_{\lambda/\beta}(\boldsymbol{w}^t)$$

where $S_{\lambda/\beta}(\cdot)$ is the soft-thresholding operator:

$$S_{\lambda/\beta}(x) = \begin{cases} x - \lambda/\beta, & x > \lambda/\beta \\ 0, & |x| \le \lambda/\beta \\ x + \lambda/\beta, & x < -\lambda/\beta \end{cases}$$

and $S_{\lambda/\beta}(\boldsymbol{w})$ applies the thresholding to each component of $\boldsymbol{w}$. Then the update of $\boldsymbol{w}$ has the form

$$\boldsymbol{w}^{t+1} = C^t\boldsymbol{w}^t + \eta\frac{k}{2\sqrt{2\pi}}\boldsymbol{w}^* + \eta\beta\boldsymbol{u}^{t+1}$$

for some constant $C^t > 0$. Suppose the initialization satisfies $\theta(\boldsymbol{w}^0, \boldsymbol{w}^*) \le \pi - \delta$, for some $\delta > 0$. It suffices to show that if

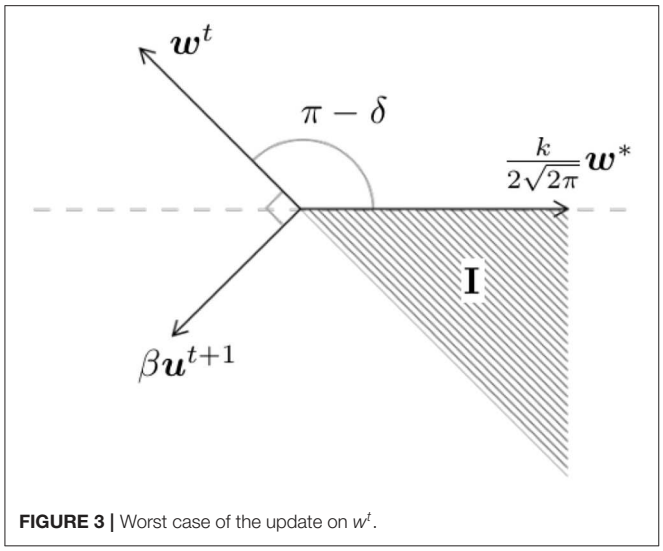

FIGURE 3 | Worst case of the update on w^t.

$\theta^t \leq \pi - \delta$, then $\theta^{t+1} \leq \pi - \delta$. To this end, since $u^{t+1} = S_{\lambda/\beta}(w^t)$, we have $\theta(w^t, u^{t+1}) \leq \frac{\pi}{2}$. Consider the worst case scenario: w^t, w^*, u^{t+1} are co-planar with $\theta(u^{t+1}, w^t) = \frac{\pi}{2}$, and w^*, u^{t+1} are on two sides of w^t (see **Figure 3**). We need $\frac{k}{2\sqrt{2\pi}} w^* + \beta u^{t+1}$ to be in region I. This condition is satisfied when β is small such that

$$\sin\delta \geq \frac{\beta\|u^{t+1}\|}{\frac{k}{2\sqrt{2\pi}}\|w^*\|} = \frac{2\sqrt{2\pi}\beta\|u^{t+1}\|}{k}$$

or

$$\beta \leq \frac{k\sin\delta}{2\sqrt{2\pi}\|u^{t+1}\|}$$

since $u^{t+1} = S_{\lambda/\beta}(w^t)$, we have $\|u^{t+1}\| \leq 1$. Thus, it suffices to have $\beta \leq \frac{k\sin\delta}{2\sqrt{2\pi}}$.

5.3. Proof of Lemma 6

By definition of the update on u, we have $\mathcal{L}_\beta(u^{t+1}, w^t) \leq \mathcal{L}_\beta(u^t, w^t)$. It remains to show $\mathcal{L}_\beta(u^{t+1}, w^{t+1}) \leq \mathcal{L}_\beta(u^{t+1}, w^t)$. First notice that since

$$w^{t+1} = C^t(w^t - \eta\mathbb{E}_Z[g(w^t, Z)] - \eta\beta(w^t - u^{t+1}))$$

where $C^t > 0$ is the normalizing constant, thus

$$\mathbb{E}_Z[g(w^t, Z)] = \frac{1}{\eta}\left(w^t - \frac{w^{t+1}}{C^t}\right) - \beta(w^t - u^{t+1})$$

