Gradient Descent Learns One-hidden-layer CNN: Don't be Afraid of Spurious Local Minima

Simon S. Du[1], Jason D. Lee[2], Yuandong Tian[3], Barnabás Póczos[1], and Aarti Singh[1]

[1]Machine Learning Department, Carnegie Mellon University
[2]Department of Data Sciences and Operations, University of Southern California
[3]Facebook Artificial Intelligence Research

December 5, 2017

Abstract

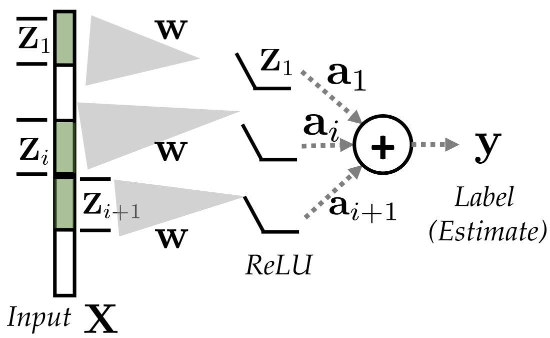
We consider the problem of learning a one-hidden-layer neural network with non-overlapping convolutional layer and ReLU activation function, i.e., $f(\mathbf{Z}; \mathbf{w}, \mathbf{a}) = \sum_j a_j \sigma(\mathbf{w}^T \mathbf{Z}_j)$, in which both the convolutional weights $\mathbf{w}$ and the output weights $\mathbf{a}$ are parameters to be learned. We prove that with Gaussian input $\mathbf{Z}$, there is a spurious local minimum that is not a global mininum. Surprisingly, in the presence of local minimum, starting from randomly initialized weights, gradient descent with weight normalization can still be proven to recover the true parameters with constant probability (which can be boosted to arbitrarily high accuracy with multiple restarts). We also show that with constant probability, the same procedure could also converge to the spurious local minimum, showing that the local minimum plays a non-trivial role in the dynamics of gradient descent. Furthermore, a quantitative analysis shows that the gradient descent dynamics has two phases: it starts off slow, but converges much faster after several iterations.

1 Introduction

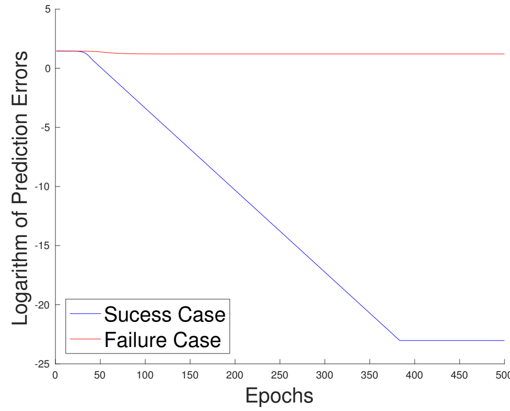
Deep convolutional neural networks (CNN) have achieved the state-of-the-art performance in many applications such as computer vision [Krizhevsky et al., 2012], natural language processing [Dauphin et al., 2016] and reinforcement learning applied in classic games like Go [Silver et al., 2016]. Despite the highly non-convex nature of the objective function, simple first-order algorithms like stochastic gradient descent and its variants often train such networks successfully. The success of such simple methods in learning convolutional neural networks remains elusive from an optimization perspective.

Recently, a line of research [Tian, 2017, Brutzkus and Globerson, 2017, Li and Yuan, 2017, Soltanolkotabi, 2017, Shalev-Shwartz et al., 2017b] assumed the input distribution is Gaussian and showed that stochastic gradient descent with random or $\mathbf{0}$ initialization is able to train a neural network $f(\mathbf{Z}; \{\mathbf{w}_j\}) = \sum_j a_j \sigma(\mathbf{w}_j^T \mathbf{Z})$ with ReLU activation $\sigma(x) = \max(x, 0)$ in polynomial time. However, these results all assume there is only one unknown layer $\{\mathbf{w}_j\}$, while $\mathbf{a}$ is a fixed vector. A natural question thus arises:

Does randomly initialized (stochastic) gradient descent learn ***deep*** *neural networks?*



(a) Convolutional neural network with an unknown non-overlapping filter and an unknown output layer. In the first (hidden) layer, a filter $\mathbf{w}$ is applied to nonoverlapping parts of the input $\mathbf{x}$, which then passes through a ReLU activation function. The final output is the inner product between an output weight vector $\mathbf{a}$ and the hidden layer outputs.



(b) The convergence of gradient descent for learning the convolutional neural network described in Figure 1a with filter size, $p = 20$ and number of non-overlapping patches, $k = 25$. The success case and the failure case correspond to convergence to the global minimum and the spurious local minimum, respectively. In the first ~ 50 iterations the convergence is slow. After that gradient descent converges at a fast linear rate.

Figure 1: Network architecture we consider in this paper and convergence of gradient descent for learning the parameters of this network.

In this paper, we take an important step by showing that randomly initialized gradient descent learns a simple non-linear convolutional neural network with *two* unknown layers $\mathbf{w}$ and $\mathbf{a}$. To our knowledge, our work is the first of its kind.

Formally, we consider the convolutional case in which $\mathbf{w}_i = \mathbf{w}$ are shared among different hidden nodes. Let $\mathbf{x} \in \mathbb{R}^d$ be an input sample, e.g., an image. We generate k patches from $\mathbf{x}$, each with size p: $\mathbf{Z} \in \mathbb{R}^{p \times k}$ where the i-th column is the i-th patch generated by some known function $\mathbf{Z}_i = \mathbf{Z}_i(\mathbf{x})$. Any two patches are non-overlapping. Thus, the neural network function has the following form:

$$f(\mathbf{Z}, \mathbf{w}, \mathbf{a}) = \sum_{i=1}^{k} \mathbf{a}_i \sigma\left(\mathbf{w}^\top \mathbf{Z}_i\right).$$

We focus on the realizable case, i.e., the label is generated according to $y = f\left(\mathbf{Z}, \mathbf{a}^*, \mathbf{w}^*\right)$ for some true parameters $\mathbf{a}^*$ and $\mathbf{w}^*$. We use ℓ_2 loss to learn the parameters:

$$\min_{\mathbf{w},\mathbf{a}} \ell(\mathbf{Z}, \mathbf{w}, \mathbf{a}) := \frac{1}{2}\left(f\left(\mathbf{Z}, \mathbf{w}, \mathbf{a}\right) - f\left(\mathbf{Z}, \mathbf{w}^*, \mathbf{a}^*\right)\right)^2.$$

We assume $\mathbf{x}$ is sampled from a Gaussian distribution and there is no overlap between patches. This assumption is equivalent to that each entry of $\mathbf{Z}$ is sampled from a Gaussian distribution [Brutzkus and Globerson, 2017, Zhong et al., 2017b]. Following [Zhong et al., 2017a,b, Li and Yuan, 2017, Tian, 2017, Brutzkus and Globerson, 2017, Shalev-Shwartz et al., 2017b], in this paper, we mainly focus on the population loss:

$$\ell\left(\mathbf{w}, \mathbf{a}\right) := \frac{1}{2}\mathbb{E}_{\mathbf{Z}}\left[\left(f\left(\mathbf{Z}, \mathbf{w}, \mathbf{a}\right) - f\left(\mathbf{Z}, \mathbf{w}^*, \mathbf{a}^*\right)\right)^2\right].$$

Algorithm 1 GD for Learning One-Hidden-Layer CNN with Weight Normalization

Input: Initialization $\mathbf{v}_0 \in \mathbb{R}^p$, $\mathbf{a}_0 \in \mathbb{R}^k$, learning rate η.
for $t = 1, 2, \ldots$ **do**
  $\mathbf{v}^{t+1} \leftarrow \mathbf{v}^t - \eta \frac{\partial \ell(\mathbf{v}^t, \mathbf{a}^t)}{\partial \mathbf{v}^t}, \qquad \mathbf{a}^{t+1} \leftarrow \mathbf{a}^t - \eta \frac{\partial \ell(\mathbf{v}^t, \mathbf{a}^t)}{\partial \mathbf{a}^t}.$
end for

We study whether the global convergence $\mathbf{w} \to \mathbf{w}_*$ and $\mathbf{a} \to \mathbf{a}_*$ can be achieved when optimizing $\ell(\mathbf{w}, \mathbf{a})$ using gradient descent.

A crucial difference between our two-layer network and previous one-layer models is there is non-uniqueness issue. That is, for any $c > 0$, $f\left(\mathbf{Z}, c\mathbf{w}, \frac{\mathbf{a}}{c}\right) = f\left(\mathbf{Z}, \mathbf{w}, \mathbf{a}\right)$. This interesting property allows the network to be rescaled without changing the function computed by the network. As reported by Neyshabur et al. [2015], it is desirable to have scaling-invariant learning algorithm to stabilize the training process. Similar observations have appeared in the matrix factorization (equivalent to linear activation) literature. For example, see Tu et al. [2016].

One commonly used technique to achieve stability is *weight-normalization* introduced by Salimans and Kingma [2016]. In our setting, we re-parametrize the first layer as $\mathbf{w} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}$ and the prediction function becomes

$$f\left(\mathbf{Z}, \mathbf{v}, \mathbf{a}\right) = \sum_{i=1}^{k} a_i \frac{\sigma\left(\mathbf{Z}_i^\top \mathbf{v}\right)}{\|\mathbf{v}\|_2}. \tag{1}$$

The loss function is

$$\ell\left(\mathbf{v}, \mathbf{a}\right) = \frac{1}{2} \mathbb{E}_{\mathbf{Z}} \left[\left(f\left(\mathbf{Z}, \mathbf{v}, a\right) - f\left(\mathbf{Z}, \mathbf{v}^*, \mathbf{a}^*\right) \right)^2 \right]. \tag{2}$$

The pseudo-code for optimizing the objective function in Equation (2) is listed in Algorithm 1.

We show that with *random initialization*, gradient descent converges to the target convolutional neural network with probability at least 0.25. By further exploiting the symmetry, we can boost up the success probability to 1 with only 3 additional deterministic restarts. Further, perhaps surprisingly, we prove that the objective function (Equation (2)) *does* have a spurious local minimum and with constant probability, using the same random initialization scheme gradient descent can converge to this bad local minimum as well. In contrast to previous works on guarantees for non-convex objective functions whose landscape satisfies "no spurious local minima" property [Li et al., 2016, Sun et al., 2016, Ge et al., 2017a, 2016, Bhojanapalli et al., 2016, Ge et al., 2017b], our result highlights a conceptually surprising phenomenon:

Randomly initialized local search can find a global minimum in presence of spurious local minima.

At the core of our analysis is a series of invariant qualitative characterizations of gradient descent dynamics, which determines the convergence to the global or the spurious local minimum. This analysis emphasizes that for non-convex optimization problems, we need to carefully characterize both the trajectory of the algorithm and the initialization. We believe that this idea is applicable to other non-convex problems.

