Training activation quantized neural networks involves minimizing a piecewise constant function whose gradient vanishes almost everywhere, which is undesirable for the standard back-propagation or chain rule. 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 “gradient” through the modified chain rule becomes non-trivial. Since this unusual “gradient” is certainly not the gradient of loss function, the following question arises: why searching in its negative direction minimizes the training loss? In this paper, we provide the theoretical justification of the concept of STE by answering this question. We consider the problem of learning a two-linear-layer network with binarized ReLU activation and Gaussian input data. We shall refer to the unusual “gradient” given by the STE-modified chain rule as coarse gradient. The choice of STE is not unique. We prove that if the STE is properly chosen, the expected coarse gradient correlates positively with the population gradient (not available for the training), and its negation is a descent direction for minimizing the population loss. We further show the associated coarse gradient descent algorithm converges to a critical point of the population loss minimization problem. Moreover, we show that a poor choice of STE may lead to instability of the training algorithm near certain local minima, which is also observed in 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 substantial memory savings and energy efficiency at inference time, many recent efforts have been made to the training of coarsely quantized DNN, meanwhile maintaining the performance of their float counterparts (Courbariaux et al., 2015; Rastegari et al., 2016; Cai et al., 2017; Hubara et al., 2018; Yin et al., 2018).

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 $f(w)$ subject to a discrete set-constraint $w \in Q$ 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; Li et al., 2017; Yin et al., 2016; 2018a; Hou & Kwok, 2018; He et al., 2018; Li & Hao 2018). On the other hand, the gradient $\nabla f(w)$ 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). In the same paper, Bengio et al. (2013) proposed an alternative approach based on stochastic neurons. In addition, Friesen & Domingos (2017) proposed the feasible target propagation algorithm for learning hard-threshold (or binary activated) networks (Lee et al., 2015) via convex combinatorial optimization.

1.1 RELATED WORKS

The idea of STE originates to the celebrated perceptron algorithm (Rosenblatt, 1957; 1962) in 1950s for learning single-layer perceptrons. The perceptron algorithm essentially does not calculate the “gradient” through the standard chain rule, but instead through a modified chain rule in which the derivative of identity function serves as the proxy of the original derivative of binary output function \(1_{\{x > 0\}}\). Its convergence has been extensive discussed in the literature; see for examples, (Widrow & Lehr, 1990; Freund & Schapire, 1999) and the references therein. Hinton (2012) extended this idea to train multi-layer networks with binary activations (a.k.a. binary neuron), namely, to backpropagate as if the activation 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 \(\pm 1\), Hubara et al. (2016) substituted the derivative of the signum activation function with \(1_{\{|x| \leq 1\}}\) in the backward pass, known as the saturated STE. Later the idea of STE was readily employed 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; Yin et al., 2018b), where some other proxies took place including the derivatives of vanilla ReLU and clipped ReLU. Despite all the empirical success of STE, there is very limited theoretical understanding of it in training DNN with stair-case activations.

Goel et al. (2018) considers leaky ReLU activation of a one-hidden-layer network. They showed the convergence of the so-called Convertron algorithm, which uses the identity STE in the backward pass through the leaky ReLU layer. Other similar scenarios, where certain layers are not desirable for back-propagation, have 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 adversarial defense (Szegedy et al., 2013), (Athalye et al., 2018) introduced the backward pass differentiable approximation, which shares the same spirit as STE, and successfully broke defenses at ICLR 2018 that rely on obfuscated gradients.

1.2 MAIN CONTRIBUTIONS

Throughout this paper, we shall refer to the “gradient” of loss function w.r.t. the weight variables through the STE-modified chain rule as coarse gradient. The coarse gradient is certainly not the gradient of loss function, and it is generally not the gradient of any function. Why searching in its negative direction minimizes the training loss, since this is not the standard gradient descent? Apparently, the choice of STE is non-unique, then what makes a good STE? From the optimization perspective, we take a step towards understanding STE in training quantized ReLU nets by attempting these questions.

On the theoretical side, we consider three representative STEs for learning a two-linear-layer network with binary activation and Gaussian data: the derivatives of the identity function (Rosenblatt, 1957; Hinton, 2012; Goel et al., 2018), vanilla ReLU and the clipped ReLUs (Cai et al., 2017; Hubara et al., 2016). We adopt the model of population loss minimization (Brutzkus & Globerson, 2017; Tran, 2017; Li & Yuan, 2017; Du et al., 2018). For the first time, we prove that proper choices of STE give rise to training algorithms that are descent. Specifically, the negative expected coarse gradients based on STEs of the vanilla and clipped ReLUs are provably descent directions for the minimizing the population loss, which yield monotonically decreasing energy in the training. In contrast, this is not true for the identity STE. We further prove that the corresponding training algorithm can be unstable sometimes near certain local minima, because the coarse gradient may simply not vanish there.
We consider a model similar to (Du et al., 2018) that outputs the prediction
\[ y(Z, v, w) := \sum_{i=1}^{m} v_i \sigma(Z_i^\top w) = v^\top \sigma(Zw) \]
for some input \( Z \in \mathbb{R}^{m \times n} \). Here \( w \in \mathbb{R}^n \) and \( v \in \mathbb{R}^m \) are the trainable weights in the first and second linear layers, respectively; \( Z_i^\top \) denotes the \( i \)th row vector of \( Z \); the activation function \( \sigma \) acts component-wise on the vector \( Zw \), i.e., \( \sigma(Zw)_i = \sigma((Zw)_i) = \sigma(Z_i^\top w) \). The first layer serves as a convolutional layer, where each row \( Z_i^\top \) can be viewed as a patch sampled from \( Z \) and the weight vector \( w \) is shared among all patches, and the second linear layer is the classifier. The label is generated according to \( y^*(Z) = (v^*)^\top \sigma(Zw^*) \) for some true (non-zero) parameters \( v^* \) and \( w^* \). Moreover, we use the following squared sample loss
\[
\ell(v, w; Z) := \frac{1}{2} (y(Z, v, w) - y^*(Z))^2 = \frac{1}{2} (v^\top \sigma(Zw) - y^*(Z))^2.
\]
Unlike in (Du et al., 2018), the activation function \( \sigma \) here is not ReLU, but the binary function \( \sigma(x) = 1_{\{x > 0\}} \).

We assume that the entries of \( Z \in \mathbb{R}^{m \times n} \) are i.i.d. sampled from the Gaussian distribution \( \mathcal{N}(0, 1) \) (Zhong et al., 2017; Brutzkus & Globerson, 2017). Since \( \ell(v, w; Z) = \ell(v, w/c; Z) \) for any scalar \( c > 0 \), without loss of generality, we take \( \|w^*\| = 1 \) and cast the learning task as the following population loss minimization problem:
\[
\min_{v \in \mathbb{R}^m, w \in \mathbb{R}^n} f(v, w) := \mathbb{E}_Z [\ell(v, w; Z)],
\]
where the sample loss \( \ell(v, w; Z) \) is given by (1).
2.1 Back-propagation and Coarse Gradient Descent

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

$$\nabla f(v, w) := \left[ \frac{\partial f}{\partial v}(v, w) \right] \cdot$$

The gradient of objective function, however, is not available for the network training. In fact, we can only access the expected sample gradient, namely,

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

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

$$\frac{\partial \ell}{\partial v}(v, w; Z) = \sigma'(Zw)(v^\top \sigma(Zw) - y^*(Z))$$

and

$$\frac{\partial \ell}{\partial w}(v, w; Z) = Z^\top (\sigma'(Zw) \odot v)(v^\top \sigma(Zw) - y^*(Z)).$$

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

$$g_\mu(v, w; Z) = Z^\top (\mu'(Zw) \odot v)(v^\top \sigma(Zw) - y^*(Z)).$$

Using the STE $\mu'$ to train the two-linear-layer convolutional neural network (CNN) with binary activation gives rise to the (full-batch) coarse gradient descent described in Algorithm 1.

Algorithm 1 Coarse gradient descent for training two-linear-layer CNN with STE $\mu'$.

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

for $t = 0, 1, \ldots$ do

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

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

end for

2.2 Preliminaries

Let us present some preliminaries about the landscape of the population loss function $f(v, w)$. To this end, we define the angle between $w$ and $w^*$ as $\theta(w, w^*) := \arccos \left( \frac{w^\top w^*}{\|w\| \|w^*\|} \right)$ for any $w \neq 0_n$. Recall that the label is given by $y^*(Z) = (w^*)^\top Z w^*$ from (1), we elaborate on the analytic expressions of $f(v, w)$ and $\nabla f(v, w)$.

Lemma 1. If $w \neq 0_n$, the population loss $f(v, w)$ is given by

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

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

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

$$\frac{\partial f}{\partial v}(v, w) = \frac{1}{4} (I_m + 1_m 1_m^\top) v - \frac{1}{4} \left( \left( 1 - \frac{2}{\pi} \theta(w, w^*) \right) I_m + 1_m 1_m^\top \right) v^* \quad (6)$$
and
\[
\frac{\partial f}{\partial w}(v, w) = -\frac{v^\top v^* + (I_n - \frac{vw^\top}{\|w\|^2})w^*}{2\|w\|\|I_n - \frac{vw^\top}{\|w\|^2}w^*\|},
\]
respectively.

For any \(v \in \mathbb{R}^m\), \((v, 0_m)\) 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.,
\[
v^\top v^* = 0 \quad \text{and} \quad v = (I_n + 1_m 1_m^\top)^{-1} \left( \left( 1 - \frac{2}{\pi} \theta(w, w^*) \right) I_n + 1_m 1_m^\top \right) v^*.
\]

2. Non-differentiable points where \(\theta(w, w^*) = 0\) and \(v = v^*\), or \(\theta(w, w^*) = \pi\) and \(v = (I_n + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^*\).

Among them, \(\{(v, w) : v = v^*, \theta(w, w^*) = 0\}\) are obviously the global minimizers of \((7)\). We show that the stationary points, if exist, can only be saddle points, and \(\{(v, w) : \theta(w, w^*) = \pi, v = (I_n + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^*\}\) are the only potential spurious local minimizers.

**Proposition 1.** If the true parameter \(v^*\) satisfies \((1_m, v^*)^2 < \frac{m+1}{2} \|v^*\|^2\), then
\[
\{(v, w) : v = (I_n + 1_m 1_m^\top)^{-1} \left( \frac{1}{(m+1)} \|v^*\|^2 - \frac{1}{(1_m, v^*)^2} I_n + 1_m 1_m^\top \right) v^*, \theta(w, w^*) = \pi, v = (I_n + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^* \}
\]
give the saddle points obeying \((8)\), and \(\{(v, w) : \theta(w, w^*) = \pi, v = (I_n + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^* \}\) are the spurious local minimizers. Otherwise, the model \((2)\) has no saddle points or spurious local minimizers.

We further prove that the population gradient \(\nabla f(v, w)\) given by \((6)\) and \((7)\), is Lipschitz continuous when restricted to bounded domains.

**Lemma 3.** For any differentiable points \((v, w)\) and \((\hat{v}, \hat{w})\) with \(\min\{\|v\|, \|\hat{v}\|\} = c_w > 0\) and \(\max\{\|v\|, \|\hat{v}\|\} = C_w\), there exists a Lipschitz constant \(L > 0\) depending on \(C_w\) and \(c_w\), such that
\[
\|\nabla f(v, w) - \nabla f(\hat{v}, \hat{w})\| \leq L\|\nabla f(v, w)\| \leq \|f(v, w) - f(\hat{v}, \hat{w})\|.
\]

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 and clipped ReLUs as well as the identity function serve as the STE, respectively. We shall prove that Algorithm \([1]\) using the derivative of vanilla or clipped ReLU converges to a critical point, whereas that with the identity STE does not.