For a fixed $u := u^{t+1}$ we have

$$\begin{aligned}
&\mathcal{L}_\beta(u, w^{t+1}) - \mathcal{L}_\beta(u, w^t)\\
=&f(w^{t+1}) - f(w^t) + \frac{\beta}{2}\left(\|w^{t+1} - u\|^2 - \|w^t - u\|^2\right)\\
\leq&\langle\mathbb{E}_Z[g(w^t, Z)], w^{t+1} - w^t\rangle + \frac{L}{2}\|w^{t+1} - w^t\|^2\\
&+ \frac{\beta}{2}\left(\|w^{t+1} - u\|^2 - \|w^t - u\|^2\right)\\
=&\frac{1}{\eta}\langle w^t - \frac{w^{t+1}}{C^t}, w^{t+1} - w^t\rangle - \beta\langle w^t - u, w^{t+1} - w^t\rangle\\
&+\frac{L}{2}\|w^{t+1} - w^t\|^2 + \frac{\beta}{2}\left(\|w^{t+1} - u\|^2 - \|w^t - u\|^2\right)\\
=&\frac{1}{\eta}\langle w^t - \frac{w^{t+1}}{C^t}, w^{t+1} - w^t\rangle + \left(\frac{L}{2} + \frac{\beta}{2}\right)\|w^{t+1} - w^t\|^2\\
&+\frac{\beta}{2}\|w^{t+1} - u\|^2 - \frac{\beta}{2}\|w^t - u\|^2 - \beta\langle w^t - u, w^{t+1} - w^t\rangle\\
&- \frac{\beta}{2}\|w^{t+1} - w^t\|^2\\
=&\frac{1}{\eta}\langle w^t - \frac{w^{t+1}}{C^t}, w^{t+1} - w^t\rangle + \left(\frac{L}{2} + \frac{\beta}{2}\right)\|w^{t+1} - w^t\|^2
\end{aligned}$$

Since $\|w^t\|, \|w^{t+1}\| = 1$, we know $(w^{t+1} - w^t)$ bisects the angle between w^{t+1} and $-w^t$. The assumption $\|\eta\mathbb{E}_Z[g(w^t, Z)] + \eta\beta(w^t - u^{t+1})\| \leq \frac{1}{2}$ guarantees $\frac{2}{3} \leq C^t \leq 2$ and $\theta(-w^t, w^{t+1}) < \pi$. It follows that $\theta(w^{t+1} - w^t, w^t)$ and $\theta(w^{t+1} - w^t, w^{t+1})$ are strictly less than $\frac{\pi}{2}$. On the other hand, $\left(\frac{w^{t+1}}{C^t} - w^t\right)$ also lies in the plane bounded by w^{t+1} and $-w^t$. Therefore,

$$\theta\left(\frac{w^{t+1}}{C^t} - w^t, w^{t+1} - w^t\right) < \frac{\pi}{2}.$$

This implies $\langle\frac{w^{t+1}}{C^t} - w^t, w^{t+1} - w^t\rangle \geq 0$. Moreover, when $C^t \geq 1$:

$$\begin{aligned}
\langle\frac{w^{t+1}}{C^t} - w^t, w^{t+1} - w^t\rangle =&\langle\frac{w^{t+1}}{C^t} - \frac{w^t}{C^t}, w^{t+1} - w^t\rangle\\
&-\langle\frac{C^t - 1}{C^t}w^t, w^{t+1} - w^t\rangle\\
\geq&\frac{1}{C^t}\|w^{t+1} - w^t\|^2
\end{aligned}$$

And when $\frac{2}{3} \leq C^t \leq 1$:

$$\begin{aligned}
\langle\frac{w^{t+1}}{C^t} - w^t, w^{t+1} - w^t\rangle =&\langle w^{t+1} - w^t, w^{t+1} - w^t\rangle\\
&+\langle\frac{1 - C^t}{C^t}w^{t+1}, w^{t+1} - w^t\rangle\\
\geq&\|w^{t+1} - w^t\|^2
\end{aligned}$$