Next, we conduct a quantitative study of the dynamics of gradient descent. We show that the dynamics of Algorithm 1 contain two phases. At the beginning (around first 50 iterations in

Figure 1b), because the magnitude of initial signal (angle between $\mathbf{v}$ and $\mathbf{w}^*$ and $\mathbf{a}^\top \mathbf{a}^*$) is small, the prediction error drops slowly. After that, when the signal becomes stronger, gradient descent converges at a much faster linear rate and the prediction error drops quickly.

1.1 Organization

This paper is organized as follows. In Section 2 we introduce necessary notations and analytical formulas of gradient updates in Algorithm 1. In Section 3, we provide our main theorems on the performances of algorithms and their implications. In Section 4, we give a proof sketch of our main theorem, laying out the key technical insights of learning one-hidden-layer CNN. We conclude and list future directions in Section 5. We place most of our detailed proofs in the appendix.

1.2 Related Works

From the point of view of learning theory, it is well known that training a neural network is hard in the worst cases [Blum and Rivest, 1989, Livni et al., 2014, Šíma, 2002, Shalev-Shwartz et al., 2017a,b] and recently, Shamir [2016] showed that assumptions on *both* the target function and the input distribution are needed for optimization algorithms used in practice to succeed. With some additional assumptions, many works tried to design algorithms that provably learn a neural network with polynomial time and sample complexity [Goel et al., 2016, Zhang et al., 2015, Sedghi and Anandkumar, 2014, Janzamin et al., 2015, Goel and Klivans, 2017a,b]. However these algorithms are specially designed for certain architectures and cannot explain why (stochastic) gradient based optimization algorithm works well in practice.

Focusing on gradient-based algorithms, a line of research analyzed the behavior of (stochastic) gradient descent for *Gaussian* input distribution. Tian [2017] showed that population gradient descent is able to find the true weight vector with random initialization for one-layer one-neuron model. Soltanolkotabi [2017] later improved this result by showing the true weights can be exactly recovered by empirical projected gradient descent with enough samples in linear time. Brutzkus and Globerson [2017] showed population gradient descent recovers the true weights of a convolution filter with non-overlapping input in polynomial time. Zhong et al. [2017b] and later work [Zhong et al., 2017a] proved that with sufficiently good initialization, which can be implemented by tensor method, gradient descent can find the true weights of a one-hidden-layer fully connected and convolutional neural network. Li and Yuan [2017] showed SGD can recover the true weights of a one-layer ResNet model with ReLU activation under the assumption that the spectral norm of the true weights is within a small constant of the identity mapping. This paper also follows this line of approach that studies the behavior of gradient descent algorithm with Gaussian inputs.

Finding the optimal weights of a neural network is non-convex problem. Recently, researchers found that if the objective functions satisfy the following two key properties:

1. All saddle points and local maxima are strict, i.e., there exists a negative curvature;

2. No spurious local minmum, i.e., all local minima are global,

then perturbed (stochastic) gradient descent [Ge et al., 2015, Jin et al., 2017, Levy, 2016] or methods with second order information [Nesterov and Polyak, 2006, Curtis et al., 2014, Carmon et al., 2016, Allen-Zhu, 2017] can find a global minimum in polynomial time. Combined with geometric analyses, these algorithmic results have shown a large number problems, including tensor decomposition [Ge et al., 2015], dictionary learning [Sun et al., 2017], phase retrieval [Sun et al., 2016], matrix sensing [Bhojanapalli et al., 2016, Park et al., 2017], matrix completion [Ge et al., 2017a, 2016] and matrix factorization [Li et al., 2016] can be solved in polynomial time with local search algorithms.

This motivates the research of studying the landscape of neural networks [Kawaguchi, 2016, Choromanska et al., 2015a, Hardt and Ma, 2016, Haeffele and Vidal, 2015, Mei et al., 2016, Freeman and Bruna, 2016, Safran and Shamir, 2016, Zhou and Feng, 2017, Nguyen and Hein, 2017a,b, Ge et al., 2017b, Zhou and Feng, 2017]. In particular, Kawaguchi [2016], Hardt and Ma [2016], Zhou and Feng [2017], Nguyen and Hein [2017a,b], Feizi et al. [2017] showed that under some conditions, all local minima are global. Recently, Ge et al. [2017b] showed using a modified objective function satisfying the two properties above, one-hidden-layer neural network can be learned by noisy perturbed gradient descent.[1] However, for nonlinear activation function where with number of samples larger than the number of nodes at every layer, the usually case in most deep neural network, and natural objective functions like ℓ_2, it is still unclear whether the strict saddle and "all locals are global" properties are satisfied. In this paper, we show even for an one-hidden-layer neural network with ReLU activation, there exists a local minimum. However, we further show that randomly initialized local search can achieve *global* minimum with constant probability.

2 Preliminaries

We use bold-faced letters for vectors and matrices. We use $\|\cdot\|_2$ to denote the Euclidean norm of a finite-dimensional vector. Let $O(\cdot)$ and $\Theta\left(\cdot\right)$ denote standard Big-O and Big-Theta notations, only hiding absolute constants. We let $\mathbf{w}^t$ and $\mathbf{a}^t$ to be the parameters at the t-th iteration and $\mathbf{w}^*$ and $\mathbf{a}^*$ be the optimal weights. a_i is the i-th coordinate of a and $\mathbf{Z}_i$ is the transpose of the i-th row of $\mathbf{Z}$ (thus a column vector). We denote $\mathcal{S}^{p-1}$ the $(p-1)$-dimensional unit sphere and $\mathcal{B}\left(\mathbf{0}, r\right)$ the ball centered at $\mathbf{0}$ with radius r.

In this paper we assume every patch $\mathbf{Z}_i$ is vector of IID Gaussian random variables. The following theorem gives an explicit formula for the population loss. The proof uses basic rotational invariant property and polar decomposition of Gaussian random variables. See Section A for details.

Theorem 2.1. *For $\mathbf{Z} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$, the population loss is*

$$\begin{aligned}\ell\left(\mathbf{v}, \mathbf{a}\right) =& \frac{1}{2}\left[\frac{(\pi-1)\left\|\mathbf{w}^*\right\|_2^2}{2\pi}\left\|\mathbf{a}^*\right\|_2^2 + \frac{(\pi-1)}{2\pi}\left\|\mathbf{a}\right\|_2^2 - \frac{2\left(g\left(\phi\right)-1\right)\left\|\mathbf{w}^*\right\|_2}{2\pi}\mathbf{a}^\top\mathbf{a}^*\right.\\ &\left. + \frac{\left\|\mathbf{w}^*\right\|_2^2}{2\pi}\left(\mathbf{1}^\top\mathbf{a}^*\right)^2 + \frac{1}{2\pi}\left(\mathbf{1}^\top\mathbf{a}\right)^2 - 2\left\|\mathbf{w}^*\right\|_2\mathbf{1}^\top\mathbf{a}\cdot\mathbf{1}^\top\mathbf{a}^*\right] \end{aligned} \tag{3}$$

where $\phi = \theta\left(\mathbf{v}, \mathbf{w}^*\right)$ and $g(\phi) = \cos(\phi)(\pi - \phi + \sin(\phi)\cos(\phi)) + \sin^3(\phi)$.

Using similar techniques, we can show the gradient also has an analytical form.

[1] Lee et al. [2016] showed vanilla gradient descent only converges to minimizers with no convergence rates guarantees. Recently, Du et al. [2017a] gave an exponential time lower bound for the vanilla gradient descent. In this paper, we give polynomial convergence guarantee on vanilla gradient descent.

Theorem 2.2. *If* $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ *and denote* $\theta(\mathbf{w}, \mathbf{w}^*) = \phi$ *then the expected gradient of* $\mathbf{w}$ *and* $\mathbf{a}$ *can be written as*

$$\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell(\mathbf{Z}, \mathbf{v}, \mathbf{a})}{\partial \mathbf{v}}\right] = -\frac{1}{2\pi \|\mathbf{v}\|_2}\left(\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right)\mathbf{a}^\top \mathbf{a}^* (\pi - \phi)\,\mathbf{w}^*$$

$$\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell(\mathbf{Z}, \mathbf{v}, \mathbf{a})}{\partial \mathbf{a}}\right] = \frac{1}{2\pi}\left(\mathbf{1}\mathbf{1}^\top + (\pi - 1)\mathbf{I}\right)\mathbf{a} - \frac{1}{2\pi}\|\mathbf{w}^*\|_2\left(\mathbf{1}\mathbf{1}^\top + (g(\phi) - 1)\mathbf{I}\right)\mathbf{a}^*.$$

As a remark, if the second layer is fixed, upon proper scaling, the formulas for the population loss and gradient of $\mathbf{v}$ are equivalent to the corresponding formulas derived in [Brutzkus and Globerson, 2017, Cho and Saul, 2009]. However, when the second layer is not fixed, the gradient of $\mathbf{v}$ depends $\mathbf{a}^\top \mathbf{a}^*$, which plays an important role in deciding whether converging to the global or the local minimum.

3 Main Result

We begin with our main theorem about the convergence of gradient descent.

Theorem 3.1. *Suppose the initialization satisfies* $\left(\mathbf{a}^0\right)^\top \mathbf{a}^* > 0$, $\left|\mathbf{1}^\top \mathbf{a}^0\right| \le \left|\mathbf{1}^\top \mathbf{a}^*\right|$, $\phi^0 < \pi/2$ *and step size satisfies* $\eta = O\left(\min\left\{\frac{\left(\mathbf{a}^0\right)^\top \mathbf{a}^* \cos\phi^0}{\left(\|\mathbf{a}^*\|_2^2 + \left(\mathbf{1}^\top \mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2}, \frac{(g(\phi_0) - 1)\|\mathbf{a}^*\|_2^2 \cos\phi^0}{\left(\|\mathbf{a}^*\|_2^2 + \left(\mathbf{1}^\top \mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2}, \frac{\cos\phi^0}{\left(\|\mathbf{a}^*\|_2^2 + \left(\mathbf{1}^\top \mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2}, \frac{1}{k}\right\}\right)$.
Then the convergence of gradient descent has two phases.
(Phase I: Slow Initial Rate) *There exists* $T_1 = O\left(\frac{1}{\eta \cos\phi^0 \beta^0} + \frac{1}{\eta}\right)$ *such that we have* $\phi^{T_1} = \Theta(1)$ *and* $\left(\mathbf{a}^{T_1}\right)^\top \mathbf{a}^* \|\mathbf{w}^*\|_2 = \Theta\left(\|\mathbf{a}^*\|_2^2 \|\mathbf{w}^*\|_2^2\right)$ *where* $\beta^0 = \min\left\{\left(\mathbf{a}^0\right)^\top \mathbf{a}^* \|\mathbf{w}^*\|_2, (g(\phi^0) - 1)\|\mathbf{a}^*\|_2^2 \|\mathbf{a}^*\|_2^2\right\}$.
(Phase II: Fast Rate) *There exists* $T_2 = \widetilde{O}(\left(\frac{1}{\eta \|\mathbf{w}^*\|_2^2 \|\mathbf{a}^*\|_2^2} + \frac{1}{\eta}\right)\log\left(\frac{1}{\epsilon}\right))^2$ *such that* $\ell\left(\mathbf{v}^{T_1+T_2}, \mathbf{a}^{T_1+T_2}\right) \le \epsilon \|\mathbf{w}^*\|_2^2 \|\mathbf{a}^*\|_2^2$.

