UNDERSTANDING STRAIGHT-THROUGH ESTIMATOR IN TRAINING ACTIVATION QUANTIZED NEURAL NETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training activation quantized neural networks involves piecewise constant loss functions with the sampled gradient vanishing almost everywhere, which is undesirable for back-propagation. An empirical way around this issue is to use a straight-through estimator (STE) (Bengio et al., 2013) in the backward pass, so that the resulting unusual "gradient" becomes non-trivial. In this paper, we make the first theoretical justification for the concept of STE, by considering the problem of learning a one-hidden-layer convolutional network with binarized ReLU activation and Gaussian input data. We refer to the unusual "gradient" based on STE as coarse gradient, which essentially is not the gradient of any function. Apparently, the choice of STE is not unique. We prove that if the STE is properly chosen, the negative expected coarse gradient is a descent direction for minimizing the population loss, and the associated coarse gradient descent algorithm converges to a local minimum (more rigorously, a critical point) of the population loss minimization problem. Moreover, we show that a relatively poor choice of STE may lead to instability of the training algorithm near certain local minima, which is also validated by our CIFAR-10 experiments.

1 INTRODUCTION

Deep neural networks (DNN) have achieved the remarkable success in many machine learning applications such as computer vision (Krizhevsky et al., 2012; Ren et al., 2015), natural language processing (Collobert & Weston, 2008) and reinforcement learning (Mnih et al., 2015; Silver et al., 2016). However, the deployment of DNN typically require hundreds of megabytes of memory storage for the trainable full-precision floating-point parameters, and billions of floating-point operations to make a single inference. To achieve the compression and acceleration, many recent efforts have been made to the training of quantized DNN, in the hope of maintaining the performance of their float counterparts (Courbariaux et al., 2015; Rastegari et al., 2016; Cai et al., 2017).

Training fully quantized DNN amounts to solving a very challenging optimization problem. It calls for minimizing a piecewise constant and highly nonconvex empirical risk function $f(\boldsymbol{w})$ subject to a discrete set-constraint $\boldsymbol{w} \in \mathcal{Q}$ that characterizes the quantized weights. In particular, weight quantization of DNN have been extensively studied in the literature; see for examples (Li et al., 2016; Zhu et al., 2016; Zhou et al., 2017; Li et al., 2017; Hou & Kwok, 2018). On the other hand, the gradient $\nabla f(\boldsymbol{w})$ in training activation quantized DNN is almost everywhere (a.e.) zero, which makes the standard back-propagation inapplicable. The arguably most effective way around this issue is nothing but to construct a non-trivial search direction by properly modifying the chain rule. Specifically, one can replace the a.e. zero derivative of quantized activation function composited in the chain rule with a related surrogate. This proxy derivative used in the backward pass only is referred as the straight-through estimator (STE) (Bengio et al., 2013).

1.1 RELATED WORKS

The concept of STE was originally introduced in lecture 9c of (Hinton, 2012) for training networks with the hard threshold activation $1_{\{x>0\}}$ (a.k.a. binary neuron). (Hinton, 2012) proposed to simply back-propagate through the hard threshold function as if it had been the identity function. (Bengio et al., 2013) proposed a STE variant which uses the derivative of the sigmoid function instead. In the

training of DNN with weights and activations constrained to ± 1, (Hubara et al., 2016) substituted the derivative of the signum activation function with $1_{\{|x|\leq 1\}}$ in the backward pass. Later the idea of STE was readily extended to the training of DNN with general quantized ReLU activations (Hubara et al., 2018; Zhou et al., 2016; Cai et al., 2017; Choi et al., 2018), where some other proxies took place including the derivatives of vanilla ReLU and clipped ReLU. Despite all the empirical success of STE, to our best knowledge, there is almost no theoretical understanding of why it works.

Similar scenarios, where the derivative of certain layer composited in the loss function is not desirable for back-propagation, have also been brought up recently by (Wang et al., 2018) and (Athalye et al., 2018). The former proposed an implicit weighted nonlocal Laplacian layer as the classifier to improve the generalization accuracy of DNN. In the backward pass, the derivative of a pre-trained fully-connected layer was used as a surrogate. To circumvent the defense in adversarial attack (Szegedy et al., 2013), (Athalye et al., 2018) introduced the so-called backward pass differentiable approximation to deal with the obfuscated gradients, which shares the same spirit as STE. Again, neither of these two papers theoretically justified the proposed training approach.

Another line of research studies the convergence of (stochastic) gradient descent algorithm for learning shallow ReLU nets with one or two linear layers and Gaussian input data. Some works consider the empirical risk minimization with finite input samples (Zhong et al., 2017; Soltanolkotabi, 2017), while some others consider the minimization of population loss averaged over the whole data space (Brutzkus & Globerson, 2017; Tian, 2017; Li & Yuan, 2017; Du et al., 2018). The advantage of using the population loss model is that it admits analytic formulas for both the objective function and gradient, which facilitates the analysis. It is of the common interest to analyze whether (or under what conditions) the (stochastic) gradient descent with the standard back-propagation converges to the global minimum of the regression problem and thus recovers the true weights.

1.2 Main Contributions

Throughout this paper, we shall refer to the resultant composite "gradient" through STE as coarse gradient. The coarse gradient is basically not the gradient of any function. Our key contribution is to provide the *first* theoretical understanding of STE by analyzing the coarse gradient descent algorithm for learning a one-hidden-layer network with binary activation and Gaussian data. We consider two representative STEs: the derivatives of the identity function (Hinton, 2012) and the vanilla ReLU (Cai et al., 2017), and adopt the model of population loss minimization. We derive the explicit form of the expected coarse gradients corresponding to the two STEs, and show that the negative expected coarse gradient based on vanilla ReLU is a *descent direction* for the minimizing the population loss, whereas this is not necessarily true for the one based on the identity function. Moreover, we prove that the former guarantees the convergence to a critical point (a saddle point or a (local) minimizer) and the latter can be unstable sometimes near certain local minima. Indeed, in our experiment on CIFAR-10 using ResNet-20, it is observed that the training algorithm using the identity STE is repelled from good minima and converges to an inferior one with higher training loss and decreased generalization accuracy. This is an implication of the poor performance of the identity STE, not because of the slow convergence of its corresponding coarse gradient descent, instead due to the fact that the algorithm can never reach a good local minimum.

Notations. $\|\cdot\|$ denotes the Euclidean norm of a vector or the spectral norm of a matrix. $\mathbf{0}_n \in \mathbb{R}^n$ represents the vector of all zeros, whereas $\mathbf{1}_n \in \mathbb{R}^n$ the vector of all ones. $\boldsymbol{I}_n$ is the identity matrix of order n. For any $\boldsymbol{w}$, $\mathbf{z} \in \mathbb{R}^n$, $\boldsymbol{w}^\top \mathbf{z} = \langle \boldsymbol{w}, \mathbf{z} \rangle = \sum_i w_i z_i$ is their inner product. $\boldsymbol{w} \odot \mathbf{z}$ denotes the Hadamard product whose ith entry is given by $(\boldsymbol{w} \odot \mathbf{z})_i = w_i z_i$.

2 Learning One-Hidden-Layer CNN with Binary Activation

In this paper, we consider a one-hidden-layer network model (Du et al., 2018) that outputs the prediction

$$y(\mathbf{Z}, \boldsymbol{v}, \boldsymbol{w}) := \sum_{i=1}^{m} v_i \sigma(\mathbf{Z}_i^\top \boldsymbol{w}) = \boldsymbol{v}^\top \sigma(\mathbf{Z}\boldsymbol{w})$$

for some input $\mathbf{Z} \in \mathbb{R}^{m\times n}$. Here $\boldsymbol{w} \in \mathbb{R}^n$ and $\boldsymbol{v} \in \mathbb{R}^m$ are the trainable weights in the first and second linear layer, respectively; $\mathbf{Z}_i^\top$ denotes the ith row vector of $\mathbf{Z}$; the activation function σ acts

component-wise on the vector $\mathbf{Z}\boldsymbol{w}$, i.e., $\sigma(\mathbf{Z}\boldsymbol{w})_i = \sigma((\mathbf{Z}\boldsymbol{w})_i) = \sigma(\mathbf{Z}_i^\top \boldsymbol{w})$. The first layer serves as a convolutional layer, where each row $\mathbf{Z}_i^\top$ can be viewed as a patch sampled from $\mathbf{Z}$ and the weight filter $\boldsymbol{w}$ is shared among all patches, and the second linear layer is the classifier. The label is generated according to $y^*(\mathbf{Z}) = (\boldsymbol{v}^*)^\top \sigma(\mathbf{Z}\boldsymbol{w}^*)$ for some true (non-zero) parameters $\boldsymbol{v}^*$ and $\boldsymbol{w}^*$. Moreover, we use the following squared sample loss

$$\ell(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z}) := \frac{1}{2}\left(y(\mathbf{Z}, \boldsymbol{v}, \boldsymbol{w}) - y^*(\mathbf{Z})\right)^2 = \frac{1}{2}\left(\boldsymbol{v}^\top \sigma(\mathbf{Z}\boldsymbol{w}) - y^*(\mathbf{Z})\right)^2. \tag{1}$$

Unlike in (Du et al., 2018), the activation function σ here is not ReLU, but the binary function $\sigma(x) = 1_{\{x>0\}}$, same as the hard threshold activation (Hinton, 2012).

We assume that the entries of $\mathbf{Z} \in \mathbb{R}^{m\times n}$ are i.i.d. sampled from the Gaussian distribution $\mathcal{N}(0, 1)$ (Zhong et al., 2017; Brutzkus & Globerson, 2017). The legitimacy of this assumption comes from the use of batch normalization (Ioffe & Szegedy, 2015) in most architectures, which sends normalized inputs to the linear layers. Since $\ell(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z}) = \ell(\boldsymbol{v}, \boldsymbol{w}/c; \mathbf{Z})$ for any scalar $c > 0$, without loss of generality, we take $\|\boldsymbol{w}^*\| = 1$ and cast the learning task as the following population loss minimization problem:

$$\min_{\boldsymbol{v}\in\mathbb{R}^m, \boldsymbol{w}\in\mathbb{R}^n} \; f(\boldsymbol{v}, \boldsymbol{w}) := \mathbb{E}_{\mathbf{Z}}\left[\ell(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right], \tag{2}$$

where the sample loss $\ell(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})$ is given by (1).

2.1 Back-propagation and Coarse Gradient Descent

With the Gaussian assumption on $\mathbf{Z}$, as will be shown in section 2.2, it is possible to find the analytic expressions of $f(\boldsymbol{v}, \boldsymbol{w})$ and its gradient

$$\nabla f(\boldsymbol{v}, \boldsymbol{w}) := \begin{bmatrix} \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}) \\ \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) \end{bmatrix}.$$

The information about $\nabla f(\boldsymbol{v}, \boldsymbol{w})$, however, is not available for the network training. In fact, we can only use the expectation of the sample gradient, namely,

$$\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right] \text{ and } \mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right].$$

We remark that the expected partial gradient $\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ is not the same as $\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) = \frac{\partial \mathbb{E}_{\mathbf{Z}}[\ell(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})]}{\partial \boldsymbol{w}}$. By the standard back-propagation or chain rule, we readily check that

$$\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z}) = \sigma(\mathbf{Z}\boldsymbol{w})\Big(\boldsymbol{v}^\top \sigma(\mathbf{Z}\boldsymbol{w}) - y^*(\mathbf{Z})\Big) \tag{3}$$

and

$$\frac{\partial \ell}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z}) = \mathbf{Z}^\top\big(\sigma'(\mathbf{Z}\boldsymbol{w}) \odot \boldsymbol{v}\big)\Big(\boldsymbol{v}^\top \sigma(\mathbf{Z}\boldsymbol{w}) - y^*(\mathbf{Z})\Big). \tag{4}$$

Note that σ' is zero a.e., which makes (4) inapplicable to the training. The idea of STE is to simply replace the a.e. zero component σ' in (4) with a related non-trivial function μ' (Hinton, 2012; Bengio et al., 2013; Hubara et al., 2016; Cai et al., 2017), which is the derivative of some (sub)differentiable function μ. More precisely, back-propagation using the STE μ' gives the following non-trivial surrogate of $\frac{\partial \ell}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})$, to which we refer as the coarse (partial) gradient

$$\boldsymbol{g}_\mu(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z}) = \mathbf{Z}^\top\big(\mu'(\mathbf{Z}\boldsymbol{w}) \odot \boldsymbol{v}\big)\Big(\boldsymbol{v}^\top \sigma(\mathbf{Z}\boldsymbol{w}) - y^*(\mathbf{Z})\Big). \tag{5}$$

Using the STE μ' to train the one-hidden-layer convolutional neural network (CNN) with binary activation gives rise to the (full-batch) coarse gradient descent described in Algorithm 1.

2.2 Preliminaries

Let us present some preliminaries about the landscape of the population loss function $f(\boldsymbol{v}, \boldsymbol{w})$. To this end, we define the angle between $\boldsymbol{w}$ and $\boldsymbol{w}^*$ as $\theta(\boldsymbol{w}, \boldsymbol{w}^*) := \arccos\left(\frac{\boldsymbol{w}^\top \boldsymbol{w}^*}{\|\boldsymbol{w}\|\|\boldsymbol{w}^*\|}\right)$ for any $\boldsymbol{w} \neq \mathbf{0}_n$. Recall that the label is given by $y^*(\mathbf{Z}) = (\boldsymbol{v}^*)^\top \mathbf{Z}\boldsymbol{w}^*$, we elaborate on the expressions of $f(\boldsymbol{v}, \boldsymbol{w})$ and $\nabla f(\boldsymbol{v}, \boldsymbol{w})$.