Thus, we have

$$\mathcal{L}_\beta(u, w^{t+1}) - \mathcal{L}_\beta(u, w^t) \leq \left(\frac{L}{2} + \frac{\beta}{2} - \frac{\chi_{\{C^t \geq 1\}}}{\eta C^t} - \frac{\chi_{\{\frac{2}{3} \leq C^t \leq 1\}}}{\eta}\right)\|w^{t+1} - w^t\|^2$$

Therefore, if η is small so that $\eta \leq \frac{2}{C^t(\beta + L)}$ and $\eta \leq \frac{2}{\beta + L}$, the update on w will decrease $\mathcal{L}_\beta$. Since $C^t \leq 2$, the condition is satisfied when $\eta \leq \frac{1}{\beta + L}$.

5.4. Proof of Lemma 7

Since $\mathcal{L}_\beta(u^t, w^t)$ is non-negative, by Lemma 5, 6, $\mathcal{L}_\beta$ converges to some limit $\mathcal{L}$. This implies (u^t, w^t) converges to some stationary point $(\bar{u}, \bar{w})$. By the update of w^t, we have

$$\bar{w} = \bar{C}(c_1\bar{w} + \eta c_2 w^* + \eta\beta\bar{u}) \tag{20}$$

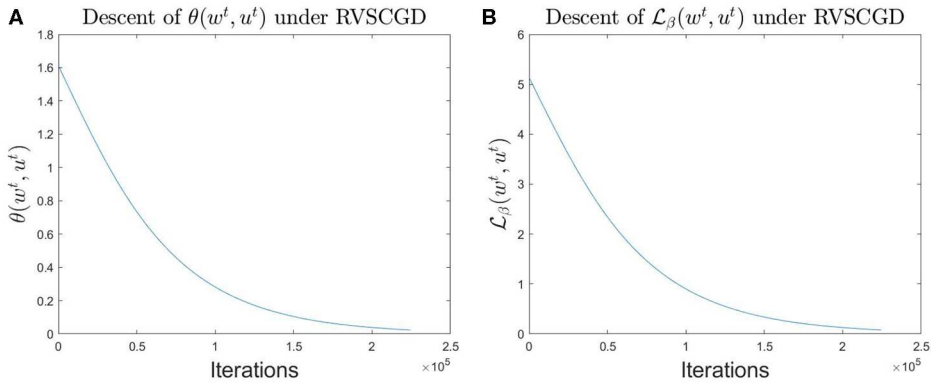

FIGURE 4 | Behavior of $\theta(\boldsymbol{w}^t, \boldsymbol{w}^*)$ and $\mathcal{L}_\beta(\boldsymbol{w}^t, \boldsymbol{u}^t)$ on the one-layer non-overlap network.

TABLE 1 | Accuracy and sparsity of RVSCGD on a LeNet variation, on the MNIST dataset.

Penalty	β	λ	Accuracy	Sparsity
Base model	1	0	89.31	0
RGSM (GL)	1	1.e-7	87.17	33.31
	1	1.e-5	85.34	66.67
	1	1.e-3	84.92	83.76

for some constant $\bar{C}, c_1, c_2 > 0$, where $c_2 = \frac{k}{2\sqrt{2\pi}}$, $c_1 > 0$ due to our assumption, and $\bar{\boldsymbol{u}} = S_{\lambda/\beta}(\bar{\boldsymbol{w}})$. For expression (20) to hold, we need

$$c_2 \boldsymbol{w}^* + \beta \bar{\boldsymbol{u}} /\!/ \bar{\boldsymbol{w}} \tag{21}$$

Expression (21) implies $\bar{\boldsymbol{w}}, \bar{\boldsymbol{u}}$, and $\boldsymbol{w}^*$ are co-planar. Let $\gamma := \theta(\bar{\boldsymbol{u}}, \bar{\boldsymbol{w}})$. From expression (21), and the fact that $\|\bar{\boldsymbol{w}}\| = \|\boldsymbol{w}^*\| = 1$, we have

$$(\langle c_2 \boldsymbol{w}^* + \beta \bar{\boldsymbol{u}}, \bar{\boldsymbol{w}} \rangle)^2 = \|c_2 \boldsymbol{w}^* + \beta \bar{\boldsymbol{u}}\|^2 \|\bar{\boldsymbol{w}}\|^2$$