Theorem 3.1 shows under certain conditions of the initialization, gradient descent converges to the global minimum. The convergence has phases, at the beginning because the initial signal is small, the convergence is quite slow. After T_1 iterations, the signal becomes stronger and we enter a regime with a faster convergence rate. More technical insights are provided in Section 4.

Initialization plays an important role in the convergence. First, Theorem 3.1 needs the initialization satisfies $\left(\mathbf{a}^0\right)^\top \mathbf{a}^* > 0$, $\left|\mathbf{1}^\top \mathbf{a}^0\right| \le \left|\mathbf{1}^\top \mathbf{a}^*\right|$ and $\phi^0 < \pi/2$. Second, the step size η and the convergence rate in the first phase also depends on the initialization. If the initial signal is very small, for example, $\phi^0 \approx \pi/2$ which makes $\cos\phi^0$ close to 0, we can only choose a very small step size and because T_1 depends on the inverse of $\cos\phi^0$, we need a large number of iterations to enter phase II. We provide the following initialization scheme which ensures the conditions required by Theorem 3.1 and a large enough initial signal.

Theorem 3.2. *Let* $\mathbf{v} \sim \operatorname{unif}\left(\mathcal{S}^{p-1}\right)$ *and* $\mathbf{a} \sim \operatorname{unif}\left(\mathcal{B}\left(\mathbf{0}, \frac{\left|\mathbf{1}^\top \mathbf{a}^*\right| \|\mathbf{w}^*\|_2}{\sqrt{k}}\right)\right)$, *then exists*

$$\left(\mathbf{v}^0, \mathbf{a}^0\right) \in \{(\mathbf{v}, \mathbf{a}), (\mathbf{v}, -\mathbf{a}), (-\mathbf{v}, \mathbf{a}), (-\mathbf{v}, -\mathbf{a})\}$$

[2] $\widetilde{O}(\cdot)$ hides factors on $\left|\mathbf{1}^\top \mathbf{a}^*\right| \|\mathbf{w}^*\|_2$ and $\|\mathbf{a}^*\|_2 \|\mathbf{w}^*\|_2$

that $\left(\mathbf{a}^0\right)^\top \mathbf{a}^* > 0$*,* $\left|\mathbf{1}^\top \mathbf{a}^0\right| \le \left|\mathbf{1}^\top \mathbf{a}^*\right|$ *and* $\phi^0 < \pi/2$*. Further, with high probability, the initialization satisfies* $\left(\mathbf{a}^0\right)^\top \mathbf{a}^* \|\mathbf{w}^*\|_2 = \Theta\left(\frac{\left|\mathbf{1}^\top \mathbf{a}^*\right| \|\mathbf{a}^*\|_2 \|\mathbf{w}^*\|_2^2}{k}\right)$*, and* $\phi^0 = \Theta\left(\frac{1}{\sqrt{p}}\right)$*.*

Theorem 3.2 shows after generating a pair of random vectors $(\mathbf{v}, \mathbf{a})$, trying out all 4 sign combinations of $(\mathbf{v}, \mathbf{a})$, we can find the global minimum by gradient descent. Further, because the initial signal is not too small, we only need to set the step size to be $\widetilde{O}(1/\text{poly}(k,p))$ and the number of iterations in phase I is at most $\widetilde{O}\left(\text{poly}\left(k,p\right)\right)$. Therefore, Theorem 3.1 and Theorem 3.2 together show that randomly initialized gradient descent learns an one-hidden-layer convolutional neural network in polynomial time. The proof of the first part of Theorem 3.2 just uses the symmetry of unit sphere and ball and the second part is a standard application of random vector in high-dimensional spaces. See Lemma 2.5 of [Hardt and Price, 2014] for example.

Remark 1: For the second layer we use $O\left(\frac{1}{\sqrt{k}}\right)$ type initialization, verifying common initialization techniques [Glorot and Bengio, 2010, He et al., 2015, LeCun et al., 1998].

Remark 2: The Gaussian input assumption is not necessarily true in practice, although this is a common assumption appeared in the previous papers [Brutzkus and Globerson, 2017, Li and Yuan, 2017, Zhong et al., 2017a,b, Tian, 2017, Choromanska et al., 2015a, Xie et al., 2017, Shalev-Shwartz et al., 2017b] and also considered plausible in [Choromanska et al., 2015b]. Our result can be easily generalized to rotation invariant distributions. However, extending to more general distributional assumption, e.g., structural conditions used in [Du et al., 2017b] remains a challenging open problem.

3.1 Gradient Descent Can Converge to the Spurious Local Minimum

Theorem 3.2 shows among $\{(\mathbf{v}, \mathbf{a}), (\mathbf{v}, -\mathbf{a}), (-\mathbf{v}, \mathbf{a}), (-\mathbf{v}, -\mathbf{a})\}$, there is one pair enables gradient descent to converge to the global minimum. Perhaps surprisingly, the next theorem shows, under some conditions of the underlying truth, there is also a pair that makes gradient descent converge to the spurious local minimum. See Figure 1b for the difference between the global minimum and the spurious local minimum.

Theorem 3.3. *Without loss of generality, we let* $\|\mathbf{w}^*\|_2 = 1$*. Suppose* $\left(\mathbf{1}^\top \mathbf{a}^*\right)^2 < \frac{1}{\text{poly}(p)} \|\mathbf{a}^*\|_2^2$ *and* η *is sufficiently small. Let* $\mathbf{v} \sim \text{unif}\left(\mathcal{S}^{p-1}\right)$ *and* $\mathbf{a} \sim \text{unif}\left(\mathcal{B}\left(\mathbf{0}, \frac{\mathbf{1}^\top \mathbf{a}^* \|\mathbf{w}^*\|_2}{\sqrt{k}}\right)\right)$*, then with high probability, there exists* $\left(\mathbf{v}^0, \mathbf{a}^0\right) \in \{(\mathbf{v}, \mathbf{a}), (\mathbf{v}, -\mathbf{a}), (-\mathbf{v}, \mathbf{a}), (-\mathbf{v}, -\mathbf{a})\}$ *that* $\left(\mathbf{a}^0\right)^\top \mathbf{a}^* < 0$*,* $\left|\mathbf{1}^\top \mathbf{a}^0\right| \le \left|\mathbf{1}^\top \mathbf{a}^*\right|$*,* $g\left(\phi^0\right) \le \frac{-2\left(\mathbf{1}^\top \mathbf{a}^*\right)^2}{\|\mathbf{a}^*\|_2^2} + 1$*. If* $\left(\mathbf{v}^0, \mathbf{a}^0\right)$ *is used as the initialization, when Algorithm 1 converges, we have* $\ell\left(\mathbf{v}, \mathbf{a}\right) = \Omega\left(\|\mathbf{w}^*\|_2^2 \|\mathbf{a}^*\|_2^2\right)$*.*

4 Overview of Proofs and Technical Insights

In this section we list our main ideas of proving Theorem 3.1. In Section 4.1, we give qualitative high level intuition on why the initial conditions are sufficient for gradient descent to converge to the global minimum. In Section 4.2, we explain why the gradient descent has two phases.

4.1 Qualitative Analysis of Convergence

The convergence to global optimum relies on a geometric characterization of saddle points and a series of invariants throughout the gradient descent dynamics. The next lemma gives the analysis of stationary points. The proof is by checking the first order condition of stationary points using Theorem 2.2.

Lemma 4.1 (Stationary Point Analysis)**.** *When the gradient descent converges, if* $\mathbf{a}^\top\mathbf{a}^* \neq 0$ *and* $\|\mathbf{v}\|_2 < \infty$*, we have either*

$$\theta\left(\mathbf{v}, \mathbf{w}^*\right) = 0, \mathbf{a} = \left\|\mathbf{w}^*\right\|_2 \mathbf{a}^*$$
$$\text{or } \theta\left(\mathbf{v}, \mathbf{w}^*\right) = \pi, \mathbf{a} = \left(\mathbf{1}\mathbf{1}^\top + (\pi - 1)\mathbf{I}\right)^{-1}\left(\mathbf{1}\mathbf{1}^\top - \mathbf{I}\right)\left\|\mathbf{w}^*\right\|_2 \mathbf{a}^*.$$

This lemma shows when the algorithm converge, and $\mathbf{a}$ and $\mathbf{a}^*$ are not orthogonal, then we arrive at either a global optimal point or a local minimum. Now recall the gradient formula of $\mathbf{v}$: $\frac{\partial \ell(\mathbf{v},\mathbf{a})}{\partial \mathbf{v}} = -\frac{1}{2\pi\|\mathbf{v}\|_2}\left(\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right)\mathbf{a}^\top\mathbf{a}^*\left(\pi - \phi\right)\mathbf{w}^*$. Notice that $\phi \leq \pi$ and $\left(\mathbf{I} - \frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right)$ is just the projection matrix onto the complement of $\mathbf{v}$. Therefore, the sign of inner product between $\mathbf{a}$ and $\mathbf{a}^*$ plays a crucial role in the dynamics of Algorithm 1 because if the inner product is positive, the gradient update will decrease the angle between $\mathbf{v}$ and $\mathbf{w}^*$ and if it is negative, the angle will increase. This observation is formalized in the lemma below.

Lemma 4.2 (Invariance I: Angle between $\mathbf{v}$ and $\mathbf{w}$ always decreases.)**.** *If* $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* > 0$*, then* $\phi^{t+1} \leq \phi^t$*.*

This lemmas shows if $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* > 0$ for all t, we have the convergence toward the global minimum. Thus, we need to study the dynamics of $\left(\mathbf{a}^t\right)^\top \mathbf{a}^*$. For the ease of presentation, without loss of generality, we assume $\|\mathbf{w}^*\|_2 = 1$. By the gradient formula of $\mathbf{a}$, we have

$$\begin{aligned}&\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* \\ =&\left(1 - \frac{\eta(\pi-1)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^* + \frac{\eta(g(\phi^t)-1)}{2\pi}\left\|\mathbf{a}^t\right\|_2^2 + \frac{\eta}{2\pi}\left(\left(\mathbf{1}^\top\mathbf{a}^*\right)^2 - \left(\mathbf{1}^\top\mathbf{a}^t\right)\left(\mathbf{1}^\top\mathbf{a}^*\right)\right).\end{aligned} \tag{4}$$

We can use the induction to prove the invariance. If $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* > 0$ and $\phi^t < \frac{\pi}{2}$ the first term of Equation (4) is non-negative. For the second term, notice that if $\phi^t < \frac{\pi}{2}$, we have $g(\phi^t) > 1$, so the second term is non-negative. Therefore, as long as $\left(\left(\mathbf{1}^\top\mathbf{a}^*\right)^2 - \left(\mathbf{1}^\top\mathbf{a}^t\right)\left(\mathbf{1}^\top\mathbf{a}^*\right)\right)$ is also non-negative, we have the desired invariance. The next lemma summarizes the above analysis.