Algorithm 1 Coarse gradient descent for learning one-hidden-layer CNN with STE μ'.

Input: initialization $\boldsymbol{v}^0 \in \mathbb{R}^m$, $\boldsymbol{w}^0 \in \mathbb{R}^n$, learning rate η.

for $t = 1, 2, \ldots$ **do**

$\quad \boldsymbol{v}^{t+1} = \boldsymbol{v}^t - \eta \mathbb{E}_{\mathbf{Z}} \left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z}) \right]$

$\quad \boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \mathbb{E}_{\mathbf{Z}} \left[\boldsymbol{g}_\mu(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z}) \right]$

end for

Lemma 1. *Suppose all entries of* $\mathbf{Z}$ *follow i.i.d.* $\mathcal{N}(0,1)$. *If* $\boldsymbol{w} \neq \mathbf{0}_n$, *the population loss* $f(\boldsymbol{v}, \boldsymbol{w})$ *is given by*

$$\frac{1}{8}\left[\boldsymbol{v}^\top \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)\boldsymbol{v} - 2\boldsymbol{v}^\top \left(\left(1 - \frac{2}{\pi}\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right)\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)\boldsymbol{v}^* + (\boldsymbol{v}^*)^\top \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)\boldsymbol{v}^*\right].$$

In addition, $f(\boldsymbol{v}, \boldsymbol{w}) = \frac{1}{8}(\boldsymbol{v}^*)^\top \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)\boldsymbol{v}^*$ *for* $\boldsymbol{w} = \mathbf{0}_n$.

All the technical proofs in this paper will be detailed in the appendix.

Lemma 2. *If* $\boldsymbol{w} \neq \mathbf{0}_n$ *and* $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \in (0, \pi)$, *the partial gradients of* $f(\boldsymbol{v}, \boldsymbol{w})$ *w.r.t.* $\boldsymbol{v}$ *and* $\boldsymbol{w}$ *are*

$$\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}) = \frac{1}{4}\left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)\boldsymbol{v} - \frac{1}{4}\left(\left(1 - \frac{2}{\pi}\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right)\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)\boldsymbol{v}^* \tag{6}$$

and

$$\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) = -\frac{\boldsymbol{v}^\top \boldsymbol{v}^*}{2\pi\|\boldsymbol{w}\|} \frac{\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|}, \tag{7}$$

respectively.

For any $\boldsymbol{v} \in \mathbb{R}^m$, $(\boldsymbol{v}, \mathbf{0}_m)$ is impossible to be a local minimizer. The only possible (local) minimizers of the model (2) are located at

1. Stationary points where the gradients given by (6) and (7) vanish simultaneously (which may not be possible), i.e.,

$$\boldsymbol{v}^\top \boldsymbol{v}^* = 0 \text{ and } \boldsymbol{v} = \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)^{-1}\left(\left(1 - \frac{2}{\pi}\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right)\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)\boldsymbol{v}^*. \tag{8}$$

2. Non-differentiable points where $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = 0$ and $\boldsymbol{v} = \boldsymbol{v}^*$, or $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi$ and $\boldsymbol{v} = \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$.

Among them, $\{(\boldsymbol{v}, \boldsymbol{w}) : \boldsymbol{v} = \boldsymbol{v}^*, \theta(\boldsymbol{w}, \boldsymbol{w}^*) = 0\}$ are obviously the global minimizers of (2). We show that the stationary points, if exist, can only be saddle points, and $\{(\boldsymbol{v}, \boldsymbol{w}) : \theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi,\ \boldsymbol{v} = \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*\}$ are the only potential spurious local minimizers.

Proposition 1. *If the true parameter* $\boldsymbol{v}^*$ *satisfies* $(\mathbf{1}_m^\top \boldsymbol{v}^*)^2 < \frac{m+1}{2}\|\boldsymbol{v}^*\|^2$, *then*

$$\left\{(\boldsymbol{v}, \boldsymbol{w}) : \boldsymbol{v} = (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1}\left(\frac{-(\mathbf{1}_m^\top \boldsymbol{v}^*)^2}{(m+1)\|\boldsymbol{v}^*\|^2 - (\mathbf{1}_m^\top \boldsymbol{v}^*)^2}\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)\boldsymbol{v}^*, \right.$$
$$\left. \theta(\boldsymbol{w}, \boldsymbol{w}^*) = \frac{\pi}{2}\frac{(m+1)\|\boldsymbol{v}^*\|^2}{(m+1)\|\boldsymbol{v}^*\|^2 - (\mathbf{1}_m^\top \boldsymbol{v}^*)^2}\right\} \tag{9}$$

give the saddle points obeying (8), and $\{(\boldsymbol{v}, \boldsymbol{w}) : \theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi,\ \boldsymbol{v} = \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*\}$ *are the spurious local minimizers. Otherwise, the model (2) has no saddle points or spurious local minimizers.*

We further prove that the underlying true gradient $\nabla f(\boldsymbol{v}, \boldsymbol{w})$ given by (6) and (7), is Lipschitz continuous under a boundedness condition.

Lemma 3. *For any differentiable points* $(\boldsymbol{v}, \boldsymbol{w})$ *and* $(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})$ *with* $\min\{\|\boldsymbol{w}\|, \|\tilde{\boldsymbol{w}}\|\} = c_{\boldsymbol{w}} > 0$ *and* $\max\{\|\boldsymbol{v}\|, \|\tilde{\boldsymbol{v}}\|\} = C_{\boldsymbol{v}}$, *there exists a Lipschitz constant* $L > 0$ *depending on* $C_{\boldsymbol{v}}$ *and* $c_{\boldsymbol{w}}$, *such that*

$$\|\nabla f(\boldsymbol{v}, \boldsymbol{w}) - \nabla f(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})\| \leq L\|(\boldsymbol{v}, \boldsymbol{w}) - (\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})\|.$$

3 MAIN RESULTS

We are most interested in the complex case where both the saddle points and spurious local minimizers are present. Our main results are concerned with the behaviors of the coarse gradient descent summarized in Algorithm 1 when the derivatives of the vanilla ReLU and identity function serve as the STE, respectively. Intuitively, the ReLU STE is supposed to outperform the identity STE, because ReLU is obviously a better approximation to the activation function $\sigma(x) = 1_{\{x>0\}}$.

Theorem 1. *Let $\{(\boldsymbol{v}^t, \boldsymbol{w}^t)\}$ be the sequence generated by Algorithm 1 with ReLU $\mu(x) = \max\{x, 0\}$. Suppose $\|\boldsymbol{v}^t\| \leq C_{\boldsymbol{v}}$ and $\|\boldsymbol{w}^t\| \geq c_{\boldsymbol{w}}$ for all t with some $C_{\boldsymbol{v}}, c_{\boldsymbol{w}} > 0$. Then if the learning rate $\eta > 0$ is sufficiently small, for any initialization $(\boldsymbol{v}^0, \boldsymbol{w}^0)$, the objective sequence $\{f(\boldsymbol{v}^t, \boldsymbol{w}^t)\}$ is monotonically decreasing, and $\{(\boldsymbol{v}^t, \boldsymbol{w}^t)\}$ converges to a saddle point or a (local) minimizer of the population loss minimization (2). In addition, if $\mathbf{1}_m^\top \boldsymbol{v}^* \neq 0$ and $m > 1$, the descent and convergence properties do not hold for Algorithm 1 with the identity function $\mu(x) = x$ near the local minimizers satisfying $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi$ and $\boldsymbol{v} = (\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$.*

Remark 1. *The convergence guarantee for the coarse gradient descent using vanilla ReLU is established under the assumption that there are infinite training samples. When there are only a few data, in a coarse scale, the empirical loss is roughly descending along the direction of negative coarse gradient, as illustrated by Figure 1. As the sample size increases, the empirical loss gains monotonicity and smoothness. This is different from the conventional gradient descent widely studied in the existing literature, which enjoys the descent property regardless of the sample size.*

In the rest of this section, we sketch the mathematical analysis for the main results.

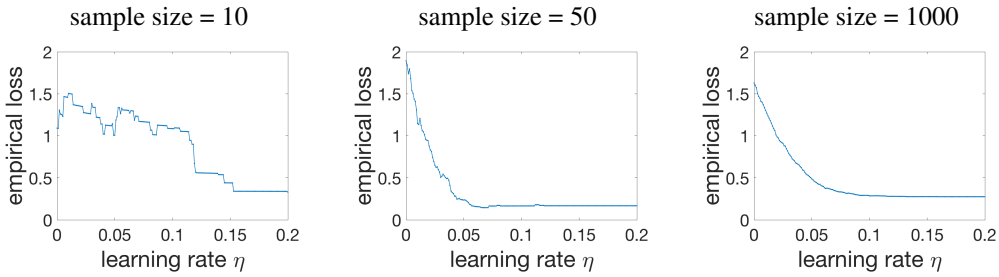

Figure 1: The plots of the empirical loss moving by one step in the direction of negative coarse gradient v.s. the learning rate (step size) η for different sample sizes.

3.1 DERIVATIVE OF THE VANILLA RELU AS STE

If we choose the derivative of ReLU $\mu(x) = \max\{x, 0\}$ as the STE in (5), it is easy to see $\mu'(x) = \sigma(x)$, and we have the following expressions of $\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ and $\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ for Algorithm 1.

Lemma 4. *The expected partial gradient of $\ell(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})$ w.r.t. $\boldsymbol{v}$ is*

$$\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right] = \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}). \tag{10}$$

Let $\mu(x) = \max\{x, 0\}$ in (5). The expected coarse gradient w.r.t. $\boldsymbol{w}$ is

$$\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right] = \frac{h(\boldsymbol{v}, \boldsymbol{v}^*)}{2\sqrt{2\pi}}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - \cos\left(\frac{\theta(\boldsymbol{w}, \boldsymbol{w}^*)}{2}\right)\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{\sqrt{2\pi}}\frac{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \boldsymbol{w}^*}{\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \boldsymbol{w}^*\right\|}, \tag{11}$$

[1]

where $h(\boldsymbol{v}, \boldsymbol{v}^*) = \|\boldsymbol{v}\|^2 + (\mathbf{1}_m^\top\boldsymbol{v})^2 - (\mathbf{1}_m^\top\boldsymbol{v})(\mathbf{1}_m^\top\boldsymbol{v}^*) + \boldsymbol{v}^\top\boldsymbol{v}^*$.

The key observation is that the coarse partial gradient $\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ has non-negative correlation with the true gradient $\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w})$, and $-\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ together with $-\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ form a descent direction for minimizing the population loss.

[1] We redefine the second term as $\mathbf{0}_n$ in the case $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi$, or equivalently, $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \boldsymbol{w}^* = \mathbf{0}_n$.

Lemma 5. *If $\boldsymbol{w} \neq \mathbf{0}_n$ and $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \in (0, \pi)$, then the inner product between the expected coarse and true gradients w.r.t. $\boldsymbol{w}$ is*

$$\left\langle \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) \right\rangle = \frac{\sin\left(\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right)}{2(\sqrt{2\pi})^3 \|\boldsymbol{w}\|} (\boldsymbol{v}^\top \boldsymbol{v}^*)^2 \geq 0.$$

Moreover, if further $\|\boldsymbol{v}\| \leq C_{\boldsymbol{v}}$, there exists a constant $A > 0$ depending only on $C_{\boldsymbol{v}}$, such that

$$\left\| \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right] \right\|^2 \leq A \left(\left\| \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}) \right\|^2 + \left\langle \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) \right\rangle \right). \tag{12}$$

Clearly, when $\left\langle \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) \right\rangle > 0$, $\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ is roughly in the same direction as $\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w})$. Moreover, since by Lemma 4, $\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right] = \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w})$, we expect that the coarse gradient descent behaves like the gradient descent directly on $f(\boldsymbol{v}, \boldsymbol{w})$. We would like to highlight the significance of the estimate (12) in guaranteeing the descent property of Algorithm 1. By the Lipschitz continuity of ∇f specified in Lemma 3, it holds that

$$\begin{aligned}
f(\boldsymbol{v}^{t+1}, \boldsymbol{w}^{t+1}) - f(\boldsymbol{v}^t, \boldsymbol{w}^t) \leq & \left\langle \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \boldsymbol{v}^{t+1} - \boldsymbol{v}^t \right\rangle + \left\langle \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \boldsymbol{w}^{t+1} - \boldsymbol{w}^t \right\rangle \\
& + \frac{L}{2}(\|\boldsymbol{v}^{t+1} - \boldsymbol{v}^t\|^2 + \|\boldsymbol{w}^{t+1} - \boldsymbol{w}^t\|^2) \\
= & -\left(\eta - \frac{L\eta^2}{2}\right) \left\| \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t, \boldsymbol{w}^t) \right\|^2 + \frac{L\eta^2}{2} \left\| \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z})\right] \right\|^2 \\
& - \eta \left\langle \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z})\right] \right\rangle \\
\overset{a)}{\leq} & -\left(\eta - (1+A)\frac{L\eta^2}{2}\right) \left\| \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t, \boldsymbol{w}^t) \right\|^2 \\
& - \left(\eta - \frac{AL\eta^2}{2}\right) \left\langle \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z})\right] \right\rangle,
\end{aligned} \tag{13}$$

where a) is due to (12). Therefore, if η is small enough, we have monotonically decreasing population loss until convergence.