which implies $c_2^2 \cos^2\theta + 2c_2\beta\|\bar{\boldsymbol{u}}\| \cos\theta \cos\gamma + \beta^2\|\bar{\boldsymbol{u}}\|^2 \cos^2\gamma = c_2^2 + 2c_2\beta\|\bar{\boldsymbol{u}}\| \cos(\theta+\gamma) + \beta^2\|\bar{\boldsymbol{u}}\|^2$ Recall $\cos(a+b) = \cos a \cos b - \sin a \sin b$. Thus,

$$c_2^2 \sin^2\theta - 2c_2\beta\|\bar{\boldsymbol{u}}\| \sin\theta \sin\gamma + \beta^2\|\bar{\boldsymbol{u}}\|^2 \sin^2\gamma = 0$$

which implies

$$\frac{k}{2\sqrt{2\pi}} \sin\theta = \beta\|\bar{\boldsymbol{u}}\| \sin\gamma \tag{22}$$

By the initialization of β, we have $\frac{k}{2\sqrt{2\pi}} \sin\theta < \frac{k}{2\sqrt{2\pi}} \sin\delta$. This implies $\theta < \delta$.

Finally, expression (20) can also be written as

$$\left(\boldsymbol{w}^* - \frac{2\sqrt{2\pi}}{k}\beta(\bar{\boldsymbol{w}} - \bar{\boldsymbol{u}})\right) /\!/ \bar{\boldsymbol{w}} \tag{23}$$

From expression (23), we see that $\boldsymbol{w}^*$, after subtracting some vector whose signs agree with $\bar{\boldsymbol{w}}$, and whose non-zero components are at most $\frac{2\sqrt{2\pi}}{k}\lambda$, is parallel to $\bar{\boldsymbol{w}}$. This implies $\bar{\boldsymbol{w}}$ is some soft-thresholded version of $\boldsymbol{w}^*$, modulo normalization. Moreover, since $\left\|\frac{2\sqrt{2\pi}}{k}\beta(\bar{\boldsymbol{w}} - \bar{\boldsymbol{u}})\right\| \leq \frac{2\sqrt{2\pi}}{k}\lambda\sqrt{d}$, for small λ such that $\frac{2\sqrt{2\pi}}{k}\lambda\sqrt{d} < 1$, we must have

$$\theta\left(\boldsymbol{w}^* - \frac{2\sqrt{2\pi}}{k}\beta(\bar{\boldsymbol{w}} - \bar{\boldsymbol{u}}), \bar{\boldsymbol{w}}\right) = 0$$

On the other hand,

$$\begin{aligned}\left\|\boldsymbol{w}^* - \frac{2\sqrt{2\pi}}{k}\beta(\bar{\boldsymbol{w}} - \bar{\boldsymbol{u}})\right\| &\geq \|\boldsymbol{w}^*\| - \left\|\frac{2\sqrt{2\pi}}{k}\beta(\bar{\boldsymbol{w}} - \bar{\boldsymbol{u}})\right\| \\ &\geq 1 - \frac{2\sqrt{2\pi}}{k}\lambda\sqrt{d}\end{aligned}$$

therefore, $\boldsymbol{w}^* - \frac{2\sqrt{2\pi}}{k}\beta(\bar{\boldsymbol{w}} - \bar{\boldsymbol{u}}) = C\bar{\boldsymbol{w}}$, for some constant C such that $C \geq \frac{k - 2\lambda\sqrt{2\pi d}}{k}$.

Finally, consider the equilateral triangle with sides $\boldsymbol{w}^*, \bar{\boldsymbol{w}}$, and $\boldsymbol{w}^* - \bar{\boldsymbol{w}}$. By the law of sines,

$$\frac{\|\boldsymbol{w}^* - \bar{\boldsymbol{w}}\|}{\sin\theta} = \frac{\|\boldsymbol{w}^*\|}{\sin\theta(\bar{\boldsymbol{w}}, \boldsymbol{w}^* - \bar{\boldsymbol{w}})} = \frac{1}{\sin\theta(\bar{\boldsymbol{w}}, \boldsymbol{w}^* - \bar{\boldsymbol{w}})}$$

as θ is small, $\theta(\bar{\boldsymbol{w}}, \boldsymbol{w}^* - \bar{\boldsymbol{w}})$ is near $\frac{\pi}{2}$. We can assume $\sin\theta(\bar{\boldsymbol{w}}, \boldsymbol{w}^* - \bar{\boldsymbol{w}}) \geq \frac{1}{2}$. Together with expression (22), we have

$$\|\boldsymbol{w}^* - \bar{\boldsymbol{w}}\| \leq 2\sin\theta = \frac{4\sqrt{2\pi}\beta\|\bar{\boldsymbol{u}}\| \sin\gamma}{k} \leq \frac{4\sqrt{2\pi}\beta \sin\gamma}{k}.$$