**Theorem 1** (Convergence). Let \(\{(v^t, w^t)\}\) be the sequence generated by Algorithm \([7]\) with ReLU \(\mu(x) = \max\{x, 0\}\) or clipped ReLU \(\mu(x) = \min\{\max\{x, 0\}, 1\}\). Suppose \(\|w^t\| \geq c_w\) for all \(t\) with some \(c_w > 0\).

Then if the learning rate \(\eta > 0\) is sufficiently small, for any initialization \((v^0, w^0)\), the objective sequence \(\{f(v^t, w^t)\}\) is monotonically decreasing, and \(\{(v^t, w^t)\}\) converges to a saddle point or a (local) minimizer of the population loss minimization \((2)\).

In addition, if \(1_m v^* \neq 0\) and \(m > 1\), the descent and convergence properties do not hold for Algorithm \([7]\) with the identity function \(\mu(x) = x\) near the local minimizers satisfying \(\theta(w, w^*) = \pi\) and \(v = (I_n + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^*\).

**Remark 1.** The convergence guarantee for the coarse gradient descent 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 roughly descends along the direction of negative coarse gradient, as illustrated by Figure \([7]\). As the sample size increases, the empirical loss gains monotonicity and smoothness. This explains why (proper) STE works so well with massive amounts of data as in deep learning.
Remark 2. The same results hold, if the Gaussian assumption on the input data is weakened to that their rows i.i.d. follow some rotation-invariant distribution. The proof will be substantially similar.

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

![Sample size plots](image)

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) \(\eta\) 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 \(\frac{\partial f}{\partial \nu}(v, w; Z)\), it is easy to see \(\mu'(x) = \sigma(x)\), and we have the following expressions of \(\mathbb{E}_Z\left[\frac{\partial \ell}{\partial v}(v, w; Z)\right]\) and \(\mathbb{E}_Z[g_{\text{relu}}(v, w; Z)]\) for Algorithm 1.

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

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

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

\[
\mathbb{E}_Z\left[g_{\text{relu}}(v, w; Z)\right] = h(v, v^*) \frac{w}{2\sqrt{2\pi}} ||w|| - \cos\left(\frac{\theta(w, w^*)}{2}\right) \frac{v^T v^*}{\sqrt{2\pi}} \frac{w}{||w||} + w^* \tag{11}
\]

where \(h(v, v^*) = ||v||^2 + (1_m^T v)^2 - (1_m^T v)(1_m^T v^*) + v^T v^*.\)

As stated in Lemma 5 below, the key observation is that the coarse partial gradient \(\mathbb{E}_Z\left[g_{\text{relu}}(v, w; Z)\right]\) has non-negative correlation with the population partial gradient \(\frac{\partial f}{\partial \nu}(v, w)\), and \(-\mathbb{E}_Z[g_{\text{relu}}(v, w; Z)]\) together with \(-\mathbb{E}_Z\left[\frac{\partial \ell}{\partial v}(v, w; Z)\right]\) form a descent direction for minimizing the population loss.

**Lemma 5.** If \(w \neq 0_v\), and \(\theta(w, w^*) \in (0, \pi)\), then the inner product between the expected coarse and population gradients w.r.t. \(w\) is

\[
\left\langle \mathbb{E}_Z\left[g_{\text{relu}}(v, w; Z)\right], \frac{\partial f}{\partial w}(v, w) \right\rangle = \frac{\sin(\theta(w, w^*))}{2(\sqrt{2\pi})^3 ||w||}(v^T v^*)^2 \geq 0.
\]

Moreover, if further \(||v|| \leq C_v\) and \(||w|| \geq c_w\), there exists a constant \(A_{\text{relu}} > 0\) depending on \(C_v\) and \(c_w\), such that

\[
\left\| \mathbb{E}_Z\left[g_{\text{relu}}(v, w; Z)\right] \right\|^2 \leq A_{\text{relu}} \left( \left\| \frac{\partial f}{\partial w}(v, w) \right\|^2 + \left\langle \mathbb{E}_Z\left[g_{\text{relu}}(v, w; Z)\right], \frac{\partial f}{\partial w}(v, w) \right\rangle \right). \tag{12}
\]

Clearly, when \(\left\langle \mathbb{E}_Z\left[g_{\text{relu}}(v, w; Z)\right], \frac{\partial f}{\partial \nu}(v, w) \right\rangle > 0\), \(\mathbb{E}_Z\left[g_{\text{relu}}(v, w; Z)\right]\) is roughly in the same direction as \(\frac{\partial f}{\partial \nu}(v, w)\). Moreover, since by Lemma 4, \(\mathbb{E}_Z\left[\frac{\partial \ell}{\partial v}(v, w; Z)\right] = \frac{\partial f}{\partial v}(v, w)\), we expect

\footnote{We redefine the second term as \(0_v\), in the case \(\theta(w, w^*) = \pi\), or equivalently, \(w^* \parallel w\) + \(w^* = 0_v\).}
that the coarse gradient descent behaves like the gradient descent directly on $f(v, w)$. Here we would like to highlight the significance of the estimate (12) in guaranteeing the descent property of Algorithm 1. By the Lipschitz continuity of $\nabla f$ specified in Lemma 3 it holds that

$$f(v^{i+1}, w^{i+1}) - f(v^i, w^i) \leq \left< \frac{\partial f}{\partial v}(v^i, w^i), v^{i+1} - v^i \right> + \left< \frac{\partial f}{\partial w}(v^i, w^i), w^{i+1} - w^i \right> + \frac{L}{2} \left( \|v^{i+1} - v^i\|^2 + \|w^{i+1} - w^i\|^2 \right)$$

$$= - \left( \eta - \frac{L\eta^2}{2} \right) \left\| \frac{\partial f}{\partial v}(v^i, w^i) \right\|^2 + \frac{L\eta^2}{2} \left\| \mathbb{E}[g_{\text{relu}}(v^i, w^i; Z)] \right\|^2 - \eta \left< \frac{\partial f}{\partial v}(v^i, w^i), \mathbb{E}[g_{\text{relu}}(v^i, w^i; Z)] \right>$$

$$\leq \left( \eta - (1 + A_{\text{relu}}) \frac{L\eta^2}{2} \right) \left\| \frac{\partial f}{\partial v}(v^i, w^i) \right\|^2 - \left( \eta - A_{\text{relu}} \frac{L\eta^2}{2} \right) \left< \frac{\partial f}{\partial w}(v^i, w^i), \mathbb{E}[g_{\text{relu}}(v^i, w^i; Z)] \right>,$$

(13)

where a) is due to (12). Therefore, if $\eta$ is small enough, we have monotonically decreasing energy until convergence.

**Lemma 6.** When Algorithm 1 converges, $\mathbb{E}[\frac{\partial f}{\partial v}(v, w; Z)]$ and $\mathbb{E}[g_{\text{relu}}(v, w; Z)]$ vanish simultaneously, which only occurs at the

1. Saddle points where (8) is satisfied according to Proposition 7
2. Minimizers of (2) where $v = v^*, \theta(w, w^*) = 0$, or $v = (I_m + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^*$.

Lemma 6 states that when Algorithm 1 using ReLU STE converges, it can only converge to a critical point of the population loss function.

### 3.2 Derivative of the Clipped ReLU as STE

For the STE using clipped ReLU, $\mu(x) = \min \{\max\{x, 0\}, 1\}$ and $\mu'(x) = 1_{\{0 < x < 1\}}(x)$. We have results similar to Lemmas 5 and 6. That is, the coarse partial gradient using clipped ReLU STE $\mathbb{E}[g_{\text{relu}}(v, w; Z)]$ generally has positive correlation with the true partial gradient of the population loss $\frac{\partial f}{\partial w}(v, w)$ (Lemma 7). Moreover, the coarse gradient vanishes and only vanishes at the critical points (Lemma 8).

**Lemma 7.** If $w \neq 0$, and $\theta(w, w^*) \in (0, \pi)$, then

$$\mathbb{E}[g_{\text{relu}}(v, w; Z)] = \frac{p(0, w)h(v, v^*)}{2 \|w\|} \frac{w}{\|w\|} - (v^\top v^*) \csc(\theta/2) \cdot q(\theta, w) \frac{w}{\|w\|} + w^*$$

$$- (v^\top v^*) (p(\theta, w) - \cot(\theta/2) \cdot q(\theta, w)) \frac{w}{\|w\|},$$

where $h(v, v^*) := \|v\|^2 + (1_m^\top v)^2 - (1_m^\top v)(1_m^\top v^*) + v^\top v^* same as in Lemma 5 and p(\theta, w) := \frac{1}{2\pi} \int_{-\frac{\pi}{2} + \theta}^{\frac{\pi}{2}} \cos(\phi) \xi \left( \frac{\sec(\phi)}{\|w\|} \right) d\phi$, $q(\theta, w) := \frac{1}{2\pi} \int_{-\frac{\pi}{2} + \theta}^{\frac{\pi}{2}} \sin(\phi) \xi \left( \frac{\sec(\phi)}{\|w\|} \right) d\phi$ with $\xi(x) := \int_0^x \exp(-\frac{r^2}{2}) dr$. The inner product between the expected coarse and true gradients w.r.t. $w$

$$\left< \mathbb{E}[g_{\text{relu}}(v, w; Z)], \frac{\partial f}{\partial w}(v, w) \right> = \frac{q(\theta, w)}{2\pi\|w\|} (v^\top v^*)^2 \geq 0.$$
Moreover, if further \( \|v\| \leq C_v \) and \( \|w\| \geq c_w \), there exists a constant \( A_{\text{crelu}} > 0 \) depending on \( C_v \) and \( c_w \), such that
\[
\left\| \mathbb{E}_Z [g_{\text{crelu}}(v, w; Z)] \right\|^2 \leq A_{\text{crelu}} \left( \left\| \frac{\partial f}{\partial v}(v, w) \right\|^2 + \left\langle \mathbb{E}_Z [g_{\text{crelu}}(v, w; Z)], \frac{\partial f}{\partial w}(v, w) \right\rangle \right).
\]

**Lemma 8.** When Algorithm 1 converges, \( \mathbb{E}_Z \left[ \frac{\partial f}{\partial v}(v, w; Z) \right] \) and \( \mathbb{E}_Z [g_{\text{crelu}}(v, w; Z)] \) vanish simultaneously, which only occurs at the

1. Saddle points where (8) is satisfied according to Proposition 4
2. Minimizers of (2) where \( v = v^*, \theta(w, w^*) = 0 \), or \( v = (I_m + 1_m1_m^\top)^{-1}(1_m1_m^\top - I_m)v^* \), \( \theta(w, w^*) = \pi \).

### 3.3 Derivative of the Identity Function as STE

Now we consider the derivative of identity function. 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 local minima, and Algorithm 1 may never converge there.

**Lemma 9.** Let \( \mu(x) = x \) in (5). Then the expected coarse partial gradient w.r.t. \( w \) is
\[
\mathbb{E}_Z [g_{\text{id}}(v, w; Z)] = \frac{1}{\sqrt{2\pi}} \left( \|v\|^2 \frac{\|w\|}{\|v\|} - (v^\top v^*)w^* \right).
\]

If \( \theta(w, w^*) = \pi \) and \( v = (I_m + 1_m1_m^\top)^{-1}(1_m1_m^\top - I_m)v^* \),
\[
\left\| \mathbb{E}_Z [g_{\text{id}}(v, w; Z)] \right\| = \frac{2(m-1)}{\sqrt{2\pi}(m+1)^2}(1_m^\top v^*)^2 \geq 0,
\]

i.e., \( \mathbb{E}_Z [g_{\text{id}}(v, w; Z)] \) does not vanish at the local minimizers if \( 1_m^\top v^* \neq 0 \) and \( m > 1 \).