Lemma 6. *When Algorithm 1 converges, $\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ and $\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ vanish simultaneously, which only occurs at the*

1. *Saddle points where (8) is satisfied according to Proposition 1.*
2. *Minimizers of (2) where $\boldsymbol{v} = \boldsymbol{v}^*$, $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = 0$, or $\boldsymbol{v} = (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$, $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi$.*

3.2 Derivative of the Identity Function as STE

Now we consider the derivative of identity function. As opposed to the ReLU case, similar results to Lemmas 5 and 6 are not valid anymore. It happens that the coarse gradient derived from the identity STE does not vanish at the local minimum, and Algorithm 1 may never converge there.

Lemma 7. *Let $\mu(x) = x$ in (5). Then the expected coarse partial gradient w.r.t. $\boldsymbol{w}$ is*

$$\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right] = \frac{1}{\sqrt{2\pi}} \left(\|\boldsymbol{v}\|^2 \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - (\boldsymbol{v}^\top \boldsymbol{v}^*)\boldsymbol{w}^* \right). \tag{14}$$

If $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi$ and $\boldsymbol{v} = \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$,

$$\left\| \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right] \right\| = \frac{2(m-1)}{\sqrt{2\pi}(m+1)^2} (\mathbf{1}_m^\top \boldsymbol{v}^*)^2 \geq 0,$$

i.e., $\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ does not vanish at the spurious local minimizers if $\mathbf{1}_m^\top \boldsymbol{v}^* \neq 0$ and $m > 1$.

Lemma 8. *If $\boldsymbol{w} \neq \mathbf{0}_n$ and $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \in (0, \pi)$, then the inner product between the expected coarse and true gradients w.r.t. $\boldsymbol{w}$ is*

$$\left\langle \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) \right\rangle = \frac{\sin\left(\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right)}{(\sqrt{2\pi})^3 \|\boldsymbol{w}\|} (\boldsymbol{v}^\top \boldsymbol{v}^*)^2 \geq 0. \tag{15}$$

When $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \to \pi$, $\boldsymbol{v} \to \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)^{-1} (\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m) \boldsymbol{v}^*$, if $\mathbf{1}_m^\top \boldsymbol{v}^* \neq 0$ and $m > 1$, we have

$$\frac{\left\| \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right] \right\|^2}{\left\| \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}) \right\|^2 + \left\langle \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) \right\rangle} \to +\infty. \tag{16}$$

Lemma 7 suggests that if $\mathbf{1}_m^\top \boldsymbol{v}^* \neq 0$, the coarse gradient descent will never converge near the spurious minimizers with $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi$ and $\boldsymbol{v} = \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)^{-1} (\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m) \boldsymbol{v}^*$, because $\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ does not vanish there. By the positive correlation implied by (15), for some proper $(\boldsymbol{v}^0, \boldsymbol{w}^0)$, the iterates $\{(\boldsymbol{v}^t, \boldsymbol{w}^t)\}$ may move towards a spurious local minimizer in the beginning. But when $\{(\boldsymbol{v}^t, \boldsymbol{w}^t)\}$ approaches it, the correlation in (15) goes to 0, and the descent property (13) does not hold with $\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]$ because of (16), hence the instability arises.

4 Experiments

In practice, the vanilla ReLU may not deliver the best empirical performance. Compared with the identity function and vanilla ReLU, the clipped ReLU proposed in (Cai et al., 2017) approximates the quantized ReLU better. The plots of the 2-bit quantized ReLU and its associated clipped ReLU are in Figure 3 in the appendix. Besides the two STEs discussed above, we include the STE using derivative of the clipped ReLU for comparisons on training 2-bit and 4-bit activation networks for MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009) classifications. Log-tailed ReLU has also been proposed, we do not consider it here since it gives similar performance to the clipped ReLU as reported in (Cai et al., 2017). We emphasize that we are not claiming the superiority of the quantization approach used here, as it is nothing but the HWGQ (Cai et al., 2017), except we consider the uniform quantization. In all of our experiments, the weights are kept float.

The optimizer we use is the stochastic (coarse) gradient descent with momentum = 0.9 for all experiments. We train 50 epochs for LeNet-5 (LeCun et al., 1998) on MNIST, and 200 epochs for VGG-11 (Simonyan & Zisserman, 2014) and ResNet-20 (He et al., 2016) on CIFAR-10. The parameters/weights are initialized with those from their pre-trained full-precision counterparts. The schedule of the learning rate is specified in Table 2 in the appendix.

The resolution α for the quantized ReLU needs to be carefully chosen to maintain the full-precision level accuracy. To this end, we follow (Cai et al., 2017) and resort to a modified batch normalization layer (Ioffe & Szegedy, 2015) without the scale and shift, whose output components approximately follow a Gaussian distribution. Then the α that fits the layer input the best can be pre-computed by a variant of Lloyd's algorithm (Lloyd, 1982; Yin et al., 2018) applied to a set of simulated 1-D half-Gaussian data. After determining the α, it will be fixed during the whole training process. Since the original LeNet-5 does not have batch normalization, we add one prior to each activation layer.

4.1 Comparison Results

The experimental results are summarized in Table 1, where we record both the training losses and validation accuracies. Among the three STEs, the derivative of clipped ReLU gives the best overall performance, followed by vanilla ReLU and then by the identity function. For deeper networks, clipped ReLU is the best performer. But on the shallow LeNet-5 network, vanilla ReLU exhibits comparable performance to clipped ReLU, which is in line with our theoretical finding that ReLU is a superior STE for learning the one-hidden-layer (shallow) CNN.

4.2 Repelled From Improved Minima

We report the phenomenon of being repelled from a good minimum on ResNet-20 with 4-bit activations when using the identity STE. By Table 1, the coarse gradient descent algorithms using the

	Network	BitWidth	Straight-through estimator		
			identity	vanilla ReLU	clipped ReLU
MNIST	LeNet5	2	2.6×10^{-2}/98.49	5.1×10^{-3}/99.24	5.4×10^{-3}/99.23
		4	6.0×10^{-3}/98.98	9.0×10^{-4}/99.32	8.8×10^{-4}/99.24
CIFAR10	VGG11	2	0.19/86.58	0.10/88.69	0.02/90.92
		4	3.1×10^{-2}/90.19	1.5×10^{-3}/92.01	1.3×10^{-3}/92.08
	ResNet20	2	1.56/46.52	1.50/48.05	0.24/88.39
		4	1.38/54.16	0.25/86.59	0.04/91.24

Table 1: Training loss/validation accuracy (%) on MNIST and CIFAR-10 with quantized activations and float weights, for STEs using derivatives of the identity function, vanilla ReLU and clipped ReLU at bit-widths 2 and 4.

vanilla and clipped ReLUs converge to the neighborhoods of the minima with validation accuracies (training losses) of 86.59% (0.25) and 91.24% (0.04), respectively, whereas that using the identity STE gives 54.16% (1.38). Note that the landscape of the empirical loss function does not depend on which STE is used in the training. Then we initialize training with the two improved minima and use the identity STE. To see if the algorithm is stable there, we start the training with a tiny learning rate of 10^{-5}. For both initializations, the training loss and validation error significantly increase within the first 20 epochs. At epoch 20, we switch to the normal schedule of learning rate and run 200 additional epochs. The training using the identity STE ends up with a much worse minimum. This is because the coarse gradient with identity STE does not vanish at the good minima in this case.

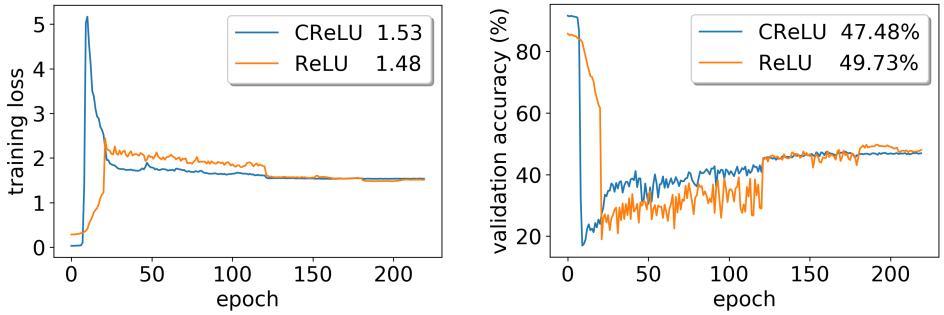

Figure 2: When initialized with the improved minima produced by the vanilla (orange) and clipped (blue) ReLUs on ResNet-20 with 4-bit activations and float weights, the coarse gradient descent using the identity STE ends up being repelled from there. The learning rate is set to 10^{-5} until epoch 20.

5 Concluding Remarks

We provided the first theoretical justification for the concept of STE. We considered two STEs: the derivatives of the identity function and the vanilla ReLU, in training one-hidden-layer CNN with binary activation. We derived the explicit formulas of the expected coarse gradients corresponding to the two STEs, and showed that the negative expected coarse gradient based on vanilla ReLU is a descent direction for minimizing the population loss, whereas the identity STE is not. Our experiments on MNIST and CIFAR-10 datasets verified the theoretical findings. Looking ahead, we believe that the closeness between the quantized ReLU and the anti-derivative of STE plays an essential role in the performance of the coarse gradient descent. Hence, we would like to explore and quantify this relationship in the future work.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Alon Brutzkus and Amir Globerson. Globally optimal gradient descent for a convnet with gaussian inputs. *arXiv preprint arXiv:1702.07966*, 2017.

Zhaowei Cai, Xiaodong He, Jian Sun, and Nuno Vasconcelos. Deep learning with low precision by half-wave gaussian quantization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*, 2018.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *International Conference on Machine Learning*, pp. 160–167. ACM, 2008.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Advances in neural information processing systems*, pp. 3123–3131, 2015.

Simon S. Du, Jason D. Lee, Yuandong Tian, Barnabas Poczos, and Aarti Singh. Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minimum. *arXiv preprint arXiv:1712.00779*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Geoffrey Hinton. Neural networks for machine learning, coursera. *Coursera, video lectures*, 2012.

Lu Hou and James T Kwok. Loss-aware weight quantization of deep networks. *arXiv preprint arXiv:1802.08635*, 2018.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks: Training neural networks with weights and activations constrained to +1 or -1. *arXiv preprint arXiv:1602.02830*, 2016.

Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *Journal of Machine Learning Research*, 18:1–30, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Tech Report*, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

Hao Li, Soham De, Zheng Xu, Christoph Studer, Hanan Samet, and Tom Goldstein. Training quantized nets: A deeper understanding. In *Advances in Neural Information Processing Systems*, pp. 5811–5821, 2017.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pp. 597–607, 2017.

Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Trans. Info. Theory*, 28:129–137, 1982.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision*, pp. 525–542. Springer, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Mahdi Soltanolkotabi. Learning relus via gradient descent. In *Advances in Neural Information Processing Systems*, pp. 2007–2017, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. *arXiv preprint arXiv:1703.00560*, 2017.

Bao Wang, Xiyang Luo, Zhen Li, Wei Zhu, Zuoqiang Shi, and Stanley J Osher. Deep neural nets with interpolating function as output activation. In *Advances in Neural Information Processing Systems*, 2018.

Penghang Yin, Shuai Zhang, Jiancheng Lyu, Stanley Osher, Yingyong Qi, and Jack Xin. Binaryrelax: A relaxation approach for training deep neural networks with quantized weights. *arXiv preprint arXiv:1801.06313*, 2018.

Kai Zhong, Zhao Song, Prateek Jain, Peter L Bartlett, and Inderjit S Dhillon. Recovery guarantees for one-hidden-layer neural networks. *arXiv preprint arXiv:1706.03175*, 2017.

Aojun Zhou, Anbang Yao, Yiwen Guo, Lin Xu, and Yurong Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. *arXiv preprint arXiv:1702.03044*, 2017.

Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.

Chenzhuo Zhu, Song Han, Huizi Mao, and William J Dally. Trained ternary quantization. *arXiv preprint arXiv:1612.01064*, 2016.

APPENDIX

A. THE PLOTS OF QUANTIZED AND CLIPPED RELUS

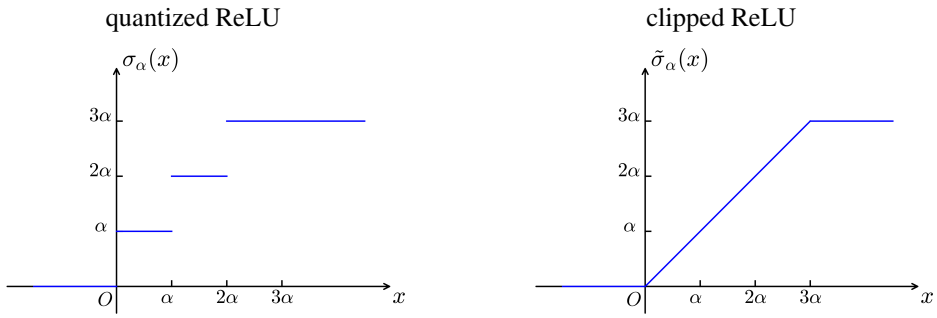

Figure 3: The plots of 2-bit quantized ReLU $\sigma_\alpha(x)$ (with $2^2 = 4$ quantization levels including 0) and the associated clipped ReLU $\tilde{\sigma}_\alpha(x)$. α is the resolution determined in advance of the network training.