5.5. Proof of Theorem 1

Combining Lemmas 2–7, Theorem 1 is proved.

5.6. Proof of Corollary

Lemma 8. *[19] Let*

$$f_{\lambda,x}(y) = \frac{1}{2}(y-x)^2 + \lambda\,\rho_a(y),$$

$$g_\lambda(x) = sgn(x)\left\{\frac{2}{3}(a+|x|)\cos\left(\frac{\phi(x)}{3}\right) - \frac{2a}{3} + \frac{|x|}{3}\right\}$$

where $\phi(x) = \arccos\left(1 - \frac{27\lambda a(a+1)}{2(a+|x|)^3}\right)$. *Then* $y^*_\lambda(x) = \arg\min_y f_{\lambda,x}(y)$ *is the* $T\ell_1$ *thresholding, equal to* $g_\lambda(x)$ *if* $|x| > t$; *zero elsewhere. Here* $t = \lambda\frac{a+1}{a}$ *if* $\lambda \le \frac{a^2}{2(a+1)}$; $t = \sqrt{2\lambda(a+1)} - \frac{a}{2}$, *elsewhere.*

Lemma 9. *[18] Let* $f_{\lambda,x}(y) = \frac{1}{2}(y-x)^2 + \lambda\,\|y\|_0$. *Then* $y^*_\lambda(x) = \arg\min_y f_{\lambda,x}(y)$ *is the* ℓ_0 *hard thresholding* $y^*_\lambda(x) = x$, *if* $|x| > \sqrt{2\lambda}$; *zero elsewhere.*

We proceed by an outline similar to the proof of Theorem 1:

Step 1. First we show that $L_{\beta,T\ell_1}(\boldsymbol{u}^t, \boldsymbol{w}^t)$ and $L_{\beta,0}(\boldsymbol{u}^t, \boldsymbol{w}^t)$ both decrease under the update of $\boldsymbol{u}^t$ and $\boldsymbol{w}^t$. To see this, notice that the update on $\boldsymbol{u}^t$ decreases $L_{\beta,T\ell_1}(\boldsymbol{u}^t, \boldsymbol{w}^t)$ and $L_{\beta,0}(\boldsymbol{u}^t, \boldsymbol{w}^t)$ by definition. Then, for a fixed $\boldsymbol{u} = \boldsymbol{u}^{t+1}$, the update on $\boldsymbol{w}^t$ decreases $L_{\beta,T\ell_1}(\boldsymbol{u}^t, \boldsymbol{w}^t)$ and $L_{\beta,0}(\boldsymbol{u}^t, \boldsymbol{w}^t)$ by a similar argument to that found in Theorem 1.

Step 2. Next, we show $\theta(\boldsymbol{w}^t, \boldsymbol{w}^*) \le \pi - \delta$, for some $\delta > 0$, for all t, with initialization $\theta(\boldsymbol{w}^0, \boldsymbol{w}^*) = \pi - \delta$. For $L_{\beta,T\ell_1}(\boldsymbol{u}^t, \boldsymbol{w}^t)$, by Lemma 8, we have

$$\boldsymbol{u}^{t+1} = (g_{\lambda/\beta}(w_1^t), g_{\lambda/\beta}(w_2^t), ..., g_{\lambda/\beta}(w_d^t))$$

And for $L_{\beta,0}(\boldsymbol{u}^t, \boldsymbol{w}^t)$, by Lemma 9,

$$\boldsymbol{u}^{t+1} = (w_1^t \chi_{\{|w_1^t| \ge t\}}, w_2^t \chi_{\{|w_2^t| \ge t\}}, ...)$$

In both cases, each component of $\boldsymbol{u}^{t+1}$ is a thresholded version of the corresponding component of $\boldsymbol{w}^t$. This implies $\theta(\boldsymbol{u}^{t+1}, \boldsymbol{w}^t) \le \frac{\pi}{2}$, and thus the argument in Theorem 1 follows through, and we have $\theta(\boldsymbol{w}^t, \boldsymbol{w}^*) \le \pi - \theta$, for all t.