**Lemma 10.** If \( w \neq 0_m \) and \( \theta(w, w^*) \in (0, \pi) \), then the inner product between the expected coarse and true gradients w.r.t. \( w \) is
\[
\left\langle \mathbb{E}_Z [g_{\text{id}}(v, w; Z)], \frac{\partial f}{\partial w}(v, w) \right\rangle = \frac{\sin(\theta(w, w^*))}{(\sqrt{2\pi})^3\|w\|}(v^\top v^*)^2 \geq 0.
\]

When \( \theta(w, w^*) \to \pi, v \to (I_m + 1_m1_m^\top)^{-1}(1_m1_m^\top - I_m)v^* \), if \( 1_m^\top v^* \neq 0 \) and \( m > 1 \), we have
\[
\frac{\left\| \mathbb{E}_Z [g_{\text{id}}(v, w; Z)] \right\|^2}{\left\| \frac{\partial f}{\partial w}(v, w) \right\|^2 + \left\langle \mathbb{E}_Z [g_{\text{id}}(v, w; Z)], \frac{\partial f}{\partial w}(v, w) \right\rangle} \to +\infty.
\]

Lemma 9 suggests that if \( 1_m^\top v^* \neq 0 \), the coarse gradient descent will never converge near the spurious minimizers with \( \theta(w, w^*) = \pi \) and \( v = (I_m + 1_m1_m^\top)^{-1}(1_m1_m^\top - I_m)v^* \), because \( \mathbb{E}_Z [g_{\text{id}}(v, w; Z)] \) does not vanish there. By the positive correlation implied by (15) of Lemma 10 for some proper \( (v^0, w^0) \), the iterates \( \{v^t, w^t\} \) may move towards a local minimizer in the beginning. But when \( \{v^t, w^t\} \) approaches it, the descent property (13) does not hold for \( \mathbb{E}_Z [g_{\text{id}}(v, w; Z)] \) because of (16), hence the training loss begins to increase and instability arises.

### 4 Experiments

While our theory implies that both vanilla and clipped ReLUs learn a two-linear-layer CNN, their empirical performances on deeper nets are different. In this section, we compare the performances of the identity, ReLU and clipped ReLU STEs on MNIST [LeCun et al. 1998] and CIFAR-10 (Krizhevsky 2009) benchmarks for 2-bit or 4-bit quantized activations. As an illustration, we plot...
the 2-bit quantized ReLU and its associated clipped ReLU in Figure 3 in the appendix. Intuitively, the clipped ReLU should be the best performer, as it best approximates the original quantized ReLU. We also report the instability issue of the training algorithm when using an improper STE in section 4.2. In all experiments, the weights are kept float.

The resolution $\alpha$ 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 unit Gaussian distribution. Then the $\alpha$ that fits the input of activation layer the best can be pre-computed by a variant of Lloyd’s algorithm (Lloyd, 1982; Yin et al., 2018a) applied to a set of simulated 1-D half-Gaussian data. After determining the $\alpha$, 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. 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.

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.

### 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 relatively shallow LeNet-5 network, vanilla ReLU exhibits comparable performance to the clipped ReLU, which is somewhat in line with our theoretical finding that ReLU is a great STE for learning the two-linear-layer (shallow) CNN.

<table>
<thead>
<tr>
<th>Network</th>
<th>BitWidth</th>
<th>Straight-through estimator</th>
<th>identity</th>
<th>vanilla ReLU</th>
<th>clipped ReLU</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LeNet5</td>
<td>2</td>
<td>$2.6 \times 10^{-2}$/98.49</td>
<td>5.1 $\times 10^{-4}$/99.24</td>
<td>5.4 $\times 10^{-9}$/99.23</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$6.0 \times 10^{-3}$/98.98</td>
<td>9.0 $\times 10^{-4}$/99.32</td>
<td>8.8 $\times 10^{-4}$/99.24</td>
<td></td>
</tr>
<tr>
<td>CIFAR10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VGG11</td>
<td>2</td>
<td>0.19/86.58</td>
<td>0.10/88.69</td>
<td>0.02/90.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$3.1 \times 10^{-3}$/90.19</td>
<td>1.5 $\times 10^{-3}$/92.01</td>
<td>1.3 $\times 10^{-3}$/92.08</td>
<td></td>
</tr>
<tr>
<td>ResNet20</td>
<td>2</td>
<td>1.56/46.52</td>
<td>1.50/48.05</td>
<td>0.24/88.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.38/54.16</td>
<td>0.25/86.59</td>
<td>0.04/91.24</td>
<td></td>
</tr>
</tbody>
</table>

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.

### 4.2 Instability

We report the phenomenon of being repelled from a good minimum on ResNet-20 with 4-bit activations when using the identity STE, to demonstrate the instability issue as predicted in Theorem 1. By Table 1 the coarse gradient descent algorithms using the 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; see Figure 2. To speed up training, at epoch 20, we switch to the normal schedule of learning rate specified in Table 2 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 (Lemma 9). Similarly, the poor performance of ReLU STE on 2-bit activated ResNet-20
is also due to the instability of the corresponding training algorithm at good minima, as illustrated by Figure 4 in Appendix C, although it diverges much slower.

![Figure 2](image.png)

Figure 2: When initialized with weights (good minima) produced by the vanilla (orange) and clipped (blue) ReLUs on ResNet-20 with 4-bit activations, 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 that it gives rise to descent training algorithm. We considered three STEs: the derivatives of the identity function, vanilla ReLU and clipped ReLU, for learning a two-linear-layer CNN with binary activation. We derived the explicit formulas of the expected coarse gradients corresponding to the STEs, and showed that the negative expected coarse gradients based on vanilla and clipped ReLUs are descent directions for minimizing the population loss, whereas the identity STE is not since it generates a coarse gradient incompatible with the energy landscape. The instability/incompatibility issue was confirmed in CIFAR experiments for improper choices of STE. In the future work, we aim further understanding of coarse gradient descent for large-scale optimization problems with intractable gradients.

ACKNOWLEDGMENTS

This work was partially supported by NSF grants DMS-1522383, IIS-1632935, ONR grant N00014-18-1-2527, AFOSR grant FA9550-18-0167, DOE grant DE-SC0013839 and STROBE STC NSF grant DMR-1548924.

REFERENCES


APPENDIX

A. THE PLOTS OF QUANTIZED AND CLIPPED ReLUs

![Quantized and Clipped ReLUs plots](image)

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 $\tilde{\sigma}_\alpha(x)$. $\alpha$ is the resolution determined in advance of the network training.

B. THE SCHEDULE OF LEARNING RATE

<table>
<thead>
<tr>
<th>Network</th>
<th># epochs</th>
<th>Batch size</th>
<th>Learning rate</th>
<th>milestone</th>
</tr>
</thead>
<tbody>
<tr>
<td>LeNet5</td>
<td>50</td>
<td>64</td>
<td>0.1 0.1</td>
<td>[20,40]</td>
</tr>
<tr>
<td>VGG11</td>
<td>200</td>
<td>128</td>
<td>0.01 0.1</td>
<td>[80,140]</td>
</tr>
<tr>
<td>ResNet20</td>
<td>200</td>
<td>128</td>
<td>0.01 0.1</td>
<td>[80,140]</td>
</tr>
</tbody>
</table>

Table 2: The schedule of the learning rate.

C. INSTABILITY OF ReLU STE ON ResNet-20 WITH 2-BIT ACTIVATIONS

![Validation accuracy and training loss plots](image)

Figure 4: When initialized with the weights produced by the clipped ReLU STE on ResNet-20 with 2-bit activations (88.38% validation accuracy), the coarse gradient descent using the ReLU STE with $10^{-5}$ learning rate is not stable there, and both classification and training errors begin to increase.

D. ADDITIONAL SUPPORTING LEMMAS

**Lemma 11.** Let $z \in \mathbb{R}^n$ be a Gaussian random vector with entries i.i.d. sampled from $\mathcal{N}(0, 1)$. Given nonzero vectors $w, \tilde{w} \in \mathbb{R}^n$ with the angle $\theta$, we have

$$
\mathbb{E} \left[ 1_{\{z^\top w > 0\}} \right] = \frac{1}{2}, \\
\mathbb{E} \left[ 1_{\{z^\top w > 0, z^\top \tilde{w} > 0\}} \right] = \frac{\pi - \theta}{2\pi}.
$$
Given nonzero vectors $w$, Lemma 12. Let $z \in \mathbb{R}^n$ be a Gaussian random vector with entries i.i.d. sampled from $N(0,1)$. Given nonzero vectors $w$, $\tilde{w} \in \mathbb{R}^n$ with the angle $\theta$, we have

$$
\mathbb{E} [z, w]) = \left( (p(\theta, w) - \cot(\theta/2) \cdot q(\theta, w)) \frac{w}{\|w\|} + \csc(\theta/2) \cdot q(\theta, w) \frac{w}{\|w\|}\right)
$$

Here

$$
p(\theta, w) = \frac{1}{2\pi} \int_{-\frac{\theta}{2}}^{\frac{\theta}{2}} \cos(\phi) \xi \left( \frac{\sec(\phi)}{\|w\|} \right) d\phi,
q(\theta, w) = \frac{1}{2\pi} \int_{-\frac{\theta}{2}}^{\frac{\theta}{2}} \sin(\phi) \xi \left( \frac{\sec(\phi)}{\|w\|} \right) d\phi
$$

with $\xi(x) = \int_0^x r^2 \exp(-\frac{r^2}{2}) dr$ satisfying $\xi(+\infty) = \sqrt{\frac{2}{\pi}}$. In addition,

$$
p(\pi, w) = q(0, w) = q(\pi, w) = 0.
$$

Proof of Lemma 11. 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 $w = [w_1, 0_{n-1}]^\top$ with $w_1 > 0$, then

$$
\mathbb{E} [1_{\{z^\top w > 0\}}] = \mathbb{P}(z_1 > 0) = \frac{1}{2}.
$$

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

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

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

$$
\mathbb{E} [z_1 1_{\{z^\top w > 0, z^\top \tilde{w} > 0\}}] = 0 \text{ for } i \geq 3. \text{ Moreover,}
$$

$$
\mathbb{E} [z_1 1_{\{z^\top w > 0, z^\top \tilde{w} > 0\}}] = \frac{1}{2\pi} \int_0^\infty r^2 \exp(-\frac{r^2}{2}) dr \int_{-\frac{\pi}{2} + \theta}^{\frac{\pi}{2}} \cos(\phi) d\phi = \frac{1 + \cos(\theta)}{2\sqrt{2\pi}}
$$

and

$$
\mathbb{E} [z_2 1_{\{z^\top w > 0, z^\top \tilde{w} > 0\}}] = \frac{1}{2\pi} \int_0^\infty r^2 \exp(-\frac{r^2}{2}) dr \int_{-\frac{\pi}{2} + \theta}^{\frac{\pi}{2}} \sin(\phi) d\phi = \frac{\sin(\theta)}{2\sqrt{2\pi}}.
$$