B. THE SCHEDULE OF LEARNING RATE

Network	Epoch	Learning rate		
		initial	decay rate	milestone
LeNet5	50	0.1	0.1	[20,40]
VGG11	200	0.01	0.1	[80,140]
ResNet20	200	0.01	0.1	[80,140]

Table 2: The schedule of the learning rate.

C. ADDITIONAL SUPPORTING LEMMAS

Lemma 9. *Let $\mathbf{z} \in \mathbb{R}^n$ be a Gaussian random vector with entries i.i.d. sampled from $\mathcal{N}(0,1)$. Given nonzero vectors $\boldsymbol{w}$, $\tilde{\boldsymbol{w}} \in \mathbb{R}^n$ with the angle θ, we have*

$$\mathbb{E}\left[1_{\{\mathbf{z}^\top \boldsymbol{w} > 0\}}\right] = \frac{1}{2},\ \mathbb{E}\left[1_{\{\mathbf{z}^\top \boldsymbol{w} > 0,\, \mathbf{z}^\top \tilde{\boldsymbol{w}} > 0\}}\right] = \frac{\pi - \theta}{2\pi},$$

and

$$\mathbb{E}\left[\mathbf{z} 1_{\{\mathbf{z}^\top \boldsymbol{w} > 0\}}\right] = \frac{1}{\sqrt{2\pi}} \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|},\ \mathbb{E}\left[\mathbf{z} 1_{\{\mathbf{z}^\top \boldsymbol{w} > 0,\, \mathbf{z}^\top \boldsymbol{w}^* > 0\}}\right] = \frac{\cos(\theta/2)}{\sqrt{2\pi}} \frac{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}}{\left\| \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|} \right\|}.^{2}$$

Proof of Lemma 9. The third identity was proved in Lemma A.1 of (Du et al., 2018). To show the first one, without loss of generality we assume $\boldsymbol{w} = [w_1, \mathbf{0}_{n-1}^\top]^\top$ with $w_1 > 0$, then $\mathbb{E}\left[1_{\{\mathbf{z}^\top \boldsymbol{w} > 0\}}\right] = \mathbb{P}(\mathrm{z}_1 > 0) = \frac{1}{2}$.

We further assume $\tilde{\boldsymbol{w}} = [\tilde{w}_1, \tilde{w}_2, \mathbf{0}_{n-2}^\top]^\top$. It is easy to see that

$$\mathbb{E}\left[1_{\{\mathbf{z}^\top \boldsymbol{w} > 0,\, \mathbf{z}^\top \tilde{\boldsymbol{w}} > 0\}}\right] = \mathbb{P}(\mathbf{z}^\top \boldsymbol{w} > 0,\, \mathbf{z}^\top \tilde{\boldsymbol{w}} > 0) = \frac{\pi - \theta}{2\pi}.$$

To prove the last identity, we use polar representation of two-dimensional Gaussian random variables, where r is the radius and ϕ is the angle with $\mathrm{d}\mathbb{P}_r = r \exp(-r^2/2)\mathrm{d}r$ and $\mathrm{d}\mathbb{P}_\phi = \frac{1}{2\pi}\mathrm{d}\phi$. Then

[2] Same as in Lemma 4, we redefine $\mathbb{E}\left[\mathbf{z} 1_{\{\mathbf{z}^\top \boldsymbol{w} > 0,\, \mathbf{z}^\top \boldsymbol{w}^* > 0\}}\right] = \mathbf{0}_n$ in the case $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi$.

$\mathbb{E}\left[\mathrm{z}_i 1_{\{\mathbf{z}^\top \boldsymbol{w}>0,\, \mathbf{z}^\top \boldsymbol{w}^*>0\}}\right] = 0$ for $i \geq 3$. Moreover,

$$\mathbb{E}\left[\mathrm{z}_1 1_{\{\mathbf{z}^\top \boldsymbol{w}>0,\, \mathbf{z}^\top \boldsymbol{w}^*>0\}}\right] = \frac{1}{2\pi}\int_0^\infty r^2 \exp\left(-\frac{r^2}{2}\right)\mathrm{d}r \int_{-\frac{\pi}{2}+\theta}^{\frac{\pi}{2}} \cos(\phi)\mathrm{d}\phi = \frac{1+\cos(\theta)}{2\sqrt{2\pi}}$$

and

$$\mathbb{E}\left[\mathrm{z}_2 1_{\{\mathbf{z}^\top \boldsymbol{w}>0,\, \mathbf{z}^\top \boldsymbol{w}^*>0\}}\right] = \frac{1}{2\pi}\int_0^\infty r^2 \exp\left(-\frac{r^2}{2}\right)\mathrm{d}r \int_{-\frac{\pi}{2}+\theta}^{\frac{\pi}{2}} \sin(\phi)\mathrm{d}\phi = \frac{\sin(\theta)}{2\sqrt{2\pi}}.$$

Therefore,

$$\mathbb{E}\left[\mathbf{z} 1_{\{\mathbf{z}^\top \boldsymbol{w}>0,\, \mathbf{z}^\top \boldsymbol{w}^*>0\}}\right] = \frac{\cos(\theta/2)}{\sqrt{2\pi}}[\cos(\theta/2), \sin(\theta/2), \mathbf{0}_{n-2}^\top]^\top = \frac{\cos(\theta/2)}{\sqrt{2\pi}} \frac{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}}{\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}\right\|},$$

where the last equality holds because $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$ and $\frac{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}}{\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}\right\|}$ are two unit-normed vectors with angle $\theta/2$. □

Lemma 10. *For any nonzero vectors $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$ with $\|\tilde{\boldsymbol{w}}\| \geq \|\boldsymbol{w}\| = c > 0$, we have*

1. $|\theta(\boldsymbol{w}, \boldsymbol{w}^*) - \theta(\tilde{\boldsymbol{w}}, \boldsymbol{w}^*)| \leq \frac{\pi}{2c}\|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|$.
2. $\left\| \frac{1}{\|\boldsymbol{w}\|} \frac{\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|} - \frac{1}{\|\tilde{\boldsymbol{w}}\|} \frac{\left(\boldsymbol{I}_n - \frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*\right\|} \right\| \leq \frac{1}{c^2}\|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|$.

Proof of Lemma 10. 1. Since by Cauchy-Schwarz inequality,

$$\left\langle \tilde{\boldsymbol{w}}, \boldsymbol{w} - \frac{c\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|} \right\rangle = \tilde{\boldsymbol{w}}^\top \boldsymbol{w} - c\|\tilde{\boldsymbol{w}}\| \leq 0,$$

we have

$$\begin{aligned}\|\tilde{\boldsymbol{w}} - \boldsymbol{w}\|^2 &= \left\|\left(1 - \frac{c}{\|\tilde{\boldsymbol{w}}\|}\right)\tilde{\boldsymbol{w}} - \left(\boldsymbol{w} - \frac{c\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}\right)\right\|^2 \geq \left\|\left(1 - \frac{c}{\|\tilde{\boldsymbol{w}}\|}\right)\tilde{\boldsymbol{w}}\right\|^2 + \left\|\boldsymbol{w} - \frac{c\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}\right\|^2 \\ &\geq \left\|\boldsymbol{w} - \frac{c\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}\right\|^2 = c^2\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}\right\|^2. \end{aligned} \tag{17}$$

Therefore,

$$\begin{aligned}|\theta(\boldsymbol{w}, \boldsymbol{w}^*) - \theta(\tilde{\boldsymbol{w}}, \boldsymbol{w}^*)| &\leq \theta(\boldsymbol{w}, \tilde{\boldsymbol{w}}) = \theta\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}\right) \\ &\leq \pi \sin\left(\frac{\theta\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}\right)}{2}\right) = \frac{\pi}{2}\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|}\right\| \leq \frac{\pi}{2c}\|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|,\end{aligned}$$

where we used the fact $\sin(x) \geq \frac{2x}{\pi}$ for $x \in [0, \frac{\pi}{2}]$ and the estimate in (17).

2. Since $\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*$ is the projection of $\boldsymbol{w}^*$ onto the complement space of $\boldsymbol{w}$, and likewise for $\left(\boldsymbol{I}_n - \frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*$, the angle between $\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*$ and $\left(\boldsymbol{I}_n - \frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*$ is equal to the angle between $\boldsymbol{w}$ and $\tilde{\boldsymbol{w}}$. Therefore,

$$\left\langle \frac{\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|}, \frac{\left(\boldsymbol{I}_n - \frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*\right\|} \right\rangle = \left\langle \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}, \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|} \right\rangle,$$

and thus

$$\left\| \frac{1}{\|\boldsymbol{w}\|} \frac{\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|} - \frac{1}{\|\tilde{\boldsymbol{w}}\|} \frac{\left(\boldsymbol{I}_n - \frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*\right\|} \right\|$$
$$= \left\| \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|^2} - \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|^2} \right\| = \frac{\|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|}{\|\boldsymbol{w}\|\|\tilde{\boldsymbol{w}}\|} \leq \frac{1}{c^2}\|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|.$$

The second equality above holds because

$$\left\| \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|^2} - \frac{\tilde{\boldsymbol{w}}}{\|\tilde{\boldsymbol{w}}\|^2} \right\|^2 = \frac{1}{\|\boldsymbol{w}\|^2} + \frac{1}{\|\tilde{\boldsymbol{w}}\|^2} - \frac{2\langle \boldsymbol{w}, \tilde{\boldsymbol{w}}\rangle}{\|\boldsymbol{w}\|^2\|\tilde{\boldsymbol{w}}\|^2} = \frac{\|\boldsymbol{w} - \tilde{\boldsymbol{w}}\|^2}{\|\boldsymbol{w}\|^2\|\tilde{\boldsymbol{w}}\|^2}.$$

□

D. Main Proofs

Lemma 1. *Suppose all entries of* $\mathbf{Z}$ *follow i.i.d.* $\mathcal{N}(0,1)$*. If* $\boldsymbol{w} \neq \mathbf{0}_n$*, the population loss* $f(\boldsymbol{v}, \boldsymbol{w})$ *is given by*

$$\frac{1}{8}\left[\boldsymbol{v}^\top\left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)\boldsymbol{v} - 2\boldsymbol{v}^\top\left(\left(1 - \frac{2}{\pi}\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right)\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)\boldsymbol{v}^* + (\boldsymbol{v}^*)^\top\left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)\boldsymbol{v}^*\right].$$

In addition, $f(\boldsymbol{v}, \boldsymbol{w}) = \frac{1}{8}(\boldsymbol{v}^*)^\top\left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)\boldsymbol{v}^*$ *for* $\boldsymbol{w} = \mathbf{0}_n$*.*

Proof of Lemma 1. We notice that

$$f(\boldsymbol{v}, \boldsymbol{w}) = \frac{1}{2}\big(\boldsymbol{v}^\top\mathbb{E}_{\mathbf{Z}}[\sigma(\mathbf{Z}\boldsymbol{w})^\top\sigma(\mathbf{Z}\boldsymbol{w})]\boldsymbol{v} - 2\boldsymbol{v}^\top\mathbb{E}_{\mathbf{Z}}[\sigma(\mathbf{Z}\boldsymbol{w})^\top\sigma(\mathbf{Z}\boldsymbol{w}^*)]\boldsymbol{v}^*$$
$$+ (\boldsymbol{v}^*)^\top\mathbb{E}_{\mathbf{Z}}[\sigma(\mathbf{Z}\boldsymbol{w}^*)^\top\sigma(\mathbf{Z}\boldsymbol{w}^*)]\boldsymbol{v}^*\big).$$

Let $\mathbf{Z}_i^\top$ be the ith row vector of $\mathbf{Z}$. Since $\boldsymbol{w} \neq \mathbf{0}_n$, using Lemma 9, we have

$$\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\boldsymbol{w})\sigma(\mathbf{Z}\boldsymbol{w})^\top\right]_{ii} = \mathbb{E}\left[\sigma(\mathbf{Z}_i^\top\boldsymbol{w})\sigma(\mathbf{Z}_i^\top\boldsymbol{w})\right] = \mathbb{E}\left[1_{\{\mathbf{Z}_i^\top\boldsymbol{w}>0\}}\right] = \frac{1}{2},$$

and for $i \neq j$,

$$\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\boldsymbol{w})\sigma(\mathbf{Z}\boldsymbol{w})^\top\right]_{ij} = \mathbb{E}\left[\sigma(\mathbf{Z}_i^\top\boldsymbol{w})\sigma(\mathbf{Z}_j^\top\boldsymbol{w})\right] = \mathbb{E}\left[1_{\{\mathbf{Z}_i^\top\boldsymbol{w}>0\}}\right]\mathbb{E}\left[1_{\{\mathbf{Z}_j^\top\boldsymbol{w}>0\}}\right] = \frac{1}{4}.$$

Therefore, $\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\boldsymbol{w})\sigma(\mathbf{Z}\boldsymbol{w})^\top\right] = \mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\boldsymbol{w}^*)\sigma(\mathbf{Z}\boldsymbol{w}^*)^\top\right] = \frac{1}{4}\left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)$. Furthermore,

$$\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\boldsymbol{w})\sigma(\mathbf{Z}\boldsymbol{w}^*)^\top\right]_{ii} = \mathbb{E}\left[1_{\{\mathbf{Z}_i^\top\boldsymbol{w}>0, \mathbf{Z}_i^\top\boldsymbol{w}^*>0\}}\right] = \frac{\pi - \theta(\boldsymbol{w}, \boldsymbol{w}^*)}{2\pi},$$

and $\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\boldsymbol{w})\sigma(\mathbf{Z}\boldsymbol{w}^*)^\top\right]_{ij} = \frac{1}{4}$. So,

$$\mathbb{E}_{\mathbf{Z}}\left[\sigma(\mathbf{Z}\boldsymbol{w})\sigma(\mathbf{Z}\boldsymbol{w}^*)^\top\right] = \frac{1}{4}\left(\left(1 - \frac{2\theta(\boldsymbol{w}, \boldsymbol{w}^*)}{\pi}\right)\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right).$$

Then it is easy to validate the first claim. Moreover, if $\boldsymbol{w} = \mathbf{0}_n$, then

$$f(\boldsymbol{v}, \boldsymbol{w}) = \frac{1}{2}(\boldsymbol{v}^*)^\top\mathbb{E}_{\mathbf{Z}}[\sigma(\mathbf{Z}\boldsymbol{w}^*)^\top\sigma(\mathbf{Z}\boldsymbol{w}^*)]\boldsymbol{v}^* = \frac{1}{8}(\boldsymbol{v}^*)^\top\left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)\boldsymbol{v}^*.$$

□

Lemma 2. *If* $\boldsymbol{w} \neq \mathbf{0}_n$ *and* $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \in (0, \pi)$*, the partial gradients of* $f(\boldsymbol{v}, \boldsymbol{w})$ *w.r.t.* $\boldsymbol{v}$ *and* $\boldsymbol{w}$ *are*

$$\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}) = \frac{1}{4}\left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)\boldsymbol{v} - \frac{1}{4}\left(\left(1 - \frac{2}{\pi}\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right)\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)\boldsymbol{v}^*$$

and

$$\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) = -\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{2\pi\|\boldsymbol{w}\|}\frac{\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|},$$

respectively.