Lemma 4.3 (Invariance II: Positive Signal from the Second Layer.)**.** *If* $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* > 0$*,* $0 \leq \mathbf{1}^\top\mathbf{a}^* \cdot \mathbf{1}^\top\mathbf{a}^t \leq \left(\mathbf{1}^\top\mathbf{a}^*\right)^2$*,* $0 < \phi^t < \pi/2$ *and* $\eta < 2$*, then* $\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* > 0$*.*

It remains to prove $\left(\left(\mathbf{1}^\top\mathbf{a}^*\right)^2 - \left(\mathbf{1}^\top\mathbf{a}^t\right)\left(\mathbf{1}^\top\mathbf{a}^*\right)\right) > 0$. Again, we study the dynamics of this quantity. Using the gradient formula and some routine algebra, we have

$$\mathbf{1}^\top\mathbf{a}^{t+1} \cdot \mathbf{1}^\top\mathbf{a}^* \leq \left(1 - \frac{\eta\left(k-\pi-1\right)}{2\pi}\right)\mathbf{1}^\top\mathbf{a}^t \cdot \mathbf{1}^\top\mathbf{a}^* + \frac{\eta\left(k + g(\phi^t) - 1\right)}{2}\left(\mathbf{1}^\top\mathbf{a}^*\right)^2$$

$$\leq \left(1 - \frac{\eta\left(k-\pi-1\right)}{2\pi}\right) \mathbf{1}^\top \mathbf{a}^t \cdot \mathbf{1}^\top \mathbf{a}^* + \frac{\eta\left(k+\pi-1\right)}{2}\left(\mathbf{1}^\top \mathbf{a}^*\right)^2$$

where have used the fact that $g(\phi) \leq \pi$ for all $0 \leq \phi \leq \frac{\pi}{2}$. Therefore we have

$$\left(\mathbf{1}^\top \mathbf{a}^* - \mathbf{1}^\top \mathbf{a}^{t+1}\right) \cdot \mathbf{1}^\top \mathbf{a}^* \geq \left(1 - \frac{\eta(k+\pi-1)}{2\pi}\right)\left(\mathbf{1}^\top \mathbf{a}^* - \mathbf{1}^\top \mathbf{a}^t\right) \mathbf{1}^\top \mathbf{a}^*.$$

Now we have the third invariance.

Lemma 4.4 (Invariance III: Summation of Second Layer Always Small.)**.** *If* $\mathbf{1}^\top \mathbf{a}^* \cdot \mathbf{1}^\top \mathbf{a}^t \leq \left(\mathbf{1}^\top \mathbf{a}^*\right)^2$ *and* $\eta < \frac{2\pi}{k+\pi-1}$ *then* $\mathbf{1}^\top \mathbf{a}^* \cdot \mathbf{1}^\top \mathbf{a}^{t+1} \leq \left(\mathbf{1}^\top \mathbf{a}^*\right)^2$.

To sum up, if the initialization satisfies (1) $\phi^0 < \frac{\pi}{2}$, (2) $\left(\mathbf{a}^0\right)^\top \mathbf{a}^* > 0$ and (3) $\mathbf{1}^\top \mathbf{a}^* \cdot \mathbf{1}^\top \mathbf{a}^0 \leq \left(\mathbf{1}^\top \mathbf{a}^*\right)^2$, with Lemma 4.2, 4.3, 4.4, by induction we can show the convergence to the global minimum. Further, Theorem 3.2 shows theses three conditions are true with constant probability using random initialization.

4.2 Quantitative Analysis of Two Phase Phenomenon

In this section we demonstrate why there is a two-phase phenomenon. Throughout this section, we assume the conditions in Section 4.1 hold. We first consider the convergence of the first layer. Because we are using weight-normalization, only the angle between $\mathbf{v}$ and $\mathbf{w}^*$ will affect the prediction. Therefore, in this paper, we study the dynamics $\sin^2 \phi^t$. The following lemma quantitatively characterize the shrinkage of this quantity of one iteration.

Lemma 4.5 (Convergence of Angle between $\mathbf{v}$ and $\mathbf{w}^*$)**.** *Under the same assumptions as in Theorem 3.1. Let* $\beta^0 = \min\left\{\left(\mathbf{a}^0\right)^\top \mathbf{a}^*, \left(g(\phi^0)-1\right)\|\mathbf{a}^*\|_2^2\right\}\|\mathbf{w}^*\|_2^2$. *If the step size satisfies* $\eta = O\left(\min\left\{\frac{\beta^* \cos\phi^0}{\left(\|\mathbf{a}^*\|_2^2+\left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2}, \frac{\cos\phi^0}{\left(\|\mathbf{a}^*\|_2^2+\left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2}, \frac{1}{k}\right\}\right)$, *we have*

$$\sin^2 \phi^{t+1} \leq \left(1 - \eta \cos\phi^t \lambda^t\right) \sin^2 \phi^t$$

where $\lambda^t = \frac{\|\mathbf{w}^*\|_2\left(\pi-\phi^t\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^*}{2\pi\|\mathbf{v}^t\|_2^2}$.

This lemma shows the convergence rate depends on two crucial quantities, $\cos\phi^t$ and λ^t. At the beginning, both $\cos\phi^t$ and λ^t are small. Nevertheless, Lemma C.3 shows λ^t is universally lower bounded by $\Omega\left(\beta^0\right)$. Therefore, after $O(\frac{1}{\eta\cos\phi^0\beta^0})$ we have $\cos\phi^t = \Omega\left(1\right)$. Once $\cos\phi^t = \Omega\left(1\right)$, Lemma C.2 shows, after $O\left(\frac{1}{\eta}\right)$ iterations, $\left(\mathbf{a}^t\right)\mathbf{a}^*\|\mathbf{w}^*\| = \Omega\left(\|\mathbf{w}^*\|_2^2\|\mathbf{a}^*\|_2^2\right)$. Combining the facts $\left\|\mathbf{v}^t\right\|_2 \leq 2$ (Lemma C.3) and $\phi^t < \pi/2$, we have $\cos\phi^t\lambda^t = \Omega\left(\|\mathbf{w}^*\|_2^2\|\mathbf{a}^*\|_2^2\right)$. Now we enter phase II.

In phase II, Lemma 4.5 shows

$$\sin^2 \phi^{t+1} \leq \left(1 - \eta C\|\mathbf{w}^*\|_2^2\|\mathbf{a}^*\|_2^2\right)\sin^2 \phi^t$$

for some positive absolute constant C. Therefore, we have much faster convergence rate than that in the Phase I. After only $O\left(\frac{1}{\eta\|\mathbf{w}^*\|_2^2\|\mathbf{a}\|_2^2}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations, we obtain $\phi^t \leq \epsilon$. Once we have this, using Lemma C.4, Lemma C.5 Lemma C.6 we can show after $\widetilde{O}(\frac{1}{\eta}\log\left(\frac{1}{\epsilon}\right))$ iterations, the population loss is smaller than $\epsilon\|\mathbf{v}^*\|_2^2\|\mathbf{a}^*\|_2^2$.

5 Conclusions and Future Works

In this paper we give the first polynomial convergence guarantee of randomly initialized gradient descent algorithm for learning a one-hidden-layer convolutional neural network. Our result reveals an interesting phenomenon that randomly initialized local search algorithm can converge to a global minimum or a spurious local minimum and both events have constant probability. We give a complete quantitative characterization of gradient descent dynamics to explain the two-phase convergence phenomenon. Here we list some future directions.

Our analysis focused on the population loss with Gaussian input. In practice one uses (stochastic) gradient descent on the empirical loss. Concentration results in [Mei et al., 2016, Soltanolkotabi, 2017, Daskalakis et al., 2016, Xu et al., 2016] are useful to generalize our results to the empirical version. A more challenging question is how to extend the analysis of gradient dynamics beyond rotationally invariant input distributions. Du et al. [2017b] proved the convergence of gradient descent under some structural input distribution assumptions in the one-layer convolutional neural network. It would be interesting to bring their insights to our setting.

Another interesting direction is to generalize our result to deeper and wider architectures. Specifically, an open problem is under what conditions randomly initialized gradient descent algorithms can learn one-hidden-layer fully connected neural network or a convolutional neural network with multiple kernels. Existing results often requires sufficiently good initialization [Zhong et al., 2017a,b]. We believe the insights from this paper, especially the invariance principles in Section 4.1 are helpful to understand the behaviors of gradient-based algorithms in these settings.

6 Acknowledgment

The authors thank Xiaolong Wang and Kai Zhong for useful discussions.

References

Zeyuan Allen-Zhu. Natasha 2: Faster non-convex optimization than SGD. *arXiv preprint arXiv:1708.08694*, 2017.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

Avrim Blum and Ronald L Rivest. Training a 3-node neural network is NP-complete. In *Advances in neural information processing systems*, pages 494–501, 1989.

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

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Accelerated methods for non-convex optimization. *arXiv preprint arXiv:1611.00756*, 2016.

Youngmin Cho and Lawrence K Saul. Kernel methods for deep learning. In *Advances in neural information processing systems*, pages 342–350, 2009.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pages 192–204, 2015a.

Anna Choromanska, Yann LeCun, and Gérard Ben Arous. Open problem: The landscape of the loss surfaces of multilayer networks. In *Conference on Learning Theory*, pages 1756–1760, 2015b.

Frank E Curtis, Daniel P Robinson, and Mohammadreza Samadi. A trust region algorithm with a worst-case iteration complexity of $O(\epsilon^{-3/2})$ for nonconvex optimization. *Mathematical Programming*, pages 1–32, 2014.

Constantinos Daskalakis, Christos Tzamos, and Manolis Zampetakis. Ten steps of EM suffice for mixtures of two gaussians. *arXiv preprint arXiv:1609.00368*, 2016.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. *arXiv preprint arXiv:1612.08083*, 2016.

Simon S Du, Chi Jin, Jason D Lee, Michael I Jordan, Barnabas Poczos, and Aarti Singh. Gradient descent can take exponential time to escape saddle points. *arXiv preprint arXiv:1705.10412*, 2017a.

Simon S Du, Jason D Lee, and Yuandong Tian. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017b.