Therefore,

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

where the last equality holds because $\frac{w}{\|w\|}$ and $\frac{\sqrt{2}}{\sqrt{2\pi}} \frac{w}{\|w\|}$ are two unit-normed vectors with angle $\theta/2$. \qed
Proof of Lemma 13. Following the proof of Lemma 11 we let \(\vec{w} = [w_1, 0_{n-1}]^T\) with \(w_1 > 0\), \(\vec{w} = [\hat{w}_1, \hat{w}_2, 0_{n-2}]^T\), and use the polar representation of two-dimensional Gaussian distribution. Then

\[
\mathbb{E}[z_1^1\{0 < z^T w < 1\}] = \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos(\phi) \int_0^{\sec(\phi)/\|w\|} r^2 \exp\left(-\frac{r^2}{2}\right) \, dr \, d\phi = p(0, \vec{w})
\]

and

\[
\mathbb{E}[z_2^1\{0 < z^T w < 1\}] = \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin(\phi) \int_0^{\sec(\phi)/\|w\|} r^2 \exp\left(-\frac{r^2}{2}\right) \, dr \, d\phi = q(0, \vec{w}) = 0,
\]

since the integrand above is an odd function in \(\phi\). Moreover, \(\mathbb{E}[z_i^1\{0 < z^T w < 1\}] = 0\) for \(i \geq 3\). So the first identity holds. For the second one, we have

\[
\mathbb{E}[z_1^1\{0 < z^T w < 1, z^T \vec{w} > 0\}] = \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos(\phi) \int_0^{\sec(\phi)/\|w\|} r^2 \exp\left(-\frac{r^2}{2}\right) \, dr \, d\phi = p(\theta, \vec{w}),
\]

and similarly, \(\mathbb{E}[z_2^1\{0 < z^T w < 1, z^T \vec{w} > 0\}] = q(\theta, \vec{w})\). Therefore,

\[
\mathbb{E}[z^1\{0 < z^T w < 1, z^T \vec{w} > 0\}] = [p(\theta, \vec{w}), q(\theta, \vec{w}), 0_{n-2}^T]^T.
\]

Since \(\frac{\vec{w}}{\|\vec{w}\|} = [1, 0_{n-1}^T]^T\) and \(\frac{\vec{w} + \vec{w}}{\|\vec{w} + \vec{w}\|} = [\cos(\theta/2), \sin(\theta/2), 0_{n-2}^T]^T\), it is easy to check the second identity is also true. \(\square\)

Lemma 13. Let \(p(\theta, \vec{w})\) and \(q(\theta, \vec{w})\) be defined in Lemma 12. Then for \(\theta \in [\frac{\pi}{2}, \pi]\), we have

1. \(p(\theta, \vec{w}) \leq q(\theta, \vec{w})\).
2. \((1 - \frac{\theta}{\pi}) p(0, \vec{w}) \leq q(\theta, \vec{w})\).

Proof of Lemma 13: 1. Let \(\theta \in [\frac{\pi}{2}, \pi]\), since \(\xi \geq 0\),

\[
p(\theta, \vec{w}) = \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos(\phi) \xi \left(\sec(\phi) / \|w\|\right) \, d\phi = \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin(\frac{\pi}{2} - \phi) \xi \left(\sec(\phi) / \|w\|\right) \, d\phi
\]

\[
\leq \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin(\theta - \phi) \xi \left(\sec(\phi) / \|w\|\right) \, d\phi \leq \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin(\phi) \xi \left(\sec(\phi) / \|w\|\right) \, d\phi = q(\theta, \vec{w}),
\]

where the last inequality is due to the rearrangement inequality since both \(\sin(\phi)\) and \(\xi \left(\sec(\phi) / \|w\|\right)\) are increasing in \(\phi\) on \([0, \frac{\pi}{2}]\).

2. Since \(\cos(\phi) \xi \left(\sec(\phi) / \|w\|\right)\) is even, we have

\[
\frac{1}{\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos(\phi) \xi \left(\sec(\phi) / \|w\|\right) \, d\phi = \frac{2}{\pi} \int_0^{\frac{\pi}{2}} \cos(\phi) \xi \left(\sec(\phi) / \|w\|\right) \, d\phi
\]

\[
\leq \frac{2}{\pi} \int_0^{\frac{\pi}{2}} \sin(\phi) \xi \left(\sec(\phi) / \|w\|\right) \, d\phi \leq \frac{1}{\pi - \theta} \int_{\theta - \frac{\pi}{2}}^{\frac{\pi}{2}} \sin(\phi) \xi \left(\sec(\phi) / \|w\|\right) \, d\phi.
\]

The first inequality is due to part 1 which gives \(p(\pi/2, \vec{w}) \leq q(\pi/2, \vec{w})\), whereas the second one holds because \(\sin(\phi) \xi \left(\sec(\phi) / \|w\|\right)\) is increasing on \([0, \frac{\pi}{2}]\). \(\square\)

Lemma 14. For any nonzero vectors \(\vec{w}\) and \(\vec{w}\) with \(\|\vec{w}\| \geq \|\vec{w}\| = c > 0\), we have
1. $|\theta(w, w^*) - \theta(\tilde{w}, w^*)| \leq \frac{\pi}{2c} \|w - \tilde{w}\|.

2. $\left\| \frac{1}{\|w\|} \begin{pmatrix} I_n - \frac{ww^T}{\|w\|^2} \\ I_n - \frac{\tilde{w}\tilde{w}^T}{\|\tilde{w}\|^2} \end{pmatrix} w^* - \frac{1}{\|w\|} \begin{pmatrix} I_n - \frac{\tilde{w}\tilde{w}^T}{\|\tilde{w}\|^2} \\ I_n - \frac{\tilde{w}\tilde{w}^T}{\|\tilde{w}\|^2} \end{pmatrix} \tilde{w}^* \right\| \leq \frac{1}{2c} \|w - \tilde{w}\|.$

**Proof of Lemma 14** 1. Since by Cauchy-Schwarz inequality,
$$\left\langle \tilde{w}, w - \frac{cw}{\|w\|} \right\rangle = \tilde{w}^T w - c\|\tilde{w}\| \leq 0,$$
we have
$$\|\tilde{w} - w\|^2 = \left\| \left(1 - \frac{c}{\|w\|}\right) \tilde{w} - \left(w - \frac{cw}{\|w\|}\right) \right\|^2 \geq \left\| \left(1 - \frac{c}{\|w\|}\right) \tilde{w} \right\|^2 + \left\| w - \frac{cw}{\|w\|} \right\|^2 \geq \left\| w - \frac{cw}{\|w\|} \right\|^2 = c^2 \left\| w - \frac{cw}{\|w\|} \right\|^2.$$
Therefore,
$$|\theta(w, w^*) - \theta(\tilde{w}, w^*)| \leq \theta(w, \tilde{w}) = \theta\left(\frac{w}{\|w\|}, \frac{\tilde{w}}{\|\tilde{w}\|}\right) \leq \pi \sin\left(\frac{\theta\left(\frac{w}{\|w\|}, \frac{\tilde{w}}{\|\tilde{w}\|}\right)}{2}\right) = \frac{\pi}{2} \left\| \frac{w}{\|w\|} - \frac{\tilde{w}}{\|\tilde{w}\|} \right\| \leq \frac{\pi}{2c} \|w - \tilde{w}\|,$$
where we used the fact $\sin(x) \geq x$ for $x \in [0, \frac{\pi}{2}]$ and the estimate in (17).

2. Since $\left( I_n - \frac{ww^T}{\|w\|^2} \right) w^*$ is the projection of $w^*$ onto the complement space of $w$, and likewise for $\left( I_n - \frac{\tilde{w}\tilde{w}^T}{\|\tilde{w}\|^2} \right) w^*$, the angle between $\left( I_n - \frac{ww^T}{\|w\|^2} \right) w^*$ and $\left( I_n - \frac{\tilde{w}\tilde{w}^T}{\|\tilde{w}\|^2} \right) w^*$ is equal to the angle between $w$ and $\tilde{w}$. Therefore,
$$\left\langle \left( I_n - \frac{ww^T}{\|w\|^2} \right) w^*, \left( I_n - \frac{\tilde{w}\tilde{w}^T}{\|\tilde{w}\|^2} \right) w^* \right\rangle = \left\langle w, \tilde{w} \right\rangle,$$
and thus
$$\left\| \frac{1}{\|w\|} \begin{pmatrix} I_n - \frac{ww^T}{\|w\|^2} \\ I_n - \frac{\tilde{w}\tilde{w}^T}{\|\tilde{w}\|^2} \end{pmatrix} w^* - \frac{1}{\|w\|} \begin{pmatrix} I_n - \frac{\tilde{w}\tilde{w}^T}{\|\tilde{w}\|^2} \\ I_n - \frac{\tilde{w}\tilde{w}^T}{\|\tilde{w}\|^2} \end{pmatrix} \tilde{w}^* \right\| = \left\| \frac{w - \tilde{w}}{\|w\|^2} \right\| \leq \frac{1}{2c} \|w - \tilde{w}\|.$$

The second equality above holds because
$$\left\| \frac{w}{\|w\|^2} - \frac{\tilde{w}}{\|\tilde{w}\|^2} \right\|^2 = \frac{1}{\|w\|^2} + \frac{1}{\|\tilde{w}\|^2} - 2\frac{\langle w, \tilde{w} \rangle}{\|w\|^2\|\tilde{w}\|^2} = \frac{\|w - \tilde{w}\|^2}{\|w\|^2\|\tilde{w}\|^2}.$$

**E. Main Proofs**

**Lemma 1.** If $w \neq 0_n$, the population loss $f(v, w)$ is given by
$$\frac{1}{8} \left( v^\top \left( I_m + 1_m 1_m^\top \right) v - 2v^\top \left( I_m + 1_m 1_m^\top \right) v^* + v^*^\top \left( I_m + 1_m 1_m^\top \right) v^* \right).$$

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

16
Proof of Lemma 1 We notice that
\[
\begin{align*}
f(v, w) &= \frac{1}{2} (v^\top \mathbb{E}_Z [\sigma(Zw)^\top \sigma(Zw)]v - 2v^\top \mathbb{E}_Z [\sigma(Zw)^\top \sigma(Zw^*)]v^* \\
&\quad + (v^*)^\top \mathbb{E}_Z [\sigma(Zw^*)^\top \sigma(Zw^*)]v^*).
\end{align*}
\]
Let \( Z_i^\top \) be the \( i \)-th row vector of \( Z \). Since \( w \neq 0_n \), using Lemma 1 we have
\[
\mathbb{E}_Z [\sigma(Zw)^\top \sigma(Zw^*)]_{ii} = \mathbb{E}_Z [\sigma(Z_i^\top w)^\top \sigma(Z_i^\top w)] = \mathbb{E} [1_{\{Z_i^\top w > 0\}}] = \frac{1}{2},
\]
and for \( i \neq j \),
\[
\mathbb{E}_Z [\sigma(Zw)^\top \sigma(Zw^*)]_{ij} = \mathbb{E}_Z [\sigma(Z_i^\top w)^\top \sigma(Z_j^\top w)] = \mathbb{E} [1_{\{Z_i^\top w > 0\} \cdot Z_j^\top w^* > 0\}}] = \frac{1}{4}.
\]
Therefore, \( \mathbb{E}_Z [\sigma(Zw)^\top \sigma(Zw^*)] = \mathbb{E}_Z [\sigma(Zw^*)^\top \sigma(Zw^*)] = \frac{1}{4} (I_m + 1_m 1_m^\top) \). Furthermore,
\[
\mathbb{E}_Z [\sigma(Zw)^\top \sigma(Zw^*)]_{ii} = \mathbb{E} [1_{\{Z_i^\top w > 0, Z_i^\top w^* > 0\}}] = \frac{\pi - \theta(w, w^*)}{2\pi},
\]
and \( \mathbb{E}_Z [\sigma(Zw)^\top \sigma(Zw^*)]_{ij} = \frac{1}{4} \). So,
\[
\mathbb{E}_Z [\sigma(Zw)^\top \sigma(Zw^*)] = \frac{1}{4} \left( 1 - \frac{2\theta(w, w^*)}{\pi} \right) I_m + 1_m 1_m^\top.
\]
Then it is easy to validate the first claim. Moreover, if \( w = 0_n \), then
\[
f(v, w) = \frac{1}{2} (v^*)^\top \mathbb{E}_Z [\sigma(Zw^*)^\top \sigma(Zw^*)]v^* = \frac{1}{8} (v^*)^\top (I_m + 1_m 1_m^\top) v^*.
\]

Lemma 2. If \( w \neq 0_n \) and \( \theta(w, w^*) \in (0, \pi) \), the partial gradients of \( f(v, w) \) w.r.t. \( v \) and \( w \) are
\[
\frac{\partial f}{\partial v} (v, w) = \frac{1}{4} (I_m + 1_m 1_m^\top) v - \frac{1}{4} \left( 1 - \frac{2}{\pi} \theta(w, w^*) \right) I_m + 1_m 1_m^\top v^*.
\]
and
\[
\frac{\partial f}{\partial w} (v, w) = -\frac{v^\top v^*}{2\pi \|w\|} \left( I_n - \frac{ww^\top}{\|w\|^2} \right) w^*.
\]
respectively.