Proof of Lemma 2. The first claim is trivial, and we only show the second one. Since $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \arccos\left(\frac{\boldsymbol{w}^\top \boldsymbol{w}^*}{\|\boldsymbol{w}\|}\right)$ is differentiable w.r.t. $\boldsymbol{w}$ at $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \in (0, \pi)$, we have

$$\begin{aligned}\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) &= \frac{\boldsymbol{v}^\top \boldsymbol{v}^*}{2\pi} \frac{\partial \theta}{\partial \boldsymbol{w}}(\boldsymbol{w}, \boldsymbol{w}^*) = -\frac{\boldsymbol{v}^\top \boldsymbol{v}^*}{2\pi} \frac{\|\boldsymbol{w}\|^2 \boldsymbol{w}^* - (\boldsymbol{w}^\top \boldsymbol{w}^*)\boldsymbol{w}}{\|\boldsymbol{w}\|^3 \sqrt{1 - \frac{(\boldsymbol{w}^\top \boldsymbol{w}^*)^2}{\|\boldsymbol{w}\|^2}}} \\ &= -\frac{\boldsymbol{v}^\top \boldsymbol{v}^*}{2\pi \|\boldsymbol{w}\|} \frac{\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|}.\end{aligned}$$

□

Proposition 1. *If the true parameter $\boldsymbol{v}^*$ satisfies $(\mathbf{1}_m^\top \boldsymbol{v}^*)^2 < \frac{m+1}{2}\|\boldsymbol{v}^*\|^2$, then*

$$\left\{(\boldsymbol{v}, \boldsymbol{w}) : \boldsymbol{v} = (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1} \left(\frac{-(\mathbf{1}_m^\top \boldsymbol{v}^*)^2}{(m+1)\|\boldsymbol{v}^*\|^2 - (\mathbf{1}_m^\top \boldsymbol{v}^*)^2} \boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right) \boldsymbol{v}^*, \right.$$
$$\left. \theta(\boldsymbol{w}, \boldsymbol{w}^*) = \frac{\pi}{2} \frac{(m+1)\|\boldsymbol{v}^*\|^2}{(m+1)\|\boldsymbol{v}^*\|^2 - (\mathbf{1}_m^\top \boldsymbol{v}^*)^2}\right\}$$

give the saddle points obeying (8), and $\{(\boldsymbol{v}, \boldsymbol{w}) : \theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi,\ \boldsymbol{v} = \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*\}$ are the spurious local minimizers. Otherwise, the model (2) has no saddle points or spurious local minimizers.

Proof of Proposition 1. Suppose $\boldsymbol{v}^\top \boldsymbol{v}^* = 0$ and $\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}) = \mathbf{0}$, then by Lemma 1,

$$0 = \boldsymbol{v}^\top \boldsymbol{v}^* = (\boldsymbol{v}^*)^\top (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1} \left(\left(1 - \frac{2}{\pi}\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right) \boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right) \boldsymbol{v}^*. \tag{18}$$

From (18) it follows that

$$\frac{2}{\pi}\theta(\boldsymbol{w}, \boldsymbol{w}^*)(\boldsymbol{v}^*)^\top (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1} \boldsymbol{v}^* = (\boldsymbol{v}^*)^\top (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1} \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right) \boldsymbol{v}^* = \|\boldsymbol{v}^*\|^2. \tag{19}$$

On the other hand, from (18) it also follows that

$$\left(\frac{2}{\pi}\theta(\boldsymbol{w}, \boldsymbol{w}^*) - 1\right)(\boldsymbol{v}^*)^\top (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1} \boldsymbol{v}^* = (\boldsymbol{v}^*)^\top (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1} \mathbf{1}_m (\mathbf{1}_m^\top \boldsymbol{v}^*) = \frac{(\mathbf{1}^\top \boldsymbol{v}^*)^2}{m+1},$$

where we used $(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)\mathbf{1}_m = (m+1)\mathbf{1}_m$. Taking the difference of the two equalities above gives

$$(\boldsymbol{v}^*)^\top (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1} \boldsymbol{v}^* = \|\boldsymbol{v}^*\|^2 - \frac{(\mathbf{1}_m^\top \boldsymbol{v}^*)^2}{m+1}. \tag{20}$$

By (19), we have $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \frac{\pi}{2} \frac{(m+1)\|\boldsymbol{v}^*\|^2}{(m+1)\|\boldsymbol{v}^*\|^2 - (\mathbf{1}_m^\top \boldsymbol{v}^*)^2}$, which requires

$$\frac{\pi}{2} \frac{(m+1)\|\boldsymbol{v}^*\|^2}{(m+1)\|\boldsymbol{v}^*\|^2 - (\mathbf{1}_m^\top \boldsymbol{v}^*)^2} < \pi, \text{ or equivalently, } (\mathbf{1}_m^\top \boldsymbol{v}^*)^2 < \frac{m+1}{2}\|\boldsymbol{v}^*\|^2.$$

Furthermore, since $\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}) = 0$, we have

$$\begin{aligned}\boldsymbol{v} &= (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1} \left(\left(1 - \frac{2}{\pi}\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right) \boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right) \boldsymbol{v}^* \\ &= (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1} \left(\frac{-(\mathbf{1}_m^\top \boldsymbol{v}^*)^2}{(m+1)\|\boldsymbol{v}^*\|^2 - (\mathbf{1}_m^\top \boldsymbol{v}^*)^2} \boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right) \boldsymbol{v}^*.\end{aligned}$$

Next, we check the local optimality of the staionary points. By ignoring the scaling and constant terms, we rewrite the objective function as

$$\tilde{f}(\boldsymbol{v}, \theta) := \boldsymbol{v}^\top \left(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right)\boldsymbol{v} - 2\boldsymbol{v}^\top \left(\left(1 - \frac{2}{\pi}\theta\right) \boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\right) \boldsymbol{v}^*, \text{ for } \theta \in [0, \pi].$$

It is easy to check that its Hessian matrix

$$\nabla^2 \tilde{f}(\boldsymbol{v}, \theta) = \begin{bmatrix} 2(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top) & \frac{4}{\pi}\boldsymbol{v}^* \\ \frac{4}{\pi}(\boldsymbol{v}^*)^\top & 0 \end{bmatrix}$$

is indefinite. Therefore, the stationary points are saddle points.

Moreover, if $(\mathbf{1}_m^\top \boldsymbol{v}^*)^2 < \frac{m+1}{2}\|\boldsymbol{v}^*\|^2$, at the point $(\boldsymbol{v}, \theta) = ((\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*, \pi)$, we have

$$\begin{aligned} \boldsymbol{v}^\top \boldsymbol{v}^* &= (\boldsymbol{v}^*)^\top (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^* \\ &= \|\boldsymbol{v}^*\|^2 - 2(\boldsymbol{v}^*)^\top (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1}\boldsymbol{v}^* = \frac{2(\mathbf{1}_m^\top \boldsymbol{v}^*)^2}{m+1} - \|\boldsymbol{v}^*\|^2 < 0, \end{aligned} \tag{21}$$

where we used (20) in the last identity. We consider an arbitrary point $(\boldsymbol{v} + \Delta\boldsymbol{v}, \pi + \Delta\theta)$ in the neighborhood of $(\boldsymbol{v}, \pi)$ with $\Delta\theta \le 0$. The perturbed objective value is

$$\begin{aligned} \tilde{f}(\boldsymbol{v} + \Delta\boldsymbol{v}, \pi + \Delta\theta) = (\boldsymbol{v} + \Delta\boldsymbol{v})^\top \big(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\big)(\boldsymbol{v} + \Delta\boldsymbol{v}) - 2(\boldsymbol{v} + \Delta\boldsymbol{v})^\top \big(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m\big)\boldsymbol{v}^* \\ + \frac{2\Delta\theta}{\pi}(\boldsymbol{v} + \Delta\boldsymbol{v})^\top \boldsymbol{v}^*. \end{aligned}$$

On the right hand side, since $\boldsymbol{v} = (\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$ is the unique minimizer to the quadratic function $\tilde{f}(\boldsymbol{v}, \pi)$, we have if $\Delta\boldsymbol{v} \neq \mathbf{0}_m$,

$$(\boldsymbol{v} + \Delta\boldsymbol{v})^\top \big(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\big)(\boldsymbol{v} + \Delta\boldsymbol{v}) - 2(\boldsymbol{v} + \Delta\boldsymbol{v})^\top \big(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m\big)\boldsymbol{v}^* > \tilde{f}(\boldsymbol{v}, \pi).$$

Moreover, for sufficently small $\|\Delta\boldsymbol{v}\|$, it holds that $\Delta\theta \cdot (\boldsymbol{v} + \Delta\boldsymbol{v})^\top \boldsymbol{v}^* > 0$ for $\Delta\theta < 0$ because of (21). Therefore, $\tilde{f}(\boldsymbol{v} + \Delta\boldsymbol{v}, \pi + \Delta\theta) > \tilde{f}(\boldsymbol{v}, \pi)$ whenever $(\Delta\boldsymbol{v}, \Delta\theta)$ is small and non-zero, and $((\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*, \pi)$ is a spurious local minimizer of $\tilde{f}$.

To prove the second claim, suppose $(\mathbf{1}_m^\top \boldsymbol{v}^*)^2 \ge \frac{m+1}{2}\|\boldsymbol{v}^*\|^2$, then either $\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w})$ does not exist, or $\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w})$ and $\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w})$ do not vanish simultaneously, and thus there is no stationary point.

At the point $(\boldsymbol{v}, \theta) = ((\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*, \pi)$, we have

$$\boldsymbol{v}^\top \boldsymbol{v}^* = \frac{2(\mathbf{1}_m^\top \boldsymbol{v}^*)^2}{m+1} - \|\boldsymbol{v}^*\|^2 \ge 0.$$

If $\boldsymbol{v}^\top \boldsymbol{v}^* > 0$, since $\nabla \tilde{f}(\boldsymbol{v}, \theta) = \frac{1}{4}[\mathbf{0}_m^\top, \frac{2}{\pi}\boldsymbol{v}^\top \boldsymbol{v}^*]^\top$, a small perturbation $[\mathbf{0}_m^\top, \Delta\theta]^\top$ with $\Delta\theta < 0$ will give a strictly decreased objective value, so $(\boldsymbol{v}, \theta) = \big((\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top)^{-1}(\mathbf{1}_m \mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*, \pi\big)$ is not a local minimizer. If $\boldsymbol{v}^\top \boldsymbol{v}^* = 0$, then $\nabla \tilde{f}(\boldsymbol{v}, \theta) = \mathbf{0}_{n+1}$, the same conclusion can be reached by examining the second order necessary condition. □

Lemma 3. *For any differentiable points $(\boldsymbol{v}, \boldsymbol{w})$ and $(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})$ with $\min\{\|\boldsymbol{w}\|, \|\tilde{\boldsymbol{w}}\|\} = c_{\boldsymbol{w}} > 0$ and $\max\{\|\boldsymbol{v}\|, \|\tilde{\boldsymbol{v}}\|\} = C_{\boldsymbol{v}}$, there exists a Lipschitz constant $L > 0$ depending on $C_{\boldsymbol{v}}$ and $c_{\boldsymbol{w}}$, such that*

$$\|\nabla f(\boldsymbol{v}, \boldsymbol{w}) - \nabla f(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})\| \le L\|(\boldsymbol{v}, \boldsymbol{w}) - (\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})\|.$$

Proof of Lemma 3. It is easy to check that $\|\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\| = m + 1$. Then

$$\begin{aligned} \left\|\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}) - \frac{\partial f}{\partial \boldsymbol{v}}(\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})\right\| &= \left\|\big(\boldsymbol{I}_m + \mathbf{1}_m \mathbf{1}_m^\top\big)(\boldsymbol{v} - \tilde{\boldsymbol{v}}) + \frac{2}{\pi}(\theta(\boldsymbol{w}, \boldsymbol{w}^*) - \theta(\tilde{\boldsymbol{w}}, \boldsymbol{w}^*))\boldsymbol{v}^*\right\| \\ &\le (m+1)\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\| + \frac{2\|\boldsymbol{v}^*\|}{\pi}|\theta(\boldsymbol{w}, \boldsymbol{w}^*) - \theta(\tilde{\boldsymbol{w}}, \boldsymbol{w}^*)| \\ &\le (m+1)\|\boldsymbol{v} - \tilde{\boldsymbol{v}}\| + \frac{\|\boldsymbol{v}^*\|}{c_{\boldsymbol{w}}}\|\boldsymbol{w} - \tilde{\boldsymbol{w}}\| \\ &\le \left(m + 1 + \frac{\|\boldsymbol{v}^*\|}{c_{\boldsymbol{w}}}\right)\|(\boldsymbol{v}, \boldsymbol{w}) - (\tilde{\boldsymbol{v}}, \tilde{\boldsymbol{w}})\|, \end{aligned}$$

where the last inequality is due to Lemma 10.1.