Soheil Feizi, Hamid Javadi, Jesse Zhang, and David Tse. Porcupine neural networks:(almost) all local optima are global. *arXiv preprint arXiv:1710.02196*, 2017.

C Daniel Freeman and Joan Bruna. Topology and geometry of half-rectified network optimization. *arXiv preprint arXiv:1611.01540*, 2016.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points − online stochastic gradient for tensor decomposition. In *Proceedings of The 28th Conference on Learning Theory*, pages 797–842, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1233–1242, 2017a.

Rong Ge, Jason D Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. *arXiv preprint arXiv:1711.00501*, 2017b.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256, 2010.

Surbhi Goel and Adam Klivans. Eigenvalue decay implies polynomial-time learnability for neural networks. *arXiv preprint arXiv:1708.03708*, 2017a.

Surbhi Goel and Adam Klivans. Learning depth-three neural networks in polynomial time. *arXiv preprint arXiv:1709.06010*, 2017b.

Surbhi Goel, Varun Kanade, Adam Klivans, and Justin Thaler. Reliably learning the ReLU in polynomial time. *arXiv preprint arXiv:1611.10258*, 2016.

Benjamin D Haeffele and René Vidal. Global optimality in tensor factorization, deep learning, and beyond. *arXiv preprint arXiv:1506.07540*, 2015.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems*, pages 2861–2869, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods. *arXiv preprint arXiv:1506.08473*, 2015.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1724–1732, 2017.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances In Neural Information Processing Systems*, pages 586–594, 2016.

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

Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 1998.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

Kfir Y Levy. The power of normalization: Faster evasion of saddle points. *arXiv preprint arXiv:1611.04831*, 2016.

Xingguo Li, Zhaoran Wang, Junwei Lu, Raman Arora, Jarvis Haupt, Han Liu, and Tuo Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*, 2016.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with ReLU activation. *arXiv preprint arXiv:1705.09886*, 2017.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

Song Mei, Yu Bai, and Andrea Montanari. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2016.

Yurii Nesterov and Boris T Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Behnam Neyshabur, Ruslan R Salakhutdinov, and Nati Srebro. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2015.

Quynh Nguyen and Matthias Hein. The loss surface of deep and wide neural networks. *arXiv preprint arXiv:1704.08045*, 2017a.

Quynh Nguyen and Matthias Hein. The loss surface and expressivity of deep convolutional neural networks. *arXiv preprint arXiv:1710.10928*, 2017b.

Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

Itay Safran and Ohad Shamir. On the quality of the initial basin in overspecified neural networks. In *International Conference on Machine Learning*, pages 774–782, 2016.

Tim Salimans and Diederik P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *Advances in Neural Information Processing Systems*, pages 901–909, 2016.

Hanie Sedghi and Anima Anandkumar. Provable methods for training neural networks with sparse connectivity. *arXiv preprint arXiv:1412.2693*, 2014.

Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Failures of gradient-based deep learning. In *International Conference on Machine Learning*, pages 3067–3075, 2017a.

Shai Shalev-Shwartz, Ohad Shamir, and Shaked Shammah. Weight sharing is crucial to succesful optimization. *arXiv preprint arXiv:1706.00687*, 2017b.

Ohad Shamir. Distribution-specific hardness of learning neural networks. *arXiv preprint arXiv:1609.01037*, 2016.

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):484–489, 2016.

Jiří Šíma. Training a single sigmoidal neuron is hard. *Neural Computation*, 14(11):2709–2728, 2002.

Mahdi Soltanolkotabi. Learning ReLUs via gradient descent. *arXiv preprint arXiv:1705.04591*, 2017.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 2379–2383. IEEE, 2016.

Ju Sun, Qing Qu, and John Wright. Complete dictionary recovery over the sphere I: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

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.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 964–973. JMLR. org, 2016.

Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. In *Artificial Intelligence and Statistics*, pages 1216–1224, 2017.

Ji Xu, Daniel J Hsu, and Arian Maleki. Global analysis of expectation maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684, 2016.

Yuchen Zhang, Jason D Lee, Martin J Wainwright, and Michael I Jordan. Learning halfspaces and neural networks with random initialization. *arXiv preprint arXiv:1511.07948*, 2015.

Kai Zhong, Zhao Song, and Inderjit S Dhillon. Learning non-overlapping convolutional neural networks with multiple kernels. *arXiv preprint arXiv:1711.03440*, 2017a.

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*, 2017b.

Pan Zhou and Jiashi Feng. The landscape of deep learning algorithms. *arXiv preprint arXiv:1705.07038*, 2017.

A Proofs of Section 2

Proof of Theorem 2.1. We first expand the loss function directly.

$$\begin{aligned}
&\ell(\mathbf{v},\mathbf{a})\\
=&\mathbb{E}\left[\frac{1}{2}\left(y-\mathbf{a}^\top\sigma(\mathbf{Z})\mathbf{w}\right)^2\right]\\
=&(\mathbf{a}^*)^\top\mathbb{E}\left[\sigma(\mathbf{Z}\mathbf{w}^*)\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]\mathbf{a}^*+\mathbf{a}^\top\mathbb{E}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w})^\top\right]\mathbf{a}-2\mathbf{a}^\top\mathbb{E}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]\mathbf{a}^*\\
=&(\mathbf{a}^*)^\top\mathbf{A}(\mathbf{w}^*)\mathbf{a}^*+\mathbf{a}^\top\mathbf{A}(\mathbf{w})\mathbf{a}-2\mathbf{a}^\top\mathbf{B}(\mathbf{w},\mathbf{w}^*)\mathbf{w}^*.
\end{aligned}$$

where for simplicity, we denote

$$\begin{aligned}
\mathbf{A}(\mathbf{w})=&\mathbb{E}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w})^\top\right] &(5)\\
\mathbf{B}(\mathbf{w},\mathbf{w}^*)=&\mathbb{E}\left[\sigma(\mathbf{Z}\mathbf{w})\sigma(\mathbf{Z}\mathbf{w}^*)^\top\right]. &(6)
\end{aligned}$$

For $i\neq j$, using the second identity of Lemma A.1, we can compute

$$\mathbf{A}(\mathbf{w})_{ij}=\mathbb{E}\left[\sigma\left(\mathbf{Z}_i^\top\mathbf{w}\right)\right]\mathbb{E}\left[\sigma\left(\mathbf{Z}_j^\top\mathbf{w}\right)\right]=\frac{1}{2\pi}\|\mathbf{w}\|_2^2$$

For $i=j$, using the second moment formula of half-Gaussian distribution we can compute

$$\mathbf{A}(\mathbf{w})_{ii}=\frac{1}{2}\|\mathbf{w}\|_2^2.$$

Therefore

$$\mathbf{A}(\mathbf{w})=\frac{1}{2\pi}\|\mathbf{w}\|_2^2\left(\mathbf{1}\mathbf{1}^\top+(\pi-1)\mathbf{I}\right).$$

Now let us compute $\mathbf{B}(\mathbf{w},\mathbf{w}_*)$. For $i\neq j$, similar to $\mathbf{A}(\mathbf{w})_{ij}$, using the independence property of Gaussian, we have

$$\mathbf{B}(\mathbf{w},\mathbf{w}_*)_{ij}=\frac{1}{2\pi}\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2.$$

Next, using the fourth identity of Lemma A.1, we have

$$\mathbf{B}(\mathbf{w},\mathbf{w}^*)_{ii}=\frac{1}{2\pi}\left(\cos\phi(\pi-\phi+\sin\phi\cos\phi)+\sin^3\phi\right)\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2.$$

Therefore, we can also write $\mathbf{B}(\mathbf{w},\mathbf{w}^*)$ in a compact form

$$\mathbf{B}(\mathbf{w},\mathbf{w}^*)=\frac{1}{2\pi}\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2\left(\mathbf{1}\mathbf{1}^\top+\left(\cos\phi(\pi-\phi+\sin\phi\cos\phi)+\sin^3\phi-1\right)\mathbf{I}\right).$$

Plugging in the formulas of $\mathbf{A}(\mathbf{w})$ and $\mathbf{B}(\mathbf{w},\mathbf{w}^*)$ and $\mathbf{w}=\frac{\mathbf{v}}{\|\mathbf{v}\|_2}$, we obtain the desired result. ☐

Proof of Theorem 2.2. We first compute the expect gradient for $\mathbf{v}$. From[Salimans and Kingma, 2016], we know

$$\frac{\partial\ell(\mathbf{v},\mathbf{a})}{\partial\mathbf{v}}=\frac{1}{\|\mathbf{v}\|_2}\left(\mathbf{I}-\frac{\mathbf{v}\mathbf{v}^\top}{\|\mathbf{v}\|_2^2}\right)\frac{\partial\ell(\mathbf{w},\mathbf{a})}{\partial\mathbf{w}}.$$

Recall the gradient formula,

$$
\begin{aligned}
&\frac{\partial \ell\left(\mathbf{Z}, \mathbf{w}, \mathbf{a}\right)}{\partial \mathbf{w}} \\
=&\left(\sum_{i=1}^{k} a_i^* \sigma\left(\mathbf{Z}_i \mathbf{w}\right) - \sum_{i=1}^{k} a_i^* \sigma\left(\mathbf{Z}\mathbf{w}^*\right)\right)\left(\sum_{i=1}^{k} a_i \mathbf{Z}_i \mathbb{I}\left\{\mathbf{Z}_i^\top \mathbf{w}\right\}\right) \\
=&\left(\sum_{i=1}^{k} a_i^2 \mathbf{Z}_i \mathbf{Z}_i^\top \mathbb{I}\left\{\mathbf{Z}_i^\top \mathbf{w} \geq 0\right\} + \sum_{i \neq j} a_i a_j \mathbf{Z}_i \mathbf{Z}_j^\top \mathbb{I}\left\{\mathbf{Z}_i^\top \mathbf{w} \geq 0, \mathbf{Z}_j^\top \mathbf{w} \geq 0\right\}\right) \mathbf{w} && (7) \\
-&\left(\sum_{i=1}^{k} a_i a_i^* \mathbf{Z}_i \mathbf{Z}_i^\top \mathbb{I}\left\{\mathbf{Z}_i^\top \mathbf{w} \geq 0, \mathbf{Z}_i^\top \mathbf{w}^* \geq 0\right\} + \sum_{i \neq j} a_i a_j^* \mathbf{Z}_i \mathbf{Z}_j^* \mathbb{I}\left\{\mathbf{Z}_i^\top \mathbf{w} \geq 0, \mathbf{Z}_j^\top \mathbf{w}^* \geq 0\right\}\right) \mathbf{w}^*. && (8)
\end{aligned}
$$