Proof of Lemma 2 The first claim is trivial, and we only show the second one. Since \( \theta(w, w^*) = \arccos \left( \frac{w^\top w^*}{\|w\| \|w^*\|} \right) \) is differentiable w.r.t. \( w \) at \( \theta(w, w^*) \in (0, \pi) \), we have
\[
\frac{\partial f}{\partial w} (v, w) = \frac{v^\top v^*}{2\pi} \frac{\partial \theta}{\partial w} (w, w^*) = -\frac{v^\top v^*}{2\pi} \frac{\|w\|^2 w^* - (w^\top w^*) w}{\|w\|^3 \sqrt{1 - (\frac{w^\top w^*}{\|w\|^2})^2}}
\]
\[
= -\frac{v^\top v^*}{2\pi \|w\|} \left( I_n - \frac{ww^\top}{\|w\|^2} \right) w^*.
\]
Proposition 1. If the true parameter $v^*$ satisfies $(1_m^\top v^*)^2 < \frac{m+1}{2} \|v^*\|^2$, then

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

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

$$
0 = v^\top v^* = (v^*)^\top (I_m + 1_m 1_m^\top)^{-1} \left( \left( 1 - \frac{2}{\pi} \theta(w, w^*) \right) I_m + 1_m 1_m^\top \right) v^*. \quad (18)
$$
From (18) it follows that

$$
\frac{2}{\pi} \theta(w, w^*) (v^*)^\top (I_m + 1_m 1_m^\top)^{-1} v^* = (v^*)^\top (I_m + 1_m 1_m^\top)^{-1} (I_m + 1_m 1_m^\top) v^* = \|v^*\|^2. \quad (19)
$$
On the other hand, from (18) it also follows that

$$
\left( \left( \frac{2}{\pi} \theta(w, w^*) - 1 \right) \right) (v^*)^\top (I_m + 1_m 1_m^\top)^{-1} v^* = (v^*)^\top (I_m + 1_m 1_m^\top)^{-1} (1_m 1_m^\top) v^* = \frac{(1_m^\top v^*)^2}{m+1},
$$
where we used $(I_m + 1_m 1_m^\top) 1_m = (m+1) 1_m$. Taking the difference of the two equalities above gives

$$
(v^*)^\top (I_m + 1_m 1_m^\top)^{-1} v^* = \|v^*\|^2 - \frac{(1_m^\top v^*)^2}{m+1}. \quad (20)
$$
By (19), we have $\theta(w, w^*) = \pi \frac{(m+1)\|v^*\|^2}{2(m+1)\|v^*\|^2 - (1_m^\top v^*)^2}$, which requires

$$
\frac{\pi}{2} \frac{(m+1)\|v^*\|^2}{(m+1)\|v^*\|^2 - (1_m^\top v^*)^2} < \pi, \text{ or equivalently, } \frac{(1_m^\top v^*)^2}{m+1} < \frac{m+1}{2} \|v^*\|^2.
$$
Furthermore, since $\frac{\partial f}{\partial w}(v, w) = 0$, we have

$$
v = (I_m + 1_m 1_m^\top)^{-1} \left( \left( 1 - \frac{2}{\pi} \theta(w, w^*) \right) I_m + 1_m 1_m^\top \right) v^* = (I_m + 1_m 1_m^\top)^{-1} \left( \frac{-(1_m^\top v^*)^2}{(m+1)\|v^*\|^2 - (1_m^\top v^*)^2} I_m + 1_m 1_m^\top \right) v^*. \quad (21)
$$
Next, we check the local optimality of the stationary points. By ignoring the scaling and constant terms, we rewrite the objective function as

$$
\hat{f}(v, \theta) := v^\top (I_m + 1_m 1_m^\top) v - 2v^\top \left( \left( 1 - \frac{2}{\pi} \theta \right) I_m + 1_m 1_m^\top \right) v^*, \text{ for } \theta \in [0, \pi].
$$
It is easy to check that its Hessian matrix

$$
\nabla^2 \hat{f}(v, \theta) = \begin{bmatrix}
2(I_m + 1_m 1_m^\top) & \frac{4}{\pi} v^*\\
\frac{4}{\pi} v^\top & 0
\end{bmatrix}
$$
is indefinite. Therefore, the stationary points are saddle points.

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

$$
v^\top v^* = (v^*)^\top (I_m + 1_m 1_m^\top)^{-1} (1_m 1_m^\top - I_m) v^* = \|v^*\|^2 - 2(v^*)^\top (I_m + 1_m 1_m^\top)^{-1} v^* = \frac{2(1_m^\top v^*)^2}{m+1} - \|v^*\|^2 < 0,
$$
where we used (20) in the last identity. We consider an arbitrary point \((v + \Delta v, \pi + \Delta \theta)\) in the neighborhood of \((\bar{v}, \pi)\) with \(\Delta \theta \leq 0\). The perturbed objective value is

\[
\tilde{f}(v + \Delta v, \pi + \Delta \theta) = (v + \Delta v)\top (I_m + 1_m 1_m\top) (v + \Delta v) - 2(v + \Delta v)\top (1_m 1_m\top - I_m) v^* + \frac{2\Delta \theta}{\pi} (v + \Delta v)\top v^*.
\]

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

\[
(v + \Delta v)\top (I_m + 1_m 1_m\top) (v + \Delta v) - 2(v + \Delta v)\top (1_m 1_m\top - I_m) v^* > \tilde{f}(v, \pi).
\]

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

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

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

\[
v\top v^* = \frac{2(1_m\top v^*)^2}{m+1} - \|v^*\|^2 \geq 0.
\]

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

\[\]

Lemma 3. For any differentiable points \((v, w)\) and \((\bar{v}, \bar{w})\) with \(\min\{|w|, |\bar{w}|\} = c_w > 0\) and \(\max\{|v|, |\bar{v}|\} = C_v\), there exists a Lipschitz constant \(L > 0\) depending on \(C_v\) and \(c_w\), such that

\[
\|\nabla f(v, w) - \nabla f(\bar{v}, \bar{w})\| \leq L\|(v, w) - (\bar{v}, \bar{w})\|.
\]

Proof of Lemma 3 It is easy to check that \(|I_m + 1_m 1_m\top| = m + 1\). Then

\[
\left\| \frac{\partial f}{\partial v}(v, w) - \frac{\partial f}{\partial v}(\bar{v}, \bar{w}) \right\| = \frac{1}{4} \left\| (I_m + 1_m 1_m\top)(v - \bar{v}) + \frac{2}{\pi}(\theta(w, w^*) - \theta(\bar{w}, \bar{w}^*))v^* \right\|
\]

\[
\leq \frac{m+1}{4}\|v - \bar{v}\| + \frac{\|v^*\|}{2\pi}\left|\theta(w, w^*) - \theta(\bar{w}, \bar{w}^*)\right|
\]

\[
\leq \frac{m+1}{4}\|v - \bar{v}\| + \frac{\|v^*\|}{4c_w}\|w - \bar{w}\|
\]

\[
\leq \frac{1}{4}\left(m + 1 + \frac{\|v^*\|}{c_w}\right)\|(v, w) - (\bar{v}, \bar{w})\|,
\]

where the last inequality is due to Lemma 14.
We further have
\[
\left\| \frac{\partial f}{\partial w}(v, w) - \frac{\partial f}{\partial w}(\tilde{v}, \tilde{w}) \right\| \leq \frac{\|v^\top v^*\|}{2\pi \|w\|} \|w - \tilde{w}\| + \frac{\|v^*\|}{2\pi \|w\|} \|v - \tilde{v}\|,
\]
where the second last inequality is to due to Lemma 14. Combining the two inequalities above validates the claim.

**Lemma 4.** The expected partial gradient of $\ell(v, w; Z)$ w.r.t. $v$ is
\[
E_Z \left[ \frac{\partial \ell}{\partial v} (v, w; Z) \right] = \frac{\partial f}{\partial v}(v, w).
\]

Let $\mu(x) = \max\{x, 0\}$ in (5). The expected coarse gradient w.r.t. $w$ is
\[
E_Z \left[ g_{\text{relu}}(v, w; Z) \right] = \frac{h(v, v^*)}{\sqrt{2\pi}} \left( \frac{\theta(w, w^*)}{\sqrt{2\pi}} \right) \frac{v^\top v^*}{\|w\|^2} + \frac{w^*}{\|w\|^2} + \frac{w^*}{\|w\|^2} + \frac{w^*}{\|w\|^2},
\]
where $h(v, v^*) = \|v\|^2 + (1_m v^2 - (1_m v)(1_m v^*) + v^\top v^*.$

**Proof of Lemma 4** The first claim is true because $\frac{\partial f}{\partial v}(v, w; Z)$ is linear in $v$. By (5),
\[
g_{\mu}(v, w; Z) = Z^\top \left( \mu'(Zw) \circ v \right) \left( v^\top \sigma(Zw) - (v^*)^\top \sigma(Zw^*) \right).
\]

Using the fact that $\mu' = \sigma = 1_{\{x > 0\}},$ we have
\[
E_Z \left[ g_{\text{relu}}(v, w; Z) \right] = E_Z \left[ \sum_{i=1}^m v_i \sigma(Z_i^\top w) - \sum_{i=1}^m v_i^* \sigma(Z_i^\top w^*) \right]
\]
\[
= E_Z \left[ \sum_{i=1}^m v_i 1_{\{Z_i^\top w > 0\}} - \sum_{i=1}^m v_i^* 1_{\{Z_i^\top w^* > 0\}} \right] \left( \sum_{i=1}^m Z_i v_i \sigma(Z_i^\top w) \right).
\]

Invoking Lemma 11 we have
\[
E \left[ Z_i 1_{\{Z_i^\top w > 0, Z_i^\top w^* > 0\}} \right] = \begin{cases} \frac{1}{2\sqrt{2\pi}} \frac{w_i}{\|w\|} & \text{if } i = j, \\ \frac{1}{2\sqrt{2\pi}} \frac{w_i}{\|w\|} & \text{if } i \neq j, \end{cases}
\]
and
\[
E \left[ Z_i 1_{\{Z_i^\top w > 0, Z_i^\top w^* > 0\}} \right] = \begin{cases} \frac{\cos(\theta(w, w^*)/2)}{\sqrt{2\pi}} \frac{w_i}{\|w\|} + \frac{w_i^*}{\|w\|} & \text{if } i = j, \\ \frac{1}{2\sqrt{2\pi}} \frac{w_i}{\|w\|} & \text{if } i \neq j. \end{cases}
\]

We redefine the second term as $O_n$ in the case $\theta(w, w^*) = \pi$, or equivalently, $\frac{w_i}{\|w\|} + w_i^* = 0_n.$
Therefore,

\[ E[Z \{ g_{\text{relu}}(v, w; Z) \}] = \sum_{i=1}^{m} v_i^2 E[Z_i 1_{Z_i > 0}] + \sum_{i=1}^{m} \sum_{j \neq i} v_i v_j E[Z_i 1_{Z_i > 0, Z_j > 0}] \]

\[ - \sum_{i=1}^{m} v_i v_i^* E[Z_i 1_{Z_i > 0, Z_i^* > 0}] \]

\[ - \sum_{i=1}^{m} \sum_{j \neq i} v_i v_j^* E[Z_i 1_{Z_i > 0, Z_j^* > 0}] \]

\[ = \frac{1}{2\sqrt{2\pi}} \left( ||v||^2 + (1_m v)^2 \right) \frac{w}{||w||} - \cos \left( \frac{\theta(w, w^*)}{2} \right) \frac{v^*}{\sqrt{2\pi}} \left( \frac{w}{||w||} + w^* \right) \]

and the result follows.