Step 3. Finally, the equilibrium condition from Equation (21) still holds for the limit point, and a similar argument shows that $\theta(\bar{\boldsymbol{w}}, \boldsymbol{w}^*) < \delta$.

6. NUMERICAL EXPERIMENTS

In this section, we demonstrate two simple experiments on implementing RVSCGD in practice.

Firstly, we numerically verify our result on the one-layer, non-overlap network, using RVSCGD with ℓ_0 penalty. The experiment was run with parameters $k = 20, d = 50, \beta = 4.e-3, \lambda = 1.e-4$, and $\eta = 1.e-5$. Results are displayed in **Figure 4**. It can be seen that the RVSCGD converges quickly for this toy model; and the quantities $\mathcal{L}_\beta(\boldsymbol{w}^t, \boldsymbol{u}^t)$, $\theta(\boldsymbol{w}^t, \boldsymbol{w}^*)$, decrease monotonically, as stated in Theorem 1.

Secondly, we extend our method to a multi-layer network. Consider a variation of LeNet [47], where we replace all ReLU activations with the binarized ReLU function. The model is then trained on the MNIST dataset for 100 epochs using SGD with momentum 0.9, weight decay 5.e-4, and learning rate 1.e-3, which is decayed by a factor of 10 at epoch 60. The RVSCGD algorithm is applied on this model using the same training setting. The results are displayed in **Table 1**. Notice that the base model has an accuracy of 89.13%, which is lower than reported in Lecun [47]; this is because of the binarized ReLU replacement. **Table 1** also shows that RVSCGD can effectively sparsify this variation of LeNet, with sparsity up to 83.76 and 4.39% loss in performance. We believe the loss in accuracy is mainly from 1-bit ReLU activation, which has too low a resolution to preserve important deep network information. We believe with higher bit quantization of weights and/or activations, networks can be more effectively pruned while still maintaining good performance (see [14]). This is a topic for our future studies.

7. CONCLUSION

We introduced a variable splitting coarse gradient descent method to learn a one-hidden layer neural network with sparse weight and binarized activation in a regression setting. The proof is based on the descent of a Lagrangian function and the angle between the sparse and true weights, and applies to ℓ_1, ℓ_0 and $T\ell_1$ sparse penalties. We plan to extend our work to a classification setting in the future.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

TD performed the analysis. All authors contributed to the discussions and production of the manuscript.

FUNDING

The work was partially supported by NSF grants IIS-1632935 and DMS-1854434.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at http://export.arxiv.org/pdf/1901.09731 [48].