Now we calculate expectation of Equation (7) and (8) separately. For (7), by first two formulas of Lemma A.1, we have

$$
\begin{aligned}
&\left(\sum_{i=1}^{k} a_i^2 \mathbf{Z}_i \mathbf{Z}_i^\top \mathbb{I}\left\{\mathbf{Z}_i^\top \mathbf{w} \geq 0\right\} + \sum_{i \neq j} a_i a_j \mathbf{Z}_i \mathbf{Z}_j^\top \mathbb{I}\left\{\mathbf{Z}_i^\top \mathbf{w} \geq 0, \mathbf{Z}_j^\top \mathbf{w} \geq 0\right\}\right) \mathbf{w} \\
=&\sum_{i=1}^{k} a_i^2 \cdot \frac{\mathbf{w}}{2} + \sum_{i \neq j} a_i a_j \frac{\mathbf{w}}{2\pi}.
\end{aligned}
$$

For (8), we use the second and third formula in Lemma A.1 to obtain

$$
\begin{aligned}
&\left(\sum_{i=1}^{k} a_i a_i^* \mathbf{Z}_i \mathbf{Z}_i^\top \mathbb{I}\left\{\mathbf{Z}_i^\top \mathbf{w} \geq 0, \mathbf{Z}_i^\top \mathbf{w}^* \geq 0\right\} + \sum_{i \neq j} a_i a_j^* \mathbf{Z}_i \mathbf{Z}_j^* \mathbb{I}\left\{\mathbf{Z}_i^\top \mathbf{w} \geq 0, \mathbf{Z}_j^\top \mathbf{w}^* \geq 0\right\}\right) \mathbf{w}^* \\
=&\mathbf{a}^\top \mathbf{a}^* \left(\frac{1}{\pi}\left(\pi - \phi\right) \mathbf{w}^* + \frac{1}{\pi} \sin\phi \frac{\left\|\mathbf{w}^*\right\|_2}{\left\|\mathbf{w}\right\|_2} \mathbf{w}\right) + \sum_{i \neq j} a_i a_j^* \frac{1}{2\pi} \frac{\left\|\mathbf{w}^*\right\|_2}{\left\|\mathbf{w}\right\|_2} \mathbf{w}.
\end{aligned}
$$

In summary, aggregating them together we have

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{Z}}\left[\frac{\partial \ell\left(\mathbf{Z}, \mathbf{w}, \mathbf{a}\right)}{\partial \mathbf{w}}\right] \\
=&\frac{1}{2\pi} \mathbf{a}^\top \mathbf{a}^* \left(\pi - \phi\right) \mathbf{w}^* + \left(\frac{\left\|\mathbf{a}\right\|_2^2}{2} + \frac{\sum_{i \neq j} a_i a_j}{2\pi} + \frac{\mathbf{a}^\top \mathbf{a}^* \sin\phi}{2\pi} \frac{\left\|\mathbf{w}^*\right\|_2}{\left\|\mathbf{w}\right\|_2} + \frac{\sum_{i \neq j} a_j a_j^*}{2\pi} \frac{\left\|\mathbf{w}_*\right\|_2}{\left\|\mathbf{w}\right\|_2}\right) \mathbf{w}.
\end{aligned}
$$

As a sanity check, this formula matches Equation (16) of [Brutzkus and Globerson, 2017] when $\mathbf{a} = \mathbf{a}^* = \mathbf{1}$.

Next, we calculate the expected gradient of $\mathbf{a}$. Recall the gradient formula of $\mathbf{a}$

$$
\begin{aligned}
\frac{\partial \ell(\mathbf{Z}, \mathbf{w}, \mathbf{a})}{\mathbf{a}} =& \left(\mathbf{a}^\top \sigma\left(\mathbf{Z}\mathbf{w}\right) - \left(\mathbf{a}^*\right)^\top \sigma\left(\mathbf{Z}\mathbf{w}^*\right)\right) \sigma\left(\mathbf{Z}\mathbf{w}\right) \\
=&\sigma\left(\mathbf{Z}\mathbf{w}\right) \sigma\left(\mathbf{Z}\mathbf{w}\right)^\top \mathbf{a} - \sigma\left(\mathbf{Z}\mathbf{w}\right) \sigma\left(\mathbf{Z}\mathbf{w}^*\right)^\top \mathbf{a}^*
\end{aligned}
$$

Taking expectation we have

$$\frac{\partial \ell(\mathbf{w}, \mathbf{a})}{\partial \mathbf{a}} = \mathbf{A}(\mathbf{w})\,\mathbf{a} - \mathbf{B}(\mathbf{w}, \mathbf{w}^*)\,\mathbf{a}^*$$

where $\mathbf{A}(\mathbf{w})$ and $\mathbf{B}(\mathbf{w}, \mathbf{w}^*)$ are defined in Equation (5) and (6). Plugging in the formulas for $\mathbf{A}(\mathbf{w})$ and $\mathbf{B}(\mathbf{w}, \mathbf{w}^*)$ derived in the proof of Theorem 2.1 we obtained the desired result.

□

Lemma A.1 (Useful Identities)**.** *Given* $\mathbf{w}$*,* $\mathbf{w}^*$ *with angle* ϕ *and* $\mathbf{Z}$ *is a Gaussian random vector, then*

$$\begin{aligned}
\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right]\mathbf{w} =& \frac{1}{2}\frac{\mathbf{w}}{\|\mathbf{w}\|_2} \\
\mathbb{E}\left[\mathbf{z}\mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right] =& \frac{1}{\sqrt{2\pi}}\frac{\mathbf{w}}{\|\mathbf{w}\|_2} \\
\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0, \mathbf{z}^\top \mathbf{w}_* \geq 0\right\}\right]\mathbf{w}_* =& \frac{1}{\pi}(\pi - \phi)\,\mathbf{w}^* + \frac{1}{\pi}\sin\phi \frac{\|\mathbf{w}^*\|_2}{\|\mathbf{w}\|_2}\mathbf{w} \\
\mathbb{E}\left[\sigma\left(\mathbf{z}^\top \mathbf{w}\right)\sigma\left(\mathbf{z}^\top \mathbf{w}_*\right)\right] =& \frac{1}{2\pi}\left(\cos\phi\left(\pi - \phi + \sin\phi\cos\phi + \sin^2\phi\right)\right)\|\mathbf{w}\|_2\|\mathbf{w}^*\|_2
\end{aligned}$$

Proof. Consider an orthonormal basis of $\mathbb{R}^{d\times d}$: $\left\{\mathbf{e}_i\mathbf{e}_j^\top\right\}$ with $\mathbf{e}_1 \parallel \mathbf{w}$. Then for $i \neq j$, we know

$$\langle \mathbf{e}_i\mathbf{e}_j, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right]\rangle = 0$$

by the independence properties of Gaussian random vector. For $i = j = 1$,

$$\langle \mathbf{e}_i\mathbf{e}_j^\top, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right]\rangle = \mathbb{E}\left[\left(\mathbf{z}^\top \mathbf{w}\right)^2 \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right] = \frac{1}{2}$$

where the last step is by the property of half-Gaussian. For $i = j \neq j$, $\langle \mathbf{e}_i\mathbf{e}_j^\top, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right]\rangle =$ 1 by standard Gaussian second moment formula. Therefore, $\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right]\mathbf{w} = \frac{1}{2}\mathbf{w}$. $\mathbb{E}\left[\mathbf{z}\mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0\right\}\right] = \frac{1}{\sqrt{2\pi}}\mathbf{w}$ can be proved by mean formula of half-normal distribution. To prove the third identity, consider an orthonormal basis of $\mathbb{R}^{d\times d}$: $\left\{\mathbf{e}_i\mathbf{e}_j^\top\right\}$ with $\mathbf{e}_1 \parallel \mathbf{w}_*$ and $\mathbf{w}$ lies in the plane spanned by $\mathbf{e}_1$ and $\mathbf{e}_2$. Using the polar representation of 2D Gaussian random variables (r is the radius and θ is the angle with $\mathrm{d}P_r = r\exp(-r^2/2)$ and $\mathrm{d}P_\theta = \frac{1}{2\pi}$):

$$\begin{aligned}
\langle \mathbf{e}_1\mathbf{e}_1^\top, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0, \mathbf{z}^\top \mathbf{w}_* \geq 0\right\}\right]\rangle =& \frac{1}{2\pi}\int_0^\infty r^3 \exp\left(-r^2/2\right)\mathrm{d}r \cdot \int_{-\pi/2+\phi}^{\pi/2} \cos^2\theta \mathrm{d}\theta \\
=& \frac{1}{2\pi}\left(\pi - \phi + \sin\phi\cos\phi\right), \\
\langle \mathbf{e}_1\mathbf{e}_2^\top, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0, \mathbf{z}^\top \mathbf{w}_* \geq 0\right\}\right]\rangle =& \frac{1}{2\pi}\int_0^\infty r^3 \exp\left(-r^2/2\right)\mathrm{d}r \cdot \int_{-\pi/2+\phi}^{\pi/2} \sin\theta\cos\theta \mathrm{d}\theta \\
=& \frac{1}{2\pi}\left(\sin^2\phi\right), \\
\langle \mathbf{e}_2\mathbf{e}_2^\top, \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top \mathbb{I}\left\{\mathbf{z}^\top \mathbf{w} \geq 0, \mathbf{z}^\top \mathbf{w}_* \geq 0\right\}\right]\rangle =& \frac{1}{2\pi}\int_0^\infty r^3 \exp\left(-r^2/2\right)\mathrm{d}r \cdot \int_{-\pi/2+\phi}^{\pi/2} \sin^2\theta \mathrm{d}\theta
\end{aligned}$$

$$=\frac{1}{2\pi}\left(\pi-\phi-\sin\phi\cos\phi\right).$$

Also note that $\mathbf{e}_2=\frac{\bar{\mathbf{w}}-\cos\phi\mathbf{e}_1}{\sin\phi}$. Therefore

$$\begin{aligned}\mathbb{E}\left[\mathbf{z}\mathbf{z}^\top\mathbb{I}\left\{\mathbf{z}^\top\mathbf{w}\geq 0,\mathbf{z}^\top\mathbf{w}_*\geq 0\right\}\right]\mathbf{w}_* =&\frac{1}{2\pi}\left(\pi-\phi+\sin\phi\cos\phi\right)\mathbf{w}^*+\frac{1}{2\pi}\sin^2\phi\cdot\frac{\bar{\mathbf{w}}-\cos\phi\mathbf{e}_1}{\sin\phi}\left\|\mathbf{w}^*\right\|_2\\ =&\frac{1}{2\pi}\left(\pi-\phi\right)\mathbf{w}^*+\frac{1}{2\pi}\sin\phi\frac{\left\|\mathbf{w}^*\right\|_2}{\left\|\mathbf{w}\right\|_2}\mathbf{w}.\end{aligned}$$