We further have

$$\begin{aligned}
\left\|\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v},\boldsymbol{w})-\frac{\partial f}{\partial \boldsymbol{w}}(\tilde{\boldsymbol{v}},\tilde{\boldsymbol{w}})\right\| &= \left\|\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{2\pi\|\boldsymbol{w}\|}\frac{\left(\boldsymbol{I}_n-\frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n-\frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|}-\frac{\tilde{\boldsymbol{v}}^\top\boldsymbol{v}^*}{2\pi\|\tilde{\boldsymbol{w}}\|}\frac{\left(\boldsymbol{I}_n-\frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n-\frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*\right\|}\right\| \\
&\le \left\|\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{2\pi\|\boldsymbol{w}\|}\frac{\left(\boldsymbol{I}_n-\frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n-\frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|}-\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{2\pi\|\tilde{\boldsymbol{w}}\|}\frac{\left(\boldsymbol{I}_n-\frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n-\frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*\right\|}\right\| \\
&\quad+\left\|\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{2\pi\|\tilde{\boldsymbol{w}}\|}\frac{\left(\boldsymbol{I}_n-\frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n-\frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*\right\|}-\frac{\tilde{\boldsymbol{v}}^\top\boldsymbol{v}^*}{2\pi\|\tilde{\boldsymbol{w}}\|}\frac{\left(\boldsymbol{I}_n-\frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n-\frac{\tilde{\boldsymbol{w}}\tilde{\boldsymbol{w}}^\top}{\|\tilde{\boldsymbol{w}}\|^2}\right)\boldsymbol{w}^*\right\|}\right\| \\
&\le \frac{|\boldsymbol{v}^\top\boldsymbol{v}^*|}{2\pi c_{\boldsymbol{w}}^2}\|\boldsymbol{w}-\tilde{\boldsymbol{w}}\|+\frac{\|\boldsymbol{v}^*\|}{2\pi c_{\boldsymbol{w}}}\|\boldsymbol{v}-\tilde{\boldsymbol{v}}\| \\
&\le \frac{(C_{\boldsymbol{v}}+c_{\boldsymbol{w}})\|\boldsymbol{v}^*\|}{2\pi c_{\boldsymbol{w}}^2}\|(\boldsymbol{v},\boldsymbol{w})-(\tilde{\boldsymbol{v}},\tilde{\boldsymbol{w}})\|,
\end{aligned}$$

where the second last inequality is to due to Lemma 10.2. Combining the two inequalities above validates the claim. □

Lemma 4. *The expected partial gradient of $\ell(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})$ w.r.t. $\boldsymbol{v}$ is*

$$\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right]=\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w}).$$

Let $\mu(x)=\max\{x,0\}$ in (5). The expected coarse gradient w.r.t. $\boldsymbol{w}$ is

$$\mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\Big]=\frac{h(\boldsymbol{v},\boldsymbol{v}^*)}{2\sqrt{2\pi}}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}-\cos\left(\frac{\theta(\boldsymbol{w},\boldsymbol{w}^*)}{2}\right)\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{\sqrt{2\pi}}\frac{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}+\boldsymbol{w}^*}{\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}+\boldsymbol{w}^*\right\|},^{3}$$

where $h(\boldsymbol{v},\boldsymbol{v}^*)=\|\boldsymbol{v}\|^2+(\mathbf{1}_m^\top\boldsymbol{v})^2-(\mathbf{1}_m^\top\boldsymbol{v})(\mathbf{1}_m^\top\boldsymbol{v}^*)+\boldsymbol{v}^\top\boldsymbol{v}^*$.

Proof of Lemma 4. The first claim is true because $\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})$ is linear in $\boldsymbol{v}$. By (5),

$$\boldsymbol{g}_\mu(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})=\mathbf{Z}^\top\big(\mu'(\mathbf{Z}\boldsymbol{w})\odot\boldsymbol{v}\big)\Big(\boldsymbol{v}^\top\sigma(\mathbf{Z}\boldsymbol{w})-(\boldsymbol{v}^*)^\top\sigma(\mathbf{Z}\boldsymbol{w}^*)\Big).$$

Using the fact that $\mu'=\sigma=1_{\{x>0\}}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right] &= \mathbb{E}_{\mathbf{Z}}\left[\left(\sum_{i=1}^m v_i\sigma(\mathbf{Z}_i^\top\boldsymbol{w})-\sum_{i=1}^m v_i^*\sigma(\mathbf{Z}_i^\top\boldsymbol{w}^*)\right)\left(\sum_{i=1}^m \mathbf{Z}_i v_i\sigma(\mathbf{Z}_i^\top\boldsymbol{w})\right)\right] \\
&= \mathbb{E}_{\mathbf{Z}}\left[\left(\sum_{i=1}^m v_i 1_{\{\mathbf{Z}_i^\top\boldsymbol{w}>0\}}-\sum_{i=1}^m v_i^* 1_{\{\mathbf{Z}_i^\top\boldsymbol{w}^*>0\}}\right)\left(\sum_{i=1}^m 1_{\{\mathbf{Z}_i^\top\boldsymbol{w}>0\}} v_i\mathbf{Z}_i\right)\right].
\end{aligned}$$

Invoking Lemma 9, we have

$$\mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top\boldsymbol{w}>0,\mathbf{Z}_j^\top\boldsymbol{w}>0\}}\right]=\begin{cases}\frac{1}{\sqrt{2\pi}}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} & \text{if } i=j,\\ \frac{1}{2\sqrt{2\pi}}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} & \text{if } i\neq j,\end{cases}$$

and

$$\mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top\boldsymbol{w}>0,\mathbf{Z}_j^\top\boldsymbol{w}^*>0\}}\right]=\begin{cases}\frac{\cos(\theta(\boldsymbol{w},\boldsymbol{w}^*)/2)}{\sqrt{2\pi}}\frac{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}+\boldsymbol{w}^*}{\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}+\boldsymbol{w}^*\right\|} & \text{if } i=j,\\ \frac{1}{2\sqrt{2\pi}}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} & \text{if } i\neq j.\end{cases}$$

[3] We redefine the second term as $\mathbf{0}_n$ in the case $\theta(\boldsymbol{w},\boldsymbol{w}^*)=\pi$, or equivalently, $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}+\boldsymbol{w}^*=\mathbf{0}_n$.

Therefore,

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right] &= \sum_{i=1}^{m} v_i^2 \mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top \boldsymbol{w}>0\}}\right] + \sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m} v_i v_j \mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top \boldsymbol{w}>0, \mathbf{Z}_j^\top \boldsymbol{w}>0\}}\right] \\
&\quad - \sum_{i=1}^{m} v_i v_i^* \mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top \boldsymbol{w}>0, \mathbf{Z}_i^\top \boldsymbol{w}^*>0\}}\right] - \sum_{i=1}^{m}\sum_{\substack{j=1\\ j\neq i}}^{m} v_i v_j^* \mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_i^\top \boldsymbol{w}>0, \mathbf{Z}_j^\top \boldsymbol{w}^*>0\}}\right] \\
&= \frac{1}{2\sqrt{2\pi}}\left(\|\boldsymbol{v}\|^2 + (\mathbf{1}_m^\top \boldsymbol{v})^2\right)\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - \cos\left(\frac{\theta(\boldsymbol{w},\boldsymbol{w}^*)}{2}\right)\frac{\boldsymbol{v}^\top \boldsymbol{v}^*}{\sqrt{2\pi}}\frac{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}+\boldsymbol{w}^*}{\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}+\boldsymbol{w}^*\right\|} \\
&\quad - \frac{1}{2\sqrt{2\pi}}\left((\mathbf{1}_m^\top \boldsymbol{v})(\mathbf{1}_m^\top \boldsymbol{v}^*) - \boldsymbol{v}^\top \boldsymbol{v}^*\right)\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|},
\end{aligned}$$

and the result follows. □

Lemma 5. *If $\boldsymbol{w} \neq \mathbf{0}_n$ and $\theta(\boldsymbol{w},\boldsymbol{w}^*) \in (0,\pi)$, then the inner product between the expected coarse and true gradients w.r.t. $\boldsymbol{w}$ is*

$$\left\langle \mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\Big], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v},\boldsymbol{w})\right\rangle = \frac{\sin\left(\theta(\boldsymbol{w},\boldsymbol{w}^*)\right)}{2(\sqrt{2\pi})^3\|\boldsymbol{w}\|}(\boldsymbol{v}^\top \boldsymbol{v}^*)^2 \geq 0.$$

Moreover, if further $\|\boldsymbol{v}\| \leq C_{\boldsymbol{v}}$, there exists a constant $A > 0$ depending only on $C_{\boldsymbol{v}}$, such that

$$\left\|\mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\Big]\right\|^2 \leq A\left(\left\|\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w})\right\|^2 + \left\langle \mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\Big], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v},\boldsymbol{w})\right\rangle\right).$$

Proof of Lemma 5. By Lemmas 2 and 5, we have

$$\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v},\boldsymbol{w}) = -\frac{\boldsymbol{v}^\top \boldsymbol{v}^*}{2\pi\|\boldsymbol{w}\|}\frac{\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|}$$

and

$$\mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\Big] = \frac{h(\boldsymbol{v},\boldsymbol{v}^*)}{2\sqrt{2\pi}}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - \cos\left(\frac{\theta(\boldsymbol{w},\boldsymbol{w}^*)}{2}\right)\frac{\boldsymbol{v}^\top \boldsymbol{v}^*}{\sqrt{2\pi}}\frac{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}+\boldsymbol{w}^*}{\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}+\boldsymbol{w}^*\right\|}.$$

Notice that $\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w} = \boldsymbol{0}_n$ and $\|\boldsymbol{w}^*\| = 1$, if $\theta(\boldsymbol{w}, \boldsymbol{w}_*) \neq 0, \pi$, then we have

$$
\begin{aligned}
&\left\langle \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) \right\rangle \\
=\ & \cos\left(\frac{\theta(\boldsymbol{w}, \boldsymbol{w}^*)}{2}\right) \frac{(\boldsymbol{v}^\top \boldsymbol{v}^*)^2}{(\sqrt{2\pi})^3} \left\langle \frac{1}{\|\boldsymbol{w}\|} \frac{\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|}, \frac{\boldsymbol{w}^*}{\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \boldsymbol{w}^*\right\|} \right\rangle \\
=\ & \cos\left(\frac{\theta(\boldsymbol{w}, \boldsymbol{w}^*)}{2}\right) \frac{(\boldsymbol{v}^\top \boldsymbol{v}^*)^2}{(\sqrt{2\pi})^3} \frac{\|\boldsymbol{w}\|^2 - (\boldsymbol{w}^\top \boldsymbol{w}^*)^2}{\left\|\|\boldsymbol{w}\|^2 \boldsymbol{w}^* - \boldsymbol{w}(\boldsymbol{w}^\top \boldsymbol{w}^*)\right\| \|\boldsymbol{w} + \|\boldsymbol{w}\| \boldsymbol{w}^*\|} \\
=\ & \cos\left(\frac{\theta(\boldsymbol{w}, \boldsymbol{w}^*)}{2}\right) \frac{(\boldsymbol{v}^\top \boldsymbol{v}^*)^2}{(\sqrt{2\pi})^3} \frac{\|\boldsymbol{w}\|^2 - (\boldsymbol{w}^\top \boldsymbol{w}^*)^2}{\sqrt{\|\boldsymbol{w}\|^4 - \|\boldsymbol{w}\|^2 (\boldsymbol{w}^\top \boldsymbol{w}^*)^2} \sqrt{2(\|\boldsymbol{w}\|^2 + \|\boldsymbol{w}\|(\boldsymbol{w}^\top \boldsymbol{w}^*))}} \\
=\ & \cos\left(\frac{\theta(\boldsymbol{w}, \boldsymbol{w}^*)}{2}\right) \frac{(\boldsymbol{v}^\top \boldsymbol{v}^*)^2}{4(\sqrt{\pi}\|\boldsymbol{w}\|)^3} \frac{\|\boldsymbol{w}\|^2 - (\boldsymbol{w}^\top \boldsymbol{w}^*)^2}{\sqrt{\|\boldsymbol{w}\|^2 - (\boldsymbol{w}^\top \boldsymbol{w}^*)^2} \sqrt{\|\boldsymbol{w}\| + (\boldsymbol{w}^\top \boldsymbol{w}^*)}} \\
=\ & \cos\left(\frac{\theta(\boldsymbol{w}, \boldsymbol{w}^*)}{2}\right) \frac{(\boldsymbol{v}^\top \boldsymbol{v}^*)^2 \sqrt{1 - \frac{\boldsymbol{w}^\top \boldsymbol{w}^*}{\|\boldsymbol{w}\|}}}{4(\sqrt{\pi})^3 \|\boldsymbol{w}\|} \\
=\ & \cos\left(\frac{\theta(\boldsymbol{w}, \boldsymbol{w}^*)}{2}\right) \frac{(\boldsymbol{v}^\top \boldsymbol{v}^*)^2 \sqrt{1 - \cos(\theta(\boldsymbol{w}, \boldsymbol{w}^*))}}{4(\sqrt{\pi})^3 \|\boldsymbol{w}\|} \\
=\ & \frac{\sin\left(\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right)}{2(\sqrt{2\pi})^3 \|\boldsymbol{w}\|} (\boldsymbol{v}^\top \boldsymbol{v}^*)^2.
\end{aligned}
$$