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

\[ \langle E[Z \{ g_{\text{relu}}(v, w; Z) \}], \frac{\partial f}{\partial w}(v, w) \rangle = \sin \left( \frac{\theta(w, w^*)}{2} \right) \left( v^* w \right)^2 \geq 0. \]

Moreover, if further \( ||v|| \leq C_v \) and \( ||w|| \geq c_w \), there exists a constant \( A_{\text{relu}} > 0 \) depending on \( C_v \) and \( c_w \), such that

\[ \left\| E[Z \{ g_{\text{relu}}(v, w; Z) \}] \right\|^2 \leq A_{\text{relu}} \left( \left\| \frac{\partial f}{\partial v}(v, w) \right\|^2 + \left\langle E[Z \{ g_{\text{relu}}(v, w; Z) \}], \frac{\partial f}{\partial w}(v, w) \right\rangle \right). \]

**Proof of Lemma 5.** By Lemmas 2 and 4, we have

\[ \frac{\partial f}{\partial w}(v, w) = -\frac{v^* v}{2\pi ||w||} \left( I_n - \frac{w w^T}{||w||^2} \right) w^* \]

and

\[ E[Z \{ g_{\text{relu}}(v, w; Z) \}] = h(v, v^*) \frac{w}{2\sqrt{2\pi}} - \cos \left( \frac{\theta(w, w^*)}{2} \right) \frac{v^* w^*}{\sqrt{2\pi}} \left( \frac{w}{||w||} \right) + w^* \]
Notice that \((I_n - \frac{ww^\top}{\|w\|^2})w = 0_n\) and \(\|w^*\| = 1\), if \(\theta(w, w_*) \neq 0, \pi\), then we have

\[
\mathbb{E}_Z [g_{\text{relu}}(v, w; Z), \frac{\partial f}{\partial w}(v, w)] = \cos \left(\frac{\theta(w, w_*)}{2}\right) \frac{(v^\top v)^2}{\sqrt{2\pi} \|w\|} \left(1 - \frac{\|w\|^2}{\|w^*\|^2}\right) \|w||v^*| - \frac{1}{4} \|w^*\|^2 \|v\|^2 - (w^\top w^*)^2 \|w\| + (w^\top v^*)^2.
\]

To show the second claim, without loss of generality, we assume \(\|w\| = 1\). Denote \(\theta := \theta(w, w^*)\).

By Lemma 1, we have

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

By Lemma 4,

\[
\mathbb{E}_Z [g_{\text{relu}}(v, w; Z)] = \frac{h(v, v^*)}{2\sqrt{2\pi} \|w\|} - \cos \left(\frac{\theta}{2}\right) \frac{v^\top v^*}{\sqrt{2\pi} \|w\|} \frac{w}{\|w\|} + w^* \tag{22}
\]

where

\[
h(v, v^*) = \|v\|^2 + (1_m v)^2 - (1_m v)(1_m v^*) + v^\top v^*
\]

\[
= v^\top (I_m + 1_m 1_m^\top) v - v^\top (1_m 1_m^\top - I_m) v^*
\]

\[
= v^\top (I_m + 1_m 1_m^\top) v - v^\top \left(1_m 1_m^\top + \left(1 - \frac{2\theta}{\pi}\right) I_m\right) v^* + 2 \left(1 - \frac{\theta}{\pi}\right) v^\top v^*.
\]

and by the first claim,

\[
\mathbb{E}_Z [g_{\text{relu}}(v, w; Z), \frac{\partial f}{\partial w}(v, w)] = \frac{\sin (\theta)}{2(\sqrt{2\pi})^3 \|w\|} (v^\top v^*)^2.
\]
Hence, for some $A_{\text{relu}}$ depending only on $C_{v}$ and $c_{w}$, we have
\[
\left\| E_{Z} \left[ g_{\text{relu}}(v, w; Z) \right] \right\| ^{2} = \\
= \left\| 2v\frac{\partial f}{\partial v}(v, w) \frac{w}{\| w \|} + \cos \left( \frac{\theta}{2} \right) \frac{v^{\top}v^{\ast}}{\sqrt{2\pi}} \left( \frac{w}{\| w \|} - \frac{w}{\| w \|} + w^{\ast} \right) \right\| ^{2} \\
+ \left( 1 - \frac{\theta}{\pi} - \cos \left( \frac{\theta}{2} \right) \right) \left( \frac{v^{\top}v^{\ast}}{\sqrt{2\pi}} \frac{w}{\| w \|} \right)^{2} \\
\leq \frac{6C_{v}^{2}}{\pi} \| \frac{\partial f}{\partial v}(v, w) \| ^{2} + \cos^{2} \left( \frac{\theta}{2} \right) \frac{3(v^{\top}v^{\ast})^{2}}{2\pi} \left\| \frac{w}{\| w \|} - \frac{w}{\| w \|} + w^{\ast} \right\| ^{2} \\
+ \left( 1 - \frac{\theta}{\pi} - \cos \left( \frac{\theta}{2} \right) \right) \frac{3(v^{\top}v^{\ast})^{2}}{2\pi} \\
\leq \frac{6C_{v}^{2}}{\pi} \| \frac{\partial f}{\partial v}(v, w) \| ^{2} + \cos^{2} \left( \frac{\theta}{2} \right) \frac{3\theta^{2}}{8\pi} (v^{\top}v^{\ast})^{2} + \left( 1 - \frac{\theta}{\pi} - \cos \left( \frac{\theta}{2} \right) \right) \frac{3(v^{\top}v^{\ast})^{2}}{2\pi} \\
\leq \frac{6C_{v}^{2}}{\pi} \| \frac{\partial f}{\partial v}(v, w) \| ^{2} + \frac{3\pi}{8} \cos^{2} \left( \frac{\theta}{2} \right) \sin^{2} \left( \frac{\theta}{2} \right) (v^{\top}v^{\ast})^{2} + \frac{3\sin \left( \theta \right)}{2\pi} (v^{\top}v^{\ast})^{2} \\
\leq A_{\text{relu}} \left( \| \frac{\partial f}{\partial v}(v, w) \| ^{2} + \left\langle \frac{\partial f}{\partial v}(v, w) \right\rangle ^{2} \right) ,
\]
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 $\frac{w}{\| w \|}$ and $\frac{w}{\| w \|} + w^{\ast}$ is $\frac{\theta}{2}$ and
\[
\left\| \frac{w}{\| w \|} - \frac{w}{\| w \|} + w^{\ast} \right\| ^{2} \leq \frac{\theta}{2} ,
\]
whereas the third inequality is due to $\sin(x) \geq \frac{2x}{\pi}$, $\cos(x) \geq 1 - \frac{2x}{\pi}$, and
\[
\left( 1 - \frac{2x}{\pi} - \cos(x) \right) ^{2} \leq \left( \cos(x) - 1 + \frac{2x}{\pi} \right) \left( \cos(x) + 1 - \frac{2x}{\pi} \right) \leq \sin(x)(2\cos(x)) = \sin(2x)
\]
for all $x \in [0, \frac{\pi}{2}]$.

\[\Box\]

**Lemma 6.** When Algorithm 1 converges, $E_{Z} \left[ \frac{\partial f}{\partial v}(v, w; Z) \right]$ and $E_{Z} \left[ g_{\text{relu}}(v, w; Z) \right]$ vanish simultaneously, which only occurs at the

1. Saddle points where (8) is satisfied according to Proposition 1
2. Minimizers of (2) where $v = v^{\ast}$, $\theta(w, w^{\ast}) = 0$, or $v = (I_{m} + 1_{m}1_{m}^{\top})^{-1}(1_{m}1_{m}^{\top} - I_{m})v^{\ast}$, $\theta(w, w^{\ast}) = \pi$.

**Proof of Lemma 6.** By Lemma 4 suppose we have
\[
E_{Z} \left[ \frac{\partial f}{\partial v}(v, w; Z) \right] = \frac{1}{4} (I_{m} + 1_{m}1_{m}^{\top})v - \frac{1}{4} \left( 1 - \frac{2}{\pi} \theta(w, w^{\ast}) \right) I_{m} + 1_{m}1_{m}^{\top} v^{\ast} = 0_{m} \quad (24)
\]
and
\[
E_{Z} \left[ g_{\text{relu}}(v, w; Z) \right] = \frac{h(v, v^{\ast})}{2\sqrt{2\pi}} \frac{w}{\| w \|} - \cos \left( \frac{\theta(w, w^{\ast})}{2} \right) \frac{v^{\top}v^{\ast}}{\sqrt{2\pi}} \frac{w}{\| w \|} + w^{\ast} = 0_{n}, \quad (25)
\]
where $h(v, v^{\ast}) = \| v \|^{2} + \frac{(1_{m}^{\top}v)^{2}}{2} - (1_{m}^{\top}v)(1_{m}^{\top}v^{\ast}) + v^{\top}v^{\ast}$. By (25), we must have $\theta(w, w^{\ast}) = 0$ or $\theta(w, w^{\ast}) = \pi$ or $v^{\top}v^{\ast} = 0$. 

23
If $\theta(w, w^*) = 0$, then by (24), $v = v^*$, and (25) is satisfied.

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

If $v^\top v^* = 0$, then by (24), we have the expressions for $v$ and $\theta(w, w^*)$ from Proposition 11 and (25) is satisfied.

Lemma 7. If $w \neq 0_n$ and $\theta(w, w^*) \in (0, \pi)$, then

$$
\mathbb{E}_Z [g_{\text{crelu}}(v, w; Z)] = \frac{p(0, w)h(v, v^*)}{2} \| w \| - (v^\top v^*) \csc(\theta/2) \cdot q(\theta, w) \| w \| + w^*
$$

$$
- (v^\top v^*) (p(\theta, w) - \cot(\theta/2) \cdot q(\theta, w)) \frac{w}{\| w \|},
$$

where $h(v, v^*) := \| v \|^2 + (1_m^\top v)^2 - (1_m^\top v) (1_m^\top v)^* + v^\top v^*$ same as in Lemma 5 and

$$
p(\theta, w) := \frac{1}{2\pi} \int_{-\frac{\pi}{2} + \theta}^{\frac{\pi}{2}} \cos(\phi) \xi \left( \frac{\sec(\phi)}{\| w \|} \right) \, d\phi, \quad q(\theta, w) := \frac{1}{2\pi} \int_{-\frac{\pi}{2} + \theta}^{\frac{\pi}{2}} \sin(\phi) \xi \left( \frac{\sec(\phi)}{\| w \|} \right) \, d\phi
$$

with $\xi(x) := \int_0^x \exp(-\frac{\phi^2}{2}) \, d\phi$. The inner product between the expected coarse and true gradients w.r.t. $w$

$$
\mathbb{E}_Z \left[ g_{\text{crelu}}(v, w; Z) \right] \cdot \frac{\partial f}{\partial w}(v, w) = \frac{q(\theta, w)}{2\pi\| w \|} (v^\top v^*)^2 \geq 0.
$$

Moreover, if further $\| v \| \leq C_v$ and $\| w \| \geq c_w$, there exists a constant $A_{\text{crelu}} > 0$ depending on $C_v$ and $c_w$, such that

$$
\| \mathbb{E}_Z [g_{\text{crelu}}(v, w; Z)] \|^2 \leq A_{\text{crelu}} \left( \| \frac{\partial f}{\partial w}(v, w) \|^2 + \mathbb{E}_Z [g_{\text{crelu}}(v, w; Z)] \cdot \frac{\partial f}{\partial w}(v, w) \right).
$$

Proof of Lemma 7 Denote $\theta := \theta(w, w^*)$. We first compute $\mathbb{E}_Z [g_{\text{crelu}}(v, w; Z)]$. By (5),

$$
g(\mu, v; Z) = Z^\top (\mu' (ZW) \circ v) \left( v^\top \sigma(Zw) - (v^*)^\top \sigma(Zw^*) \right).
$$

Since $\mu' = 1_{\{0 < \theta < 1\}}$ and $\sigma = 1_{\{\theta > 0\}}$, we have

$$
\mathbb{E}_Z [g_{\text{crelu}}(v, w; Z)] = \mathbb{E}_Z \left[ \sum_{i=1}^m v_i \mu' (Z_i^\top \mu) \sum_{i=1}^m v_i^* \mu' (Z_i^\top w^*) \sum_{i=1}^m Z_i v_i \sigma(Z_i^\top w) \right]
$$

$$
= \mathbb{E}_Z \left[ \sum_{i=1}^m v_i 1_{\{0 < Z_i^\top w < 1\}} - \sum_{i=1}^m v_i^* 1_{\{0 < Z_i^\top w^* < 1\}} \right] \left( \sum_{i=1}^m 1_{\{Z_i^\top w > 0\}} v_i Z_i \right)
$$

$$
p(0, w)h(v, v^*) \frac{w}{\| w \|} - (v^\top v^*) \csc(\theta/2) \cdot q(\theta, w) \| w \| + w^*
$$

$$
- (v^\top v^*) (p(\theta, \| w \|) - \cot(\theta/2) \cdot q(\theta, w)) \frac{w}{\| w \|},
$$

In the last equality above, we called Lemma 12.