For the fourth identity, focusing on the plane spanned by $\mathbf{w}$ and $\mathbf{w}_*$, using the polar decomposition, we have

$$\begin{aligned}\mathbb{E}\left[\sigma\left(\mathbf{z}^\top\mathbf{w}\right)\sigma\left(\mathbf{z}^\top\mathbf{w}_*\right)\right] =& \frac{1}{2\pi}\int_0^\infty r^3\exp\left(-r^2/2\right)\mathrm{d}r\cdot\int_{-\pi/2+\phi}^{\pi/2}\left(\cos\theta\cos\phi+\sin\theta\sin\phi\right)\cos\theta\mathrm{d}\theta\left\|\mathbf{w}\right\|_2\left\|\mathbf{w}^*\right\|_2\\ =& \frac{1}{2\pi}\left(\cos\phi\left(\pi-\phi+\sin\phi\cos\phi\right)+\sin^3\phi\right)\left\|\mathbf{w}\right\|_2\left\|\mathbf{w}^*\right\|_2.\end{aligned}$$

□

B Proofs of Qualitative Convergence Results

Proof of Lemma 4.1. When Algorithm 1 converges, since $\mathbf{a}^\top\mathbf{a}^*\neq 0$ and $\left\|\mathbf{v}\right\|_2<\infty$, using the gradient formula in Theorem 2.2, we know that either $\pi-\phi=0$ or $\left(\mathbf{I}-\frac{\mathbf{v}\mathbf{v}^\top}{\left\|\mathbf{v}\right\|_2^2}\right)\mathbf{w}^*=\mathbf{0}$. For the second case, since $\mathbf{I}-\frac{\mathbf{v}\mathbf{v}^\top}{\left\|\mathbf{v}\right\|_2^2}$ is a projection matrix on the complement space of $\mathbf{v}$, $\left(\mathbf{I}-\frac{\mathbf{v}\mathbf{v}^\top}{\left\|\mathbf{v}\right\|_2^2}\right)\mathbf{w}^*=\mathbf{0}$ is equivalent to $\theta\left(\mathbf{v},\mathbf{w}^*\right)=0$. Once the angle between $\mathbf{v}$ and $\mathbf{w}^*$ is fixed, using the gradient formula for $\mathbf{a}$ we have the desired formulas for saddle points. □

Proof of Lemma 4.2. By the gradient formula of $\mathbf{w}$, if $\mathbf{a}^\top\mathbf{a}^*>0$, the gradient is of the form $c\left(\mathbf{I}-\frac{\mathbf{v}\mathbf{v}^\top}{\left\|\mathbf{v}\right\|_2^2}\right)\mathbf{w}^*$ where $c>0$. Thus because $\mathbf{I}-\frac{\mathbf{v}\mathbf{v}^\top}{\left\|\mathbf{v}\right\|_2^2}$ is the projection matrix onto the complement space of $\mathbf{v}$, the gradient update always makes the angle smaller. □

C Proofs of Quantitative Convergence Results

C.1 Useful Technical Lemmas

We first prove the lemma about the convergence of ϕ^t.

Proof of Lemma 4.5. We consider the dynamics of $\sin^2\phi^t$.

$$\begin{aligned}&\sin^2\phi^{t+1}\\ =&1-\frac{\left(\left(\mathbf{v}^{t+1}\right)^\top\mathbf{w}^*\right)^2}{\left\|\mathbf{v}^{t+1}\right\|_2^2\left\|\mathbf{w}^*\right\|_2^2}\\ =&1-\frac{\left(\left(\mathbf{v}^{\mathbf{t}}-\eta\frac{\partial\ell}{\partial\mathbf{v}^{\mathbf{t}}}\right)^\top\mathbf{w}^*\right)^2}{\left(\left\|\mathbf{v}^t\right\|_2^2+\eta^2\left(\frac{\partial\ell}{\partial\mathbf{v}^{\mathbf{t}}}\right)^2\right)\left\|\mathbf{w}^*\right\|_2^2}\end{aligned}$$

$$
\begin{aligned}
=&1 - \frac{\left(\left(\mathbf{v}^t\right)^\top \mathbf{v} + \eta \frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\left(\pi - \phi^t\right)}{2\pi \|\mathbf{v}\|_2} \cdot \sin^2 \phi^t \|\mathbf{w}\|_2^2\right)^2}{\|\mathbf{v}^t\|_2^2 \|\mathbf{w}^*\|_2^2 + \eta^2 \left(\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\left(\pi - \phi^t\right)}{2\pi}\right)^2 \frac{\sin^2 \phi^t \|\mathbf{w}^*\|_2^4}{\|\mathbf{v}^t\|_2^2}} \\
\le&1 - \frac{\left\|\mathbf{v}^t\right\|_2^2 \|\mathbf{w}^*\|_2^2 \cos^2 \phi^t + 2\eta \|\mathbf{w}^*\|_2^3 \cdot \frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}(\pi - \phi)}{2\pi} \cdot \sin^2 \phi^t \cos \phi^t}{\|\mathbf{v}^t\|_2^2 \|\mathbf{w}^*\|_2^2 + \eta^2 \left(\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\left(\pi - \phi^t\right)}{2\pi}\right)^2 \frac{\sin^2 \phi^t \|\mathbf{w}^*\|_2^4}{\|\mathbf{v}^t\|_2^2}} \\
=&\frac{\sin^2 \phi^t - 2\eta \frac{\|\mathbf{w}^*\|_2}{\|\mathbf{v}^t\|_2^2} \cdot \frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}(\pi - \phi)}{2\pi} \cdot \sin^2 \phi^t \cos \phi^t + \eta^2 \left(\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*(\pi - \phi)}{2\pi}\right)^2 \sin^2 \phi^t \left(\frac{\|\mathbf{w}^*\|_2}{\|\mathbf{v}\|_2^2}\right)^2}{1 + \eta^2 \left(\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*(\pi - \phi)}{2\pi}\right)^2 \sin^2 \phi^t \left(\frac{\|\mathbf{w}^*\|_2}{\|\mathbf{v}^t\|_2^2}\right)^2} \\
\le& \sin^2 \phi^t - 2\eta \frac{\|\mathbf{w}^*\|_2}{\|\mathbf{v}^t\|_2^2} \cdot \frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}\left(\pi - \phi\right)}{2\pi} \cdot \sin^2 \phi^t \cos \phi^t + \eta^2 \left(\frac{\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\left(\pi - \phi\right)}{2\pi}\right)^2 \sin^2 \phi^t \left(\frac{\|\mathbf{w}^*\|_2}{\|\mathbf{v}^t\|_2^2}\right)^2
\end{aligned}
$$

where in the first inequality we dropped term proportional to $O(\eta^4)$ because it is negative, in the last equality, we divided numerator and denominator by $\left\|\mathbf{v}^t\right\|_2^2 \|\mathbf{w}^*\|_2^2$ and the last inequality we dropped the denominator because it is bigger than 1. Therefore, recall $\lambda^t = \frac{\|\mathbf{w}^*\|_2 \left(\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\right)\left(\pi - \phi^t\right)}{2\pi \|\mathbf{v}^t\|_2^2}$ and we have

$$
\sin^2 \phi^{t+1} \le \left(1 - 2\eta \cos \phi^t \lambda^t + \eta^2 \left(\lambda^t\right)^2\right) \sin^2 \phi^t. \tag{9}
$$

To this end, we need to make sure $\eta \le \frac{\cos \phi^t}{\lambda^t}$. Note that since $\left\|\mathbf{v}^t\right\|_2^2$ is monotonically increasing, it is lower bounded by 1. Next notice $\phi^t \le \pi/2$. Finally, from Lemma C.2, we know $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \le \left(\|\mathbf{a}^*\|_2^2 + \left(\mathbf{1}^\top \mathbf{a}^*\right)^2\right) \|\mathbf{w}\|_2^2$. Combining these, we have an upper bound

$$
\lambda^t \le \frac{\left(\|\mathbf{a}^*\|_2^2 + \left(\mathbf{1}^\top \mathbf{a}^*\right)^2\right) \|\mathbf{w}^*\|_2^2}{4}.
$$

Plugging this back to Equation (9) and use our assumption on η, we have

$$
\sin^2 \phi^{t+1} \le \left(1 - \eta \cos \phi^t \lambda^t\right) \sin^2 \phi^t.
$$

□

Lemma C.1. $\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* \ge \min\left\{\left(\mathbf{a}^t\right)^\top \mathbf{a}^* + \eta\left(\frac{g(\phi^t) - 1}{\pi - 1} \|\mathbf{a}^*\|_2^2 - \left(\mathbf{a}^t\right)^\top \mathbf{a}^*\right), \frac{g(\phi^t) - 1}{\pi - 1} \|\mathbf{a}^*\|_2^2\right\}$

Proof. Recall the dynamics of $\left(\mathbf{a}^t\right)^\top \mathbf{a}^*$.

$$
\begin{aligned}
\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* &= \left(1 - \frac{\eta\left(\pi - 1\right)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^* + \frac{\eta\left(g(\phi^t) - 1\right)}{2\pi} \|\mathbf{a}^*\|_2^2 + \frac{\eta}{2\pi}\left(\left(\mathbf{1}^\top \mathbf{a}^*\right)^2 - \left(\mathbf{1}^\top \mathbf{a}^*\right)\left(\mathbf{1}^\top \mathbf{a}^t\right)\right) \\
&\ge \left(1 - \frac{\eta\left(\pi - 1\right)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^* + \frac{\eta\left(g(\phi^t) - 1\right)}{2\pi} \|\mathbf{a}^*\|_2^2
\end{aligned}
$$

where the inequality is due to Lemma 4.4. If $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \geq \frac{g(\phi^t)-1}{\pi-1}\left\|\mathbf{a}^*\right\|_2^2$,

$$\begin{aligned}\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* \geq & \left(1-\frac{\eta\left(\pi-1\right)}{2\pi}\right)\frac{g(\pi^t)-1}{\pi-1}\left\|\mathbf{a}^*\right\|_2^2+\frac{\eta\left(g(\phi^t)\right)}{\pi-1}\left\|\mathbf{a}^*\right\|_2^2 \\ = & \frac{g(\phi^t)-1}{\pi-1}\left\|\mathbf{a}^*\right\|_2^2.\end{aligned}$$

If $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \leq \frac{g(\phi^t)-1}{\pi-1}\left\|\mathbf{a}^*\right\|_2^2$, simple algebra shows $\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^*$ increases by at least

$$\eta\left(\frac{g(\phi^t)-1}{\pi-1}\left\|\mathbf{a}^*\right\|_2^2-\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\right).$$

□

A simple corollary is $\mathbf{a}^\top \mathbf{a}^*$ is uniformly lower bounded.

Corollary C.1. *For all* $t = 1, 2, \ldots$, $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \geq \min\left\{\left(\mathbf{a}^0\right)^\top \mathbf{a}^*, \frac{g(\phi^0)-1}{\pi-1}\left\|\mathbf{a}^*\right\|_2^2\right\}$.

This lemma also gives an upper bound of number of iterations to make $\mathbf{a}^\top \mathbf{a}^* = \Theta\left(\left\|\mathbf{a}^*\right\|_2^2\right)$.