To show the second claim, without loss of generality, we assume $\|\boldsymbol{w}\| = 1$. Denote $\theta := \theta(\boldsymbol{w}, \boldsymbol{w}^*)$. By Lemma 1, we have

$$
\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}) = \frac{1}{4}(\boldsymbol{I}_m + \boldsymbol{1}_m \boldsymbol{1}_m^\top)\boldsymbol{v} - \frac{1}{4}\left(\left(1 - \frac{2\theta}{\pi}\right)\boldsymbol{I}_m + \boldsymbol{1}_m \boldsymbol{1}_m^\top\right)\boldsymbol{v}^*.
$$

By Lemma 4,

$$
\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right] = \frac{h(\boldsymbol{v}, \boldsymbol{v}^*)}{2\sqrt{2\pi}}\boldsymbol{w} - \cos\left(\frac{\theta}{2}\right) \frac{\boldsymbol{v}^\top \boldsymbol{v}^*}{\sqrt{2\pi}} \frac{\boldsymbol{w} + \boldsymbol{w}^*}{\|\boldsymbol{w} + \boldsymbol{w}^*\|}, \tag{22}
$$

where

$$
\begin{aligned}
h(\boldsymbol{v}, \boldsymbol{v}^*) &= \|\boldsymbol{v}\|^2 + (\boldsymbol{1}_m^\top \boldsymbol{v})^2 - (\boldsymbol{1}_m^\top \boldsymbol{v})(\boldsymbol{1}_m^\top \boldsymbol{v}^*) + \boldsymbol{v}^\top \boldsymbol{v}^* \\
&= \boldsymbol{v}^\top \left(\boldsymbol{I}_m + \boldsymbol{1}_m \boldsymbol{1}_m^\top\right)\boldsymbol{v} - \boldsymbol{v}^\top (\boldsymbol{1}_m \boldsymbol{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^* \\
&= \boldsymbol{v}^\top \left(\boldsymbol{I}_m + \boldsymbol{1}_m \boldsymbol{1}_m^\top\right)\boldsymbol{v} - \boldsymbol{v}^\top \left(\boldsymbol{1}_m \boldsymbol{1}_m^\top + \left(1 - \frac{2\theta}{\pi}\right)\boldsymbol{I}_m\right)\boldsymbol{v}^* + 2\left(1 - \frac{\theta}{\pi}\right)\boldsymbol{v}^\top \boldsymbol{v}^* \\
&= 4\boldsymbol{v}^\top \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w}) + 2\left(1 - \frac{\theta}{\pi}\right)\boldsymbol{v}^\top \boldsymbol{v}^*,
\end{aligned} \tag{23}
$$

and by the first claim,

$$
\left\langle \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w}) \right\rangle = \frac{\sin(\theta)}{2(\sqrt{2\pi})^3}(\boldsymbol{v}^\top \boldsymbol{v}^*)^2.
$$

Hence, for some A depending only on $C_{\boldsymbol{v}}$, we have

$$
\begin{aligned}
&\left\|\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right]\right\|^2 \\
&= \left\|\frac{2\boldsymbol{v}^\top \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w})}{\sqrt{2\pi}}\boldsymbol{w} + \cos\left(\frac{\theta}{2}\right)\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{\sqrt{2\pi}}\left(\boldsymbol{w} - \frac{\boldsymbol{w}+\boldsymbol{w}^*}{\|\boldsymbol{w}+\boldsymbol{w}^*\|}\right) + \left(1 - \frac{\theta}{\pi} - \cos\left(\frac{\theta}{2}\right)\right)\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{\sqrt{2\pi}}\boldsymbol{w}\right\|^2 \\
&\le \frac{6C_{\boldsymbol{v}}^2}{\pi}\left\|\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w})\right\|^2 + \cos^2\left(\frac{\theta}{2}\right)\frac{3(\boldsymbol{v}^\top\boldsymbol{v}^*)^2}{2\pi}\left\|\boldsymbol{w} - \frac{\boldsymbol{w}+\boldsymbol{w}^*}{\|\boldsymbol{w}+\boldsymbol{w}^*\|}\right\|^2 \\
&\quad + \left(1 - \frac{\theta}{\pi} - \cos\left(\frac{\theta}{2}\right)\right)^2\frac{3(\boldsymbol{v}^\top\boldsymbol{v}^*)^2}{2\pi} \\
&\le \frac{6C_{\boldsymbol{v}}^2}{\pi}\left\|\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w})\right\|^2 + \cos^2\left(\frac{\theta}{2}\right)\frac{3\theta^2}{8\pi}(\boldsymbol{v}^\top\boldsymbol{v}^*)^2 + \left(1 - \frac{\theta}{\pi} - \cos\left(\frac{\theta}{2}\right)\right)^2\frac{3(\boldsymbol{v}^\top\boldsymbol{v}^*)^2}{2\pi} \\
&\le \frac{6C_{\boldsymbol{v}}^2}{\pi}\left\|\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w})\right\|^2 + \frac{3\pi}{8}\cos^2\left(\frac{\theta}{2}\right)\sin^2\left(\frac{\theta}{2}\right)(\boldsymbol{v}^\top\boldsymbol{v}^*)^2 + \frac{3\sin(\theta)}{2\pi}(\boldsymbol{v}^\top\boldsymbol{v}^*)^2 \\
&\le A\left(\left\|\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w})\right\|^2 + \left\langle \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v},\boldsymbol{w})\right\rangle\right),
\end{aligned}
$$

where the equality is due to (22) and (23), the first inequality is due to Cauchy-Schwarz inequality, the second inequality holds because the angle between $\boldsymbol{w}$ and $\frac{\boldsymbol{w}+\boldsymbol{w}^*}{\|\boldsymbol{w}+\boldsymbol{w}^*\|}$ is $\frac{\theta}{2}$ and $\left\|\boldsymbol{w} - \frac{\boldsymbol{w}+\boldsymbol{w}^*}{\|\boldsymbol{w}+\boldsymbol{w}^*\|}\right\| \le \frac{\theta}{2}$, whereas the third inequality is due to $\sin(x) \ge \frac{2x}{\pi}$, $\cos(x) \ge 1 - \frac{2x}{\pi}$, and

$$
\left(1 - \frac{2x}{\pi} - \cos(x)\right)^2 \le \left(\cos(x) - 1 + \frac{2x}{\pi}\right)\left(\cos(x) + 1 - \frac{2x}{\pi}\right) \le \sin(x)(2\cos(x)) = \sin(2x)
$$

for all $x \in [0, \frac{\pi}{2}]$. □

Lemma 6. *When Algorithm 1 converges,* $\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right]$ *and* $\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right]$ *vanish simultaneously, which only occurs at the*

1. *Saddle points where (8) is satisfied according to Proposition 1.*
2. *Minimizers of (2) where* $\boldsymbol{v} = \boldsymbol{v}^*$, $\theta(\boldsymbol{w},\boldsymbol{w}^*) = 0$, *or* $\boldsymbol{v} = (\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$, $\theta(\boldsymbol{w},\boldsymbol{w}^*) = \pi$.

Proof of Lemma 6. By Lemma 4, suppose we have

$$
\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right] = \frac{1}{4}(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top)\boldsymbol{v} - \frac{1}{4}\left(\left(1 - \frac{2}{\pi}\theta(\boldsymbol{w},\boldsymbol{w}^*)\right)\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)\boldsymbol{v}^* = \mathbf{0}_m \quad (24)
$$

and

$$
\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right] = \frac{h(\boldsymbol{v},\boldsymbol{v}^*)}{2\sqrt{2\pi}}\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - \cos\left(\frac{\theta(\boldsymbol{w},\boldsymbol{w}^*)}{2}\right)\frac{\boldsymbol{v}^\top\boldsymbol{v}^*}{\sqrt{2\pi}}\frac{\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \boldsymbol{w}^*}{\left\|\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} + \boldsymbol{w}^*\right\|} = \mathbf{0}_n, \quad (25)
$$

where $h(\boldsymbol{v},\boldsymbol{v}^*) = \|\boldsymbol{v}\|^2 + (\mathbf{1}_m^\top\boldsymbol{v})^2 - (\mathbf{1}_m^\top\boldsymbol{v})(\mathbf{1}_m^\top\boldsymbol{v}^*) + \boldsymbol{v}^\top\boldsymbol{v}^*$. By (25), we must have $\theta(\boldsymbol{w},\boldsymbol{w}^*) = 0$ or $\theta(\boldsymbol{w},\boldsymbol{w}^*) = \pi$ or $\boldsymbol{v}^\top\boldsymbol{v}^* = 0$.

If $\theta(\boldsymbol{w},\boldsymbol{w}^*) = 0$, then by (24), $\boldsymbol{v} = \boldsymbol{v}^*$, and (25) is satisfied.

If $\theta(\boldsymbol{w},\boldsymbol{w}^*) = \pi$, then by (24), $\boldsymbol{v} = (\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$, and (25) is satisfied.

If $\boldsymbol{v}^\top\boldsymbol{v}^* = 0$, then by (24), we have the expressions for $\boldsymbol{v}$ and $\theta(\boldsymbol{w},\boldsymbol{w}^*)$ from Proposition 1, and (25) is satisfied. □

Lemma 7. *Let $\mu(x) = x$ in (5). Then the expected coarse partial gradient w.r.t. $\boldsymbol{w}$ is*

$$\mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\Big] = \frac{1}{\sqrt{2\pi}}\left(\|\boldsymbol{v}\|^2 \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - (\boldsymbol{v}^\top \boldsymbol{v}^*)\boldsymbol{w}^*\right).$$

If $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi$ and $\boldsymbol{v} = \left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$,

$$\left\|\mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\Big]\right\| = \frac{2(m-1)}{\sqrt{2\pi}(m+1)^2}(\mathbf{1}_m^\top \boldsymbol{v}^*)^2 \geq 0,$$

i.e., $\mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\Big]$ does not vanish at the spurious local minimizers if $\mathbf{1}_m^\top \boldsymbol{v}^* \neq 0$ and $m > 1$,.

Proof of Lemma 7. By (5),

$$\boldsymbol{g}_\mu(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z}) = \mathbf{Z}^\top\big(\mu'(\mathbf{Z}\boldsymbol{w}) \odot \boldsymbol{v}\big)\Big(\boldsymbol{v}^\top \sigma(\mathbf{Z}\boldsymbol{w}) - (\boldsymbol{v}^*)^\top \sigma(\mathbf{Z}\boldsymbol{w}^*)\Big).$$

Using the facts that $\mu' = 1$ and $\sigma = 1_{\{x>0\}}$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right] &= \mathbb{E}_{\mathbf{Z}}\left[\left(\sum_{i=1}^m v_i 1_{\{\mathbf{Z}_i^\top \boldsymbol{w} > 0\}} - \sum_{i=1}^m v_i^* 1_{\{\mathbf{Z}_i^\top \boldsymbol{w}^* > 0\}}\right)\left(\sum_{i=1}^m v_i \mathbf{Z}_i\right)\right] \\
&= \sum_{i=1}^m \sum_{j=1}^m v_i v_j \mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_j^\top \boldsymbol{w} > 0\}}\right] - \sum_{i=1}^m \sum_{j=1}^m v_i^* v_j \mathbb{E}\left[\mathbf{Z}_i 1_{\{\mathbf{Z}_j^\top \boldsymbol{w}^* > 0\}}\right] \\
&= \frac{1}{\sqrt{2\pi}}\left(\|\boldsymbol{v}\|^2 \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - (\boldsymbol{v}^\top \boldsymbol{v}^*)\boldsymbol{w}^*\right).
\end{aligned}$$

In the last equality above, we called the third identity in Lemma 9. If $\theta(\boldsymbol{w}, \boldsymbol{w}^*) = \pi$ and $\boldsymbol{v} = \left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$, then

$$\begin{aligned}
&\|\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\right]\| = \frac{1}{\sqrt{2\pi}}|\boldsymbol{v}^\top(\boldsymbol{v} + \boldsymbol{v}^*)| \\
&= \frac{1}{\sqrt{2\pi}}\left|(\boldsymbol{v}^*)^\top(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)^{-1}\left(\left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m) + \boldsymbol{I}_m\right)\boldsymbol{v}^*\right| \\
&= \frac{2}{\sqrt{2\pi}}\left|(\boldsymbol{v}^*)^\top(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)^{-1}\left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)^{-1}\mathbf{1}_m(\mathbf{1}_m^\top \boldsymbol{v}^*)\right| \\
&= \frac{2}{\sqrt{2\pi}(m+1)^2}\left|(\boldsymbol{v}^*)^\top(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\mathbf{1}_m(\mathbf{1}_m^\top \boldsymbol{v}^*)\right| \\
&= \frac{2(m-1)}{\sqrt{2\pi}(m+1)^2}(\mathbf{1}_m^\top \boldsymbol{v}^*)^2.
\end{aligned}$$