Notice that $(I_n - \frac{w w^\top}{\| w \|^2}) w = 0_n$, and $\| w^\top \| = 1$. If $\theta(w, w^*) \neq 0, \pi$, then the inner product between $\mathbb{E}_Z [g_{\text{crelu}}(v, w; Z)]$ and $\frac{\partial f}{\partial w}(v, w)$ is given by

$$
\mathbb{E}_Z \left[ g_{\text{crelu}}(v, w; Z) \right] \cdot \frac{\partial f}{\partial w}(v, w)
$$

$$
= \csc \left( \frac{\theta}{2} \right) q(\theta, w) \frac{w^*}{2\pi\| w \|} (v^\top v^*)^2 \left( 1_{\| w \|}^\top \frac{(I_n - \frac{w w^\top}{\| w \|^2}) w^*}{\| w \| + w^*} \right)
$$

$$
= \frac{q(\theta, w)}{2\pi\| w \|} (v^\top v^*)^2 \geq 0.
$$

24
In the last line, \( q(\theta, w) \geq 0 \) because \( \sin(\phi) \xi \left( \frac{\sec(\phi)}{||w||} \right) \) is odd in \( \phi \) and positive for \( \phi \in (0, \frac{\pi}{2}) \).

Next, we bound \( \left\| \mathbb{E}_z [g_{\text{crelu}}(v, w; Z)] \right\|^2 \). Since (23) gives
\[
h(v, v^*) = 4v^\top \frac{\partial f}{\partial v}(v, w) + 2 \left( 1 - \frac{\theta}{\pi} \right) v^\top v^*,
\]
where according to Lemma [1]
\[
\frac{\partial f}{\partial v}(v, w) = \frac{1}{4} \left( (I_m + 1_m) v - \frac{1}{4} \left( 1 - \frac{2\theta}{\pi} \right) I_m + 1_m^\top \right) v^*.
\]
We rewrite the coarse partial gradient w.r.t. \( w \) in (26) as
\[
\mathbb{E}_z [g_{\text{crelu}}(v, w; Z)] = p(0, w) \left( 2v^\top \frac{\partial f}{\partial v}(v, w) + \left( 1 - \frac{\theta}{\pi} \right) v^\top v^* \right) \frac{w}{||w||}
\] 
\[
- (v^\top v^*) \csc(\theta/2) \cdot q(\theta, w) \frac{w}{||w||} + w^*
\] 
\[
- (v^\top v^*) (p(\theta, w) - \cot(\theta/2) \cdot q(\theta, w)) \frac{w}{||w||},
\]
\[
= 2p(0, w)v^\top \frac{\partial f}{\partial v}(v, w) \frac{w}{||w||}
\]
\[
+ (v^\top v^*) \left( \left( 1 - \frac{\theta}{\pi} \right) p(0, w) - p(\theta, w) \right) \frac{w}{||w||}
\]
\[
+ (v^\top v^*) \csc(\theta/2) \cdot q(\theta, w) \left( \frac{w}{||w||} - \frac{w}{||w||} + w^* \right)
\]
\[
- (v^\top v^*) (\csc(\theta/2) - \cot(\theta/2))q(\theta, w) \frac{w}{||w||}.
\]

To prove the last claim, we notice that in the above equality,
\[
\left\| v^\top \frac{\partial f}{\partial v}(v, w) \frac{w}{||w||} \right\|^2 \leq C_v \left\| \frac{\partial f}{\partial v}(v, w) \right\|^2,
\]
(27)
\[
\left\| (\csc(\theta/2) - \cot(\theta/2))q(\theta, w) \frac{w}{||w||} \right\|^2 = (\csc(\theta/2) - \cot(\theta/2))^2 q(\theta, w)^2 \leq q(\theta, w)^2,
\]
(28)
\[
\left\| \csc(\theta/2) \cdot q(\theta, w) \left( \frac{w}{||w||} - \frac{w}{||w||} + w^* \right) \right\|^2 \leq \left( \frac{\theta/2}{\sin(\theta/2)} \right)^2 q(\theta, w)^2 \leq \frac{\pi^2}{4} q(\theta, w)^2.
\]
(29)

Now, what is left is to bound \( \left( 1 - \frac{\theta}{\pi} \right) p(0, w) - p(\theta, w) \)^2, using a multiple of \( q(\theta, w) \). Recall that
\[
p(\theta, w) = \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^\frac{\pi}{2} \cos(\phi) \xi \left( \frac{\sec(\phi)}{||w||} \right) d\phi,
\]
\[
q(\theta, w) = \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^\frac{\pi}{2} \sin(\phi) \xi \left( \frac{\sec(\phi)}{||w||} \right) d\phi.
\]

We first show that both \( \left( 1 - \frac{\theta}{\pi} \right) p(0, w) - p(\theta, w) \)^2 and \( q(\theta, w) \) are symmetric with respect to \( \theta = \frac{\pi}{2} \) on \([0, \pi]\). This is because
\[
q(\pi - \theta, w) = \frac{1}{2\pi} \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \sin(\phi) \xi \left( \frac{\sec(\phi)}{||w||} \right) d\phi
\]
\[
= q(\theta, w) + \frac{1}{2\pi} \int_{\frac{\pi}{2}}^{\frac{\pi}{2} + \theta} \sin(\phi) \xi \left( \frac{\sec(\phi)}{||w||} \right) d\phi = q(\theta, w).
\]
and
\[
\left(1 - \frac{\pi - \theta}{\pi}\right) p(0, w) - p(\pi, \theta, w) \\
= \frac{\theta}{\pi} \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\phi) \xi \left(\frac{\text{sec}(\phi)}{\|w\|}\right) d\phi - \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\phi) \xi \left(\frac{\text{sec}(\phi)}{\|w\|}\right) d\phi \\
= \left(\frac{\theta}{\pi} - 1\right) \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\phi) \xi \left(\frac{\text{sec}(\phi)}{\|w\|}\right) d\phi + \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\phi) \xi \left(\frac{\text{sec}(\phi)}{\|w\|}\right) d\phi \\
= - \left(1 - \frac{\theta}{\pi}\right) p(0, w) - p(\theta, w).
\]

Therefore, it suffices to consider \(\theta \in [\frac{\pi}{2}, \pi]\) only. Then calling Lemma 13 for \(\theta \in [\frac{\pi}{2}, \pi]\), we have \(p(\theta, w) \leq q(\theta, w)\) and \((1 - \frac{\theta}{\pi}) p(0, w) \leq q(\theta, w)\). Therefore,
\[
\left(p(\theta, w) - \left(1 - \frac{\theta}{\pi}\right) p(0, w)\right)^2 \leq q(\theta, w)^2.
\]

Combining the above estimate together with (27), (28) and (29), and using Cauchy-Schwarz inequality, we have
\[
\|E_g(v, w; Z)\|^2 \leq 4 \left(4p(0, w)C_2^2 \left\|\frac{\partial f}{\partial v}(v, w)\right\|^2 + \left(2 + \frac{\pi^2}{4}\right) q(\theta, w)^2 (v^T v^*)^2\right),
\]
where \(p(0, w)\) and \(q(\theta, w)\) are uniformly bounded. This completes the proof.

\[\square\]

**Lemma 8.** When Algorithm 1 converges, \(E_g\) and \(E_\theta\) vanish simultaneously, which only occurs at the

1. Saddle points where (8) is satisfied according to Proposition 7
2. Minimizers of (2) where \(v = v^*, \theta(w, w^*) = 0\), or \(v = (I_m + 1_m I_m^T)^{-1}(1_m I_m - I_m)v^*\).

**Proof of Lemma 8** The proof of Lemma 8 is similar to that of Lemma 6 and we omit it here. The core part is that \(q(\theta, w)\) defined in Lemma 12 is non-negative and equals 0 only at \(\theta = 0, \pi\), as well as \(p(0, w) \geq p(\theta, w) \geq p(\pi, w) = 0\). \[\square\]

**Lemma 9.** Let \(\mu(x) = x\) in (3). Then the expected coarse partial gradient w.r.t. \(w\) is
\[
E_g = \frac{1}{\sqrt{2\pi}} \left(\|w\|^2 \frac{w}{\|w\|} - (v^T v^*) w^*\right).
\]

If \(\theta(w, w^*) = \pi\) and \(v = (I_m + 1_m I_m^T)^{-1}(1_m I_m - I_m)v^*\),
\[
\|E_g\| = \frac{2(m-1)}{\sqrt{2\pi}(m+1)} (1_m v^*)^2 \geq 0,
\]
i.e., \(E_g\) does not vanish at the local minimizers if \(1_m v^* \neq 0\) and \(m > 1\).

**Proof of Lemma 9** By (3),
\[
g_s(v, w; Z) = Z^T \left(\mu'(Zw) \circ v\right) \left(v^T \sigma(Zw) - (v^*)^T \sigma(Zw^*)\right),
\]

Published as a conference paper at ICLR 2019
Using the facts that $\mu^\prime = 1$ and $\sigma = 1_{\{z > 0\}}$, we have
\[
\mathbb{E}_z [g_{id}(v, w; Z)] = \mathbb{E}_z \left[ \sum_{i=1}^{m} v_i 1_{\{Z_i^\top w > 0\}} - \sum_{i=1}^{m} v_i^* 1_{\{Z_i^\top w^* > 0\}} \right] \left( \sum_{i=1}^{m} v_i Z_i \right) \\
= \sum_{i=1}^{m} \sum_{j=1}^{m} v_i v_j \mathbb{E} \left[ Z_i 1_{\{Z_i^\top w > 0\}} \right] - \sum_{i=1}^{m} \sum_{j=1}^{m} v_i v_j \mathbb{E} \left[ Z_i 1_{\{Z_i^\top w^* > 0\}} \right] \\
= \frac{1}{\sqrt{2\pi}} \left( \|v\|^2 \frac{w}{\|w\|} - (v^\top v^*) w^* \right).
\]
In the last equality above, we called the third identity in Lemma 11. If $\theta(w, w^*) = \pi$ and $v = (I_m + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^*$, then
\[
\|\mathbb{E}_z [g_{id}(v, w; Z)]\| = \frac{1}{\sqrt{2\pi}} \|v^\top (v + v^*)\| \\
= \frac{1}{\sqrt{2\pi}} \left| (v^*)^\top (I_m + 1_m 1_m^\top - I_m)(I_m + 1_m 1_m^\top)^{-1} (I_m + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m) + I_m \right| v^* \\
= \frac{2}{\sqrt{2\pi}} \frac{1}{(m + 1)^2} \left| (v^*)^\top (I_m + 1_m 1_m^\top - I_m)1_m(1_m^\top v^*) \right| \\
= \frac{2(m - 1)}{\sqrt{2\pi} (m + 1)^2} (1_m^\top v^*)^2.
\]
In the third equality, we used the identity $(I_m + 1_m 1_m^\top)1_m = (m + 1)1_m$ twice.