REFERENCES

1. Hinton G, Deng L, Yu D, Dahl GE, Mohamed A, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag*. (2012) **29**:82–97. doi: 10.1109/MSP.2012.2205597
2. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Bartlett PL, Pereira FCN, Burges CJC, Bottou L, Weinberger KQ, editors. *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012* (2012). p. 1106–14. Available online at: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks
3. Dauphin YN, Fan A, Auli M, Grangier D. Language modeling with gated convolutional networks. In: *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. Sydney, NSW (2017). p. 933–41. doi: 10.5555/3305381.3305478
4. Blum AL, Rivest RL. Training a 3-node neural network is NP-complete. In: José HS, Werner R, Rivest RL, editors. *Machine Learning: From Theory to Applications: Cooperative Research at Siemens and MIT*. Berlin; Heidelberg: Springer (1993). p. 9–28. doi: 10.1007/3-540-56483-7_20
5. Shamir O. Distribution-specific hardness of learning neural networks. *J Mach Learn Res*. (2018) 19:1532–35. doi: 10.5555/3291125.3291157
6. Tian Y. An analytical formula of population gradient for two-layered ReLU network and its applications in convergence and critical point analysis. In: *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. Sydney, NSW: JMLR.org (2017). p. 3404–13. doi: 10.5555/3305890.3306033
7. Brutzkus A, Globerson A. Globally optimal gradient descent for a convnet with Gaussian inputs. In: *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. Sydney, NSW: JMLR.org (2017). p. 605–14.
8. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. In: *5th International Conference on Learning Representations, ICLR 2017*. Toulon (2017). Available online at: https://openreview.net/forum?id=Sy8gdB9xx
9. LeCun Y, Denker J, Solla S. Optimal brain damage. In: *Proceedings of the 2nd International Conference on Neural Information Processing Systems*. Cambridge, MA: MIT Press (1989). p. 589–605. doi: 10.5555/109230.109298
10. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural network with pruning, trained quantization and huffman coding. In: Bengio Y, LeCun Y, editors. *4th International Conference on Learning Representations, ICLR 2016*. San Juan (2016). Available online at: http://arxiv.org/abs/1510.00149
11. Ullrich K, Meeds E, Welling M. Soft weight-sharing for neural network compression. In: *5th International Conference on Learning Representations, ICLR 2017*. Toulon (2017). Available online at: https://openreview.net/forum?id=HJGwcKclx
12. Molchanov D, Ashukha A, Vetrov D. Variational dropout sparsifies deep neural networks. In: *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. Sydney, NSW: JMLR.org (2017). p. 2498–507. doi: 10.5555/3305890.3305939
13. Louizos C, Welling M, Kingma D. Learning sparse neural networks through L_0 regularization. In: *6th International Conference on Learning Representations, ICLR 2018*. Vancouver, BC (2018). Available online at: https://openreview.net/forum?id=H1Y8hhg0b
14. Yin P, Zhang S, Lyu J, Osher S, Qi Y, Xin J. Blended coarse gradient descent for full quantization of deep neural networks. *Res Math Sci*. (2019) **6**:14. doi: 10.1007/s40687-018-0177-6
15. Daubechies I, Defrise M, Mol CD. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun Pure Appl Math*. (2004) **57**:1413–57. doi: 10.1002/cpa.20042
16. Candés E, Romberg J, Tao T. Stable signal recovery from incomplete and inaccurate measurements. *Commun Pure Appl Math*. (2006) **59**:1207–23. doi: 10.1002/cpa.20124
17. Blumensath T, Davies M. Iterative thresholding for sparse approximations. *J Fourier Anal Appl*. (2008) **14**:629–54. doi: 10.1007/s00041-008-9035-z
18. Blumensath T. Accelerated iterative hard thresholding. *Signal Process*. (2012) **92**:752–6. doi: 10.1016/j.sigpro.2011.09.017
19. Zhang S, Xin J. Minimization of transformed l_1 penalty: closed form representation and iterative thresholding algorithms. *Commun Math Sci*. (2017) **15**:511–37. doi: 10.4310/CMS.2017.v15.n2.a9
20. Nikolova M. Local strong homogeneity of a regularized estimator. *SIAM J Appl Math*. (2000) **61**:633–58. doi: 10.1137/S0036139997327794
21. Zhang S, Xin J. Minimization of transformed l_1 penalty: theory, difference of convex function algorithm, and robust application in compressed sensing. *Math Program Ser B*. (2018) **169**:307–36. doi: 10.1007/s10107-018-1236-x
22. Donoho D. Denoising by soft-thresholding. *IEEE Trans Inform Theor*. (1995) **41**:613–27. doi: 10.1109/18.382009
23. Moreau JJ. Proximité et dualité dans un espace hilbertien. *Bull Soc Math France*. (1965) **93**:273–99. doi: 10.24033/bsmf.1625
24. Livni R, Shalev-Shwartz S, Shamir O. On the computational efficiency of training neural networks. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Vol. 1. Cambridge, MA: MIT Press (2014). p. 855–63.
25. Shalev-Shwartz S, Shamir O, Shammah S. Failures of gradient-based deep learning. In: Precup D, Teh YW, editors. *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. Sydney, NSW: PMLR (2017). p. 3067–75. Available online at: http://proceedings.mlr.press/v70/shalev-shwartz17a.html
26. Robbins H, Monro S. A stochastic approximation method. *Ann Math Stat*. (1951) **22**:400–7. doi: 10.1214/aoms/1177729586
27. Rumelhart D, Hinton G, Williams R. Learning representations by back-propagating errors. *Nature*. (1986) **323**:533–6. doi: 10.1038/323533a0
28. Polyak B. Some methods of speeding up the convergence of iteration methods. *USSR Comput Math Math Phys*. (1964) **4**:1–17. doi: 10.1016/0041-5553(64)90137-5
29. Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. *J Machine Learn Res*. (2011) **12**:2121–59. doi: 10.5555/1953048.2021068
30. Tieleman T, Hinton G. *Divide the Gradient by a Running Average of Its Recent Magnitude*. Technical report. Coursera: Neural networks for machine learning (2017).
31. Kingma D, Ba J. Adam: a method for stochastic optimization. In: Bengio Y, LeCun Y, editors. *3rd International Conference on Learning Representations, ICLR 2015*. San Diego, CA (2015). Available online at: http://arxiv.org/abs/1412.6980
32. Reddi SJ, Kale S, Kumar S. On the convergence of adam and beyond. *In: 6th International Conference on Learning Representations, ICLR 2018*. Vancouver, BC (2018). Available online at: https://openreview.net/forum?id=ryQu7f-RZ
33. Du S, Lee J, Tian Y. When is a convolutional filter easy to learn? *arXiv [preprint] arXiv:1709.06129*. (2017).
34. Du S, Lee J, Tian Y, Singh A, Poczos B. Gradient descent learns one-hidden-layer CNN: don't be afraid of spurious local minima. In: Dy J, Krause A, editors. *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. Stockholm: PMLR (2018). p. 1339–48. Available online at: http://proceedings.mlr.press/v80/du18b.html
35. Courbariaux M, Bengio Y, David J-P. BinaryConnect: training deep neural networks with binary weights during propagations. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Vol. 2 (Cambridge, MA: MIT Press (2015). p. 3123–31.
36. Yin P, Zhang S, Qi Y, Xin J. Quantization and training of low bit-width convolutional neural networks for object detection. *J Comput Math*. (2019) 37:349–59. doi: 10.4208/jcm.1803-m2017-0301
37. Yin P, Zhang S, Lyu J, Osher S, Qi Y, Xin J. BinaryRelax: a relaxation approach for training deep neural networks with quantized weights. *SIAM J Imag Sci*. (2018) 11:2205–23. doi: 10.1137/18M1166134
38. Hinton G. *Neural Networks for Machine Learning, Coursera*. Coursera, Video Lectures (2012).
39. Hubara I, Courbariaux M, Soudry D, El-Yaniv R, Bengio Y. Binarized neural networks: training neural networks with weights and activations constrained to $+1$ or -1. *arXiv [preprint] arXiv:160202830*. (2016).
40. Cai Z, He X, Sun J, Vasconcelos N. Deep learning with low precision by half-wave Gaussian quantization. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI (2017). p. 5406–14.