In the third equality, we used the identity $(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top)\mathbf{1}_m = (m+1)\mathbf{1}_m$ twice. □

Lemma 8. *If $\boldsymbol{w} \neq \mathbf{0}_n$ and $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \in (0, \pi)$, then the inner product between the expected coarse and true gradients w.r.t. $\boldsymbol{w}$ is*

$$\left\langle \mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\Big], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w})\right\rangle = \frac{\sin\left(\theta(\boldsymbol{w}, \boldsymbol{w}^*)\right)}{(\sqrt{2\pi})^3\|\boldsymbol{w}\|}(\boldsymbol{v}^\top \boldsymbol{v}^*)^2 \geq 0.$$

When $\theta(\boldsymbol{w}, \boldsymbol{w}^*) \to \pi$, $\boldsymbol{v} \to \left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$, if $\mathbf{1}_m^\top \boldsymbol{v}^* \neq 0$ and $m > 1$, we have

$$\frac{\left\|\mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\Big]\right\|^2}{\left\|\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}, \boldsymbol{w})\right\|^2 + \left\langle \mathbb{E}_{\mathbf{Z}}\Big[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v}, \boldsymbol{w}; \mathbf{Z})\Big], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}, \boldsymbol{w})\right\rangle} \to +\infty.$$

Proof of Lemma 8. By Lemmas 2 and 4, we have

$$\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v},\boldsymbol{w}) = -\frac{\boldsymbol{v}^\top \boldsymbol{v}^*}{2\pi\|\boldsymbol{w}\|}\frac{\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|}$$

and

$$\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right] = \frac{1}{\sqrt{2\pi}}\left(\|\boldsymbol{v}\|^2\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} - (\boldsymbol{v}^\top\boldsymbol{v}^*)\boldsymbol{w}^*\right).$$

Since $\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w} = \mathbf{0}_n$ and $\|\boldsymbol{w}^*\| = 1$, if $\theta(\boldsymbol{w},\boldsymbol{w}_*) \neq 0, \pi$, then we have

$$\begin{aligned}
\left\langle \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v},\boldsymbol{w})\right\rangle &= \frac{(\boldsymbol{v}^\top\boldsymbol{v}^*)^2}{(\sqrt{2\pi})^3\|\boldsymbol{w}\|}\frac{(\boldsymbol{w}^*)^\top\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*}{\left\|\left(\boldsymbol{I}_n - \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\|\boldsymbol{w}\|^2}\right)\boldsymbol{w}^*\right\|} \\
&= \frac{(\boldsymbol{v}^\top\boldsymbol{v}^*)^2}{(\sqrt{2\pi})^3\|\boldsymbol{w}\|}\frac{1-\frac{(\boldsymbol{w}^\top\boldsymbol{w}^*)^2}{\|\boldsymbol{w}\|^2}}{\sqrt{1-\frac{(\boldsymbol{w}^\top\boldsymbol{w}^*)^2}{\|\boldsymbol{w}\|^2}}} = \frac{(\boldsymbol{v}^\top\boldsymbol{v}^*)^2}{(\sqrt{2\pi})^3\|\boldsymbol{w}\|}\sqrt{1-\frac{(\boldsymbol{w}^\top\boldsymbol{w}^*)^2}{\|\boldsymbol{w}\|^2}} \\
&= \frac{(\boldsymbol{v}^\top\boldsymbol{v}^*)^2}{(\sqrt{2\pi})^3\|\boldsymbol{w}\|}\sin(\theta(\boldsymbol{w},\boldsymbol{w}^*)).
\end{aligned}$$

When $\theta(\boldsymbol{w},\boldsymbol{w}^*) \to \pi$, $\boldsymbol{v} \to \left(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top\right)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$, both $\left\|\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v},\boldsymbol{w})\right\|$ and $\left\langle \mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right], \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v},\boldsymbol{w})\right\rangle$ converge to 0. But if $\mathbf{1}_m^\top\boldsymbol{v}^* \neq 0$ and $m > 1$, $\left\|\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{id}}(\boldsymbol{v},\boldsymbol{w};\mathbf{Z})\right]\right\| \to \frac{2(m-1)}{\sqrt{2\pi}(m+1)^2}(\mathbf{1}_m^\top\boldsymbol{v}^*)^2 > 0$, which completes the proof. □

Theorem 1. *Let $\{(\boldsymbol{v}^t,\boldsymbol{w}^t)\}$ be the sequence generated by Algorithm 1 with ReLU $\mu(x) = \max\{x,0\}$. Suppose $\|\boldsymbol{v}^t\| \leq C_{\boldsymbol{v}}$ and $\|\boldsymbol{w}^t\| \geq c_{\boldsymbol{w}}$ for all t with some $C_{\boldsymbol{v}}, c_{\boldsymbol{w}} > 0$. Then if the learning rate $\eta > 0$ is sufficiently small, for any initialization $(\boldsymbol{v}^0,\boldsymbol{w}^0)$, the objective sequence $\{f(\boldsymbol{v}^t,\boldsymbol{w}^t)\}$ is monotonically decreasing, and $\{(\boldsymbol{v}^t,\boldsymbol{w}^t)\}$ converges to a saddle point or a (local) minimizer of the population loss minimization (2). In addition, if $\mathbf{1}_m^\top\boldsymbol{v}^* \neq 0$ and $m > 1$, the descent and convergence properties do not hold for Algorithm 1 with the identity function $\mu(x) = x$ near the local minimizers satisfying $\theta(\boldsymbol{w},\boldsymbol{w}^*) = \pi$ and $\boldsymbol{v} = (\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$.*

Proof of Theorem 1 . If $\theta(\boldsymbol{w}^t,\boldsymbol{w}^*) = 0$ or π, then $\theta(\boldsymbol{w}^T,\boldsymbol{w}^*) = 0$ or π for all $T \geq t$, and the original problem reduces to a quadratic program in terms of $\boldsymbol{v}$. So $\{\boldsymbol{v}^t\}$ will converge to $\boldsymbol{v}^*$ or $(\boldsymbol{I}_m + \mathbf{1}_m\mathbf{1}_m^\top)^{-1}(\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m)\boldsymbol{v}^*$ by choosing a suitable step size η. In either case, we have $\left\|\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v}^t,\boldsymbol{w}^t;\mathbf{Z})\right]\right\|$ and $\left\|\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t,\boldsymbol{w}^t;\mathbf{Z})\right]\right\|$ both converge to 0. Else if $\theta(\boldsymbol{w}^t,\boldsymbol{w}^*) \in (0,\pi)$, we define for any $a \in [0,1]$ that

$$\boldsymbol{v}^t(a) := \boldsymbol{v}^t - a(\boldsymbol{v}^{t+1} - \boldsymbol{v}^t) = \boldsymbol{v}^t - a\eta\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell}{\partial \boldsymbol{v}}(\boldsymbol{v}^t,\boldsymbol{w}^t;\mathbf{Z})\right]$$

and

$$\boldsymbol{w}^t(a) := \boldsymbol{w}^t - a(\boldsymbol{w}^{t+1} - \boldsymbol{w}^t) = \boldsymbol{w}^t - a\eta\mathbb{E}_{\mathbf{Z}}\left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t,\boldsymbol{w}^t;\mathbf{Z})\right],$$

which satisfy

$$\boldsymbol{v}^t(0) = \boldsymbol{v}^t,\ \boldsymbol{v}^t(1) = \boldsymbol{v}^{t+1},\ \boldsymbol{w}^t(0) = \boldsymbol{w}^t,\ \boldsymbol{w}^t(1) = \boldsymbol{w}^{t+1}.$$

Since by the assumption, $\|\boldsymbol{v}^t\| \leq C_{\boldsymbol{v}}$ and $\|\boldsymbol{w}^t\| \geq c_{\boldsymbol{w}}$ for all t, for sufficiently small $\eta > 0$, it holds that $\|\boldsymbol{v}^t(a)\| \leq C_{\boldsymbol{v}}$ and $\|\boldsymbol{w}^t(a)\| \geq c_{\boldsymbol{w}}$ for all $a \in [0,1]$. Possibly at some point a_0 where $\theta(\boldsymbol{w}^t(a_0),\boldsymbol{w}^*) = 0$ or π, the partial gradient $\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t(a_0),\boldsymbol{w}^t(a_0))$ does not exist. Otherwise, $\left\|\frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t(a),\boldsymbol{w}^t(a))\right\|$ is uniformly bounded for all $a \in [0,1]/\{a_0\}$, which makes it integrable over the interval $[0,1]$.

Then for some constants L and A depending on $C_{\boldsymbol{v}}$ and $c_{\boldsymbol{w}}$, we have

$$\begin{aligned}
f(\boldsymbol{v}^{t+1}, \boldsymbol{w}^{t+1}) &= f(\boldsymbol{v}^t + (\boldsymbol{v}^{t+1} - \boldsymbol{v}^t), \boldsymbol{w}^t + (\boldsymbol{w}^{t+1} - \boldsymbol{w}^t)) \\
&= f(\boldsymbol{v}^t, \boldsymbol{w}^t) + \int_0^1 \left\langle \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t(a), \boldsymbol{w}^t(a)), \boldsymbol{v}^{t+1} - \boldsymbol{v}^t \right\rangle \mathrm{d}a \\
&\quad + \int_0^1 \left\langle \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t(a), \boldsymbol{w}^t(a)), \boldsymbol{w}^{t+1/2} - \boldsymbol{w}^t \right\rangle \mathrm{d}a \\
&= f(\boldsymbol{v}^t, \boldsymbol{w}^t) + \left\langle \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \boldsymbol{v}^{t+1} - \boldsymbol{v}^t \right\rangle + \left\langle \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \boldsymbol{w}^{t+1/2} - \boldsymbol{w}^t \right\rangle \\
&\quad + \int_0^1 \left\langle \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t(a), \boldsymbol{w}^t(a)) - \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \boldsymbol{v}^{t+1} - \boldsymbol{v}^t \right\rangle \mathrm{d}a \\
&\qquad\qquad + \int_0^1 \left\langle \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t(a), \boldsymbol{w}^t(a)) - \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \boldsymbol{w}^{t+1/2} - \boldsymbol{w}^t \right\rangle \mathrm{d}a \\
&\le f(\boldsymbol{v}^t, \boldsymbol{w}^t) - \left(\eta - \frac{L\eta^2}{2}\right) \left\| \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t, \boldsymbol{w}^t) \right\|^2 \\
&\quad - \eta \left\langle \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \mathbb{E}_{\mathbf{Z}} \left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z}) \right] \right\rangle + \frac{L\eta^2}{2} \left\| \mathbb{E}_{\mathbf{Z}} \left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z}) \right] \right\|^2 \\
&\le f(\boldsymbol{v}^t, \boldsymbol{w}^t) - \left(\eta - (1+A)\frac{L\eta^2}{2}\right) \left\| \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t, \boldsymbol{w}^t) \right\|^2 \\
&\quad - \left(\eta - \frac{AL\eta^2}{2}\right) \left\langle \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \mathbb{E}_{\mathbf{Z}} \left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z}) \right] \right\rangle . \qquad (26)
\end{aligned}$$

The third equality is due to the fundamental theorem of calculus. In the first inequality, we called Lemma 3 for $(\boldsymbol{v}^t, \boldsymbol{w}^t)$ and $(\boldsymbol{v}^t(a), \boldsymbol{w}^t(a))$ with $a \in [0,1]/\{a_0\}$. In the last inequality, we used Lemma 5. So when $\eta < \frac{2}{(1+A)L}$, we have $f(\boldsymbol{v}^{t+1}, \boldsymbol{w}^{t+1}) \le f(\boldsymbol{v}^t, \boldsymbol{w}^t)$.

Summing up the inequality (26) over t from 0 to ∞ and using $f \ge 0$, we have

$$\eta \sum_{t=0}^{\infty} \left(1 - (1+A)\frac{L\eta}{2}\right) \left\| \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t, \boldsymbol{w}^t) \right\|^2 + \left(1 - \frac{AL\eta}{2}\right) \left\langle \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \mathbb{E}_{\mathbf{Z}} \left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z}) \right] \right\rangle$$
$$\le f(\boldsymbol{v}^0, \boldsymbol{w}^0) < \infty.$$

Hence,

$$\lim_{t \to \infty} \left\| \frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{v}^t, \boldsymbol{w}^t) \right\| = 0$$

and

$$\lim_{t \to \infty} \left\langle \frac{\partial f}{\partial \boldsymbol{w}}(\boldsymbol{v}^t, \boldsymbol{w}^t), \mathbb{E}_{\mathbf{Z}} \left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z}) \right] \right\rangle = 0.$$

Invoking Lemma 5 again, we further have

$$\lim_{t \to \infty} \left\| \mathbb{E}_{\mathbf{Z}} \left[\boldsymbol{g}_{\mathrm{relu}}(\boldsymbol{v}^t, \boldsymbol{w}^t; \mathbf{Z}) \right] \right\| = 0.$$

Invoking Lemma 6, we have that coarse gradient descent with ReLU $\mu(x)$ only converges to a saddle point or a minimizer.

The second claim follows from Lemmas 7 and 8. □