**Lemma 10.** If $w \neq 0_n$ and $\theta(w, w^*) \in (0, \pi)$, then the inner product between the expected coarse and true gradients w.r.t. $w$ is
\[
\left\langle \mathbb{E}_z [g_{id}(v, w; Z)], \frac{\partial f}{\partial w}(v, w) \right\rangle = \frac{\sin(\theta(w, w^*))}{(\sqrt{2\pi})^3 \|w\|} (v^\top v^*)^2 \geq 0.
\]
When $\theta(w, w^*) \to \pi$, $v \to (I_m + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^*$, if $1_m v^* \neq 0$ and $m > 1$, we have
\[
\left\| \mathbb{E}_z [g_{id}(v, w; Z)] \right\|^2 \to +\infty.
\]

**Proof of Lemma 10.** By Lemmas 2 and 4, we have
\[
\frac{\partial f}{\partial w}(v, w) = -\frac{v^\top v^*}{2\|w\|} \frac{(I_n - \frac{w v^*}{\|w\|^2}) w^*}{\|w\|^2} \\
\text{and} \\
\mathbb{E}_z [g_{id}(v, w; Z)] = \frac{1}{\sqrt{2\pi}} \left( \|v\|^2 \frac{w}{\|w\|} - (v^\top v^*) w^* \right).
\]
Since $(I_n - \frac{w v^*}{\|w\|^2}) w = 0_n$ and $\|w^*\| = 1$, if $\theta(w, w^*) \neq 0, \pi$, then we have
\[
\left\langle \mathbb{E}_z [g_{id}(v, w; Z)], \frac{\partial f}{\partial w}(v, w) \right\rangle = \frac{(v^\top v^*)^2}{(\sqrt{2\pi})^3 \|w\|} \frac{\left( \frac{w}{\|w\|^2} \right)^2}{\|w\|^2} \frac{(I_n - \frac{w v^*}{\|w\|^2}) w^*}{\|w\|^2} \frac{(I_n - \frac{w v^*}{\|w\|^2}) w^*}{\|w\|^2} \\
= \frac{(v^\top v^*)^2}{(\sqrt{2\pi})^6 \|w\|^2} \frac{1 - (\frac{w v^*}{\|w\|^2})^2}{\|w\|^2} \frac{(v^\top v^*)^2}{(\sqrt{2\pi})^3 \|w\|} \frac{1 - (\frac{w v^*}{\|w\|^2})^2}{\|w\|^2} \\
= \frac{(v^\top v^*)^2}{(\sqrt{2\pi})^6 \|w\|^2} \sin(\theta(w, w^*)).
\]
When \( \theta(w, w^*) \rightarrow \pi, \ v \rightarrow (I_m + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^* \), both \( \left\| \frac{\partial f}{\partial w}(v, w) \right\| \) and \( \left\| \mathbb{E}_Z[g_{id}(v, w; Z)] \frac{\partial f}{\partial w}(v, w) \right\| \) converge to 0. But if \( 1_m^\top v^* \neq 0 \) and \( m > 1 \), \[ \left\| \mathbb{E}_Z[g_{id}(v, w; Z)] \right\| \rightarrow \frac{2(m-1)}{\sqrt{2\pi(m+1)^2}} (1_m^\top v^*)^2 > 0, \] which completes the proof. \( \square \)

**Theorem 1.** Let \( \{(v^t, w^t)\} \) be the sequence generated by Algorithm 1 with ReLU \( \mu(x) = \max\{x, 0\} \) or clipped ReLU \( \mu(x) = \min\{\max\{x, 0\}, 1\} \). Suppose \( \|w\| \geq c_w \) for all \( t \) with some \( c_w > 0 \). Then if the learning rate \( \eta > 0 \) is sufficiently small, for any initialization \((v^0, w^0)\), the objective sequence \( \{f(v^t, w^t)\} \) is monotonically decreasing, and \( \{(v^t, w^t)\} \) converges to a saddle point or a (local) minimizer of the population loss minimization. In addition, if \( 1_m^\top v^* \neq 0 \) 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 \( \theta(w, w^*) = \pi \) and \( v = (I_m + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^* \).

**Proof of Theorem 1.** We first prove the upper boundedness of \( \{v^t\} \). Due to the coerciveness of \( f(v, w) \) w.r.t \( v \), there exists \( C_v > 0 \), such that \( \|v\| \leq C_v \) for any \( v \in \mathbb{R}^m : f(v, w) \leq f(v^0, w^0) \) for some \( w \). In particular, \( \|v^0\| \leq C_v \). Using induction, suppose we already have \( f(v^t, w^t) \leq f(v^0, w^0) \) and \( \|v^t\| \leq C_v \). If \( \theta(w^t, w^*) = 0 \) or \( \pi \), then \( \theta(w^t, w^*) = 0 \) or \( \pi \) for all \( T \geq t \), and the original problem reduces to a quadratic program in terms of \( v \). So \( \{v^t\} \) will converge to \( v^* \) or \( (I_m + 1_m 1_m^\top)^{-1}(1_m 1_m^\top - I_m)v^* \) by choosing a suitable step size \( \eta \). In either case, we have \( \left\| \mathbb{E}_Z[\frac{\partial f}{\partial v}(v^t, w^t; Z)] \right\| \) and \( \left\| \mathbb{E}_Z[g_{relu}(v^t, w^t; Z)] \right\| \) both converge to 0. Else if \( \theta(w^t, w^*) \in (0, \pi) \), we define for any \( a \in [0, 1] \) that

\[
v^t(a) := v^t - a(v^{t+1} - v^t) = v^t - a\eta \mathbb{E}_Z \left[ \frac{\partial f}{\partial v}(v^t, w^t; Z) \right]
\]

and

\[
w^t(a) := w^t - a(w^{t+1} - w^t) = w^t - a\eta \mathbb{E}_Z \left[ g_{relu}(v^t, w^t; Z) \right],
\]

which satisfy

\[
v^t(0) = v^t, \ v^t(1) = v^{t+1}, \ w^t(0) = w^t, \ w^t(1) = w^{t+1}.
\]

Let us fix \( 0 < c \leq c_w \) and \( C \geq C_v \). By the expressions of \( \mathbb{E}_Z \left[ \frac{\partial f}{\partial v}(v^t, w^t; Z) \right] \) and \( \mathbb{E}_Z \left[ g_{relu}(v^t, w^t; Z) \right] \) given in Lemma 4 and since \( \|w^t\| = 1 \), for sufficiently small \( \tilde{\eta} \) depending on \( C_0 \) and \( c_w \), with \( \eta \leq \tilde{\eta} \), it holds that \( \|v^t(a)\| \leq C \) and \( \|w^t(a)\| \geq c \) for all \( a \in [0, 1] \). Possibly at some point \( a_0 \) where \( \theta(w^t(a_0), w^*) = 0 \) or \( \pi \), the partial gradient \( \frac{\partial f}{\partial w}(v^t(a_0), w^t(a_0)) \) does not exist. Otherwise, \( \left\| \frac{\partial f}{\partial w}(v^t(a), w^t(a)) \right\| \) 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 \( A_{relu} \) depending on \( C \) and \( c \), we
We prove that if the initialization weights $(v^t, w^t)$ satisfy $(v^0, w^t)^\top v^* > 0$, $\theta(w^0, w^t) < \frac{\eta}{2}$ and $(1_w^\top v^0) (1_m^\top v^0) \leq (1_m^\top v^* )^2$, then we have convergence guarantee to global optima by using the vanilla or clipped ReLU STE.

The second claim follows from Lemmas 9 and 10.

F. CONVERGENCE TO GLOBAL MINIMIZERS

We prove that if the initialization weights $(v^0, w^0)$ satisfy $(v^0)^\top v^* > 0$, $\theta(w^0, w^*) < \frac{\eta}{2}$ and $(1_w^\top v^0) (1_m^\top v^0) \leq (1_m^\top v^* )^2$, then we have convergence guarantee to global optima by using the vanilla or clipped ReLU STE.
Theorem 2. Under the assumptions of Theorem\[7\] if further the initialization $(v^0, w^0)$ satisfies $(v^0)^\top v^* > 0$, $\theta(w^0, w^*) < \frac{\pi}{2}$ and $(1_m^\top v^*)(1_m^\top v^0) \leq (1_m^\top v^*)^2$, then by using the vanilla or clipped ReLU STE for sufficiently learning rate $\eta > 0$, we have $(v^t)^\top v^* > 0$ and $\theta(w^t, w^*) < \frac{\pi}{2}$ for all $t > 0$, and $\{(v^t, w^t)\}$ converges to a global minimizer.

Proof of Theorem[2] Proof by induction. Suppose $(v^t)^\top v^* > 0$, $\theta(w^t, w^*) < \frac{\pi}{2}$ and $(1_m^\top v^*)(1_m^\top v^t) \leq (1_m^\top v^*)^2$. Then for small enough $\eta$, we have

$$(v^{t+1})^\top v^* = \left( v^t - \frac{\eta}{4} \left( (I_m + 1_m 1_m^\top) v^t - \left( 1 - \frac{2}{\pi} \theta(w^t, w^*) \right) I_m + 1_m 1_m^\top \right) v^* \right)$$

$$= \left( 1 - \frac{\eta}{4} \right) (v^t)^\top v^* + \eta \left( (1_m^\top v^*)^2 - (1_m^\top v^*)(1_m^\top v^t) \right)$$

$$+ \eta \left( 1 - \frac{2}{\pi} \theta(w^t, w^*) \right) ||v^*||^2 > 0.$$ 

and

$$(1_m^\top v^t)^\top (1_m^\top v^*) = \left( 1 - \frac{m+1}{4} \eta \right) (1_m^\top v^t)(1_m^\top v^*) + \eta \left( m + 1 - \frac{2}{\pi} \theta(w^t, w^*) \right) (1_m^\top v^*)^2$$

$$\leq \left( 1 - \frac{m+1}{2\pi} \theta(w^t, w^*) \right) (1_m^\top v^*)^2 \leq (1_m^\top v^*)^2.$$ 

Moreover, by Lemmas[4,5] and[7] both $\mathbb{E}_Z[g_{relu}(v, w; Z)]$ and $\mathbb{E}_Z[g_{crelu}(v, w; Z)]$ can be written in the form of $\alpha_1 \left( I_m - \frac{ww^\top}{||w||^2} \right) w^* + \alpha_2 w$, where $\alpha_1 \leq 0$ and $\alpha_2$ is bounded by a constant depending on $C_v$ and $c_w$. Therefore,

$$(w^{t+1})^\top w^* = (1 - \eta\alpha_2)(w^t)^\top w^* - \alpha_1(w^*)^\top \left( I_m - \frac{ww^\top}{||w||^2} \right) w^* \geq (1 - \eta\alpha_2)(w^t)^\top w^* > 0,$$

and thus $\theta(w^{t+1}, w^*) < \frac{\pi}{2}$. Finally, since $\{(v^t, w^t)\}$ converges, it can only converge to a global minimizer. \(\square\)