41. Taylor G, Burmeister R, Xu Z, Singh B, Patel A, Goldstein T. Training neural networks without gradients: a scalable ADMM approach. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, Vol. 48. New York, NY: JMLR.org (2016). p. 2722–31.
42. Carreira-Perpinan M, Wang W. Distributed optimization of deeply nested systems. In: Kaski S, Corander J, editors. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*. Reykjavik: PMLR (2014). p. 10–19. Available online at: http://proceedings.mlr.press/v33/carreira-perpinan14.html
43. Zhang Z, Chen Y, Saligrama V. Efficient training of very deep neural networks for supervised hashing. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV (2016). p. 1487–95.
44. Attouch H, Bolte J, Redont P, Soubeyran A. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Lojasiewicz inequality. *Math Oper Res*. (2010) **35**:438–57. doi: 10.1287/moor.1100.0449
45. Wu T. Variable splitting based method for image restoration with impulse plus Gaussian noise. *Math Probl Eng*. (2016) **2016**:3151303. doi: 10.1155/2016/3151303
46. Wang Y, Zeng J, Yin W. Global convergence of ADMM in nonconvex nonsmooth optimization. *J Sci Comput*. (2019) **78**:29–63. doi: 10.1007/s10915-018-0757-z
47. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. (1998) **86**:2278–324. doi: 10.1109/5.726791
48. Dinh T, Xin J. Convergence of a relaxed variable splitting coarse gradient descent method for learning parse weight binarized activation neural network. *arXiv [preprint] arXiv:1901.09731*. (2019).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.