Corollary C.2. *If* $g(\phi)-1=\Omega\left(1\right)$, *then after* $\frac{1}{\eta}$ *iterations,* $\mathbf{a}^\top \mathbf{a}^* = \Theta\left(\left\|\mathbf{a}^*\right\|_2^2\right)$.

Proof. Note if $g(\phi)-1=\Omega\left(1\right)$ and $\mathbf{a}^\top \mathbf{a}^* \leq \frac{1}{2}\cdot\frac{g(\phi)}{\pi-1}\left\|\mathbf{a}^*\right\|_2^2$, each iteration $\mathbf{a}^\top \mathbf{a}^*$ increases by $\eta\frac{g(\phi)}{\pi-1}\left\|\mathbf{a}^*\right\|_2^2$. □

We also need an upper bound of $\left(\mathbf{a}^t\right)^\top \mathbf{a}^*$.

Lemma C.2. *For* $t = 0, 1, \ldots$, $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \leq \left(\left\|\mathbf{a}^*\right\|_2^2+\left(\mathbf{1}^\top \mathbf{a}^*\right)^2\right)\left\|\mathbf{w}^*\right\|_2^2$.

Proof. Without loss of generality, assume $\left\|\mathbf{w}^*\right\|_2 = 1$. Again, recall the dynamics of $\left(\mathbf{a}^t\right)^\top \mathbf{a}^*$.

$$\begin{aligned}\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* = & \left(1-\frac{\eta\left(\pi-1\right)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^*+\frac{\eta\left(g(\phi^t)-1\right)}{2\pi}\left\|\mathbf{a}^*\right\|_2^2+\frac{\eta}{2\pi}\left(\left(\mathbf{1}^\top \mathbf{a}^*\right)^2-\left(\mathbf{1}^\top \mathbf{a}^*\right)\left(\mathbf{1}^\top \mathbf{a}^t\right)\right) \\ \leq & \left(1-\frac{\eta\left(\pi-1\right)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^*+\frac{\eta\left(\pi-1\right)}{2\pi}\left\|\mathbf{a}^*\right\|_2^2+\frac{\eta\left(\pi-1\right)}{2\pi}\left(\mathbf{1}^\top \mathbf{a}^*\right)^2.\end{aligned}$$

Now we prove by induction, suppose the conclusion holds at iteration t, $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* \leq \left\|\mathbf{a}^*\right\|_2^2+\left(\mathbf{1}^\top \mathbf{a}^*\right)^2$. Plugging in we have the desired result. □

C.2 Convergence of Phase I

In this section we prove the convergence of Phase I.

Proof of Convergence of Phase I. Lemma C.3 implies after $O\left(\frac{1}{\cos\phi^0\beta^0}\right)$ iterations, $\cos\phi^t = \Omega\left(1\right)$, which implies $\frac{g(\phi^t)-1}{\pi-1} = \Omega\left(1\right)$. Using Corollary C.2, we know after $O\left(\frac{1}{\eta}\right)$ iterations we have $\left(\mathbf{a}^t\right)^\top \mathbf{a}^*\left\|\mathbf{w}^*\right\| = \Omega\left(\left\|\mathbf{w}^*\right\|_2^2\left\|\mathbf{a}^*\right\|_2^2\right)$. □

The main ingredient of the proof of phase I is the follow lemma where we use a joint induction argument to show the convergence of ϕ^t and a uniform upper bound of $\left\|\mathbf{v}^t\right\|_2$.

Lemma C.3. *Let* $\beta^0 = \min\left\{\left(\mathbf{a}^0\right)^\top \mathbf{a}^*, \left(g(\phi^0)-1\right)\|\mathbf{a}^*\|_2^2\right\}\|\mathbf{w}^*\|_2^2$. *If the step size satisfies* $\eta \leq \min\left\{\frac{\beta^*\cos\phi^0}{8\left(\|\mathbf{a}^*\|_2^2+\left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2}, \frac{\cos\phi^0}{\left(\|\mathbf{a}^*\|_2^2+\left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2}, \frac{2\pi}{k+\pi-1}\right\}$, *we have for* $t = 0, 1, \ldots$

$$\sin^2\phi^t \leq \left(1-\eta\cdot\frac{\cos\phi^0\beta^0}{8}\right)^t \text{ and } \left\|\mathbf{v}^t\right\|_2 \leq 2.$$

Proof. We prove by induction. The initialization ensure when $t=0$, the conclusion is correct. Now we consider the dynamics of $\left\|\mathbf{v}^t\right\|_2^2$. Note because the gradient of $\mathbf{v}$ is orthogonal to $\mathbf{v}$ [Salimans and Kingma, 2016], we have a simple dynamic of $\left\|\mathbf{v}^t\right\|_2^2$.

$$\begin{aligned}
\left\|\mathbf{v}^t\right\|_2^2 &= \left\|\mathbf{v}^{t-1}\right\|_2^2 + \eta^2\left\|\frac{\partial\ell\left(\mathbf{v},\mathbf{a}\right)}{\partial\mathbf{v}}\right\|_2^2 \\
&= \left\|\mathbf{v}^{t-1}\right\|_2^2 + \eta^2\left(\frac{\left(\mathbf{a}^t\right)^\top\mathbf{a}^*\left(\pi-\phi^{t-1}\right)}{2\pi}\right)^2\frac{\sin^2\phi^t\|\mathbf{w}^*\|_2^2}{\left\|\mathbf{v}^t\right\|_2^2} \\
&\leq \left\|\mathbf{v}^{t-1}\right\|_2^2 + \eta^2\left(\|\mathbf{a}^*\|_2^2+\left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2\sin^2\phi^{t-1} \\
&= 1 + \eta^2\left(\|\mathbf{a}^*\|_2^2+\left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2\sum_{i=1}^{t-1}\sin^2\phi^i \\
&\leq 1 + \eta^2\left(\|\mathbf{a}^*\|_2^2+\left(\mathbf{1}^\top\mathbf{a}^*\right)^2\right)\|\mathbf{w}^*\|_2^2\frac{8}{\eta\cos\phi^0\beta^0} \\
&\leq 2
\end{aligned}$$

where the first inequality is by Lemma C.2 and the second inequality we use our induction hypothesis. Recall $\lambda^t = \frac{\|\mathbf{w}^*\|_2\left(\left(\mathbf{a}^t\right)^\top\mathbf{a}^*\right)\left(\pi-\phi^t\right)}{2\pi\|\mathbf{v}^t\|_2^2}$. The uniform upper bound of $\|\mathbf{v}\|_2$ and the fact that $\phi^t \leq \pi/2$ imply a lower bound $\lambda^t \geq \frac{\beta^0}{8}$. Plugging in Lemma 4.5, we have

$$\sin^2\phi^{t+1} \leq \left(1-\eta\frac{\cos\phi^0\beta^0}{8}\right)\sin^2\phi^t \leq \left(1-\eta\frac{\cos\phi^0\beta^0}{8}\right)^{t+1}.$$

We finish our joint induction proof. □

C.3 Analysis of Phase II

In this section we prove the convergence of phase II and necessary auxiliary lemmas.

Proof of Convergence of Phase II. At the beginning of Phase II, $\left(\mathbf{a}^{T_1}\right)^\top\mathbf{a}^*\|\mathbf{w}^*\| = \Omega\left(\|\mathbf{w}^*\|_2^2\|\mathbf{a}^*\|_2^2\right)$ and $g(\phi^{T_1})-1 = \Omega(1)$. Therefore, Lemma C.1 implies for all $t = T_1, T_1+1, \ldots$, $\left(\mathbf{a}^t\right)^\top\mathbf{a}^*\|\mathbf{w}^*\| = \Omega\left(\|\mathbf{w}^*\|_2^2\|\mathbf{a}^*\|_2^2\right)$. Combining with the fact that $\|\mathbf{v}\|_2 \leq 2$ (c.f. Lemma C.3), we obtain a lower bound $\lambda_t \geq \Omega\left(\|\mathbf{w}^*\|_2^2 \|\mathbf{a}^*\|_2^2\right)$ We also know that $\cos\phi^{T_1} = \Omega(1)$ and $\cos\phi^t$ is monotinically increasing (c.f. Lemma 4.2), so for all $t = T_1, T_1+1, \ldots$, $\cos\phi^t = \Omega(1)$. Plugging in these two lower bounds into Theorem 4.5, we have

$$\sin^2\phi^{t+1} \leq \left(1 - \eta C \|\mathbf{w}^*\|_2^2 \|\mathbf{a}^*\|_2^2\right)\sin^2\phi^t.$$

for some absolute constant C. Thus, after $O\left(\frac{1}{\eta\|\mathbf{w}^*\|_2^2\|\mathbf{a}^*\|_2^2}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations, we have $\sin^2\phi^t \leq \min\left\{\epsilon^{10}, \left(\epsilon\frac{\|\mathbf{a}^*\|_2}{|\mathbf{1}^\top\mathbf{a}^*|}\right)^{10}\right\}$, which implies $\pi - g(\phi^t) \leq \min\left\{\epsilon, \epsilon\frac{\|\mathbf{a}^*\|_2}{|\mathbf{1}^\top\mathbf{a}^*|}\right\}$. Now using Lemma C.4,Lemma C.5 and Lemma C.6, we have after $\widetilde{O}\left(\frac{1}{\eta k}\log\left(\frac{1}{\epsilon}\right)\right)$ iterations $\ell(\mathbf{v}, \mathbf{a}) \leq C_1\epsilon\|\mathbf{a}^*\|_2^2\|\mathbf{w}^*\|_2^2$ for some absolute constant C_1. Rescaling ϵ properly we obtain the desired result. □

C.3.1 Technical Lemmas for Analyzing Phase II

In this section we provide some technical lemmas for analyzing Phase II. Because of the positive homogeneity property, without loss of generality, we assume $\|\mathbf{w}^*\|_2 = 1$.

Lemma C.4. *If* $\pi - g(\phi^0) \leq \epsilon\frac{\|\mathbf{a}^*\|_2}{|\mathbf{1}^\top\mathbf{a}^*|}$, *after* $T = O\left(\frac{1}{\eta k}\log\left(\frac{|\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^0|}{\epsilon\|\mathbf{a}^*\|_2}\right)\right)$ *iterations,* $\left|\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^T\right| \leq 2\epsilon\|\mathbf{a}^*\|_2$.

Proof. Recall the dynamics of $\mathbf{1}^\top\mathbf{a}^t$.

$$\begin{aligned}
\mathbf{1}^\top\mathbf{a}^{t+1} &= \left(1 - \frac{\eta(k+\pi-1)}{2\pi}\right)\mathbf{1}^\top\mathbf{a}^t + \frac{\eta\left(k + g(\phi^t) - 1\right)}{2\pi}\mathbf{1}^\top\mathbf{a}^* \\
&= \left(1 - \frac{\eta(k+\pi-1)}{2\pi}\right)\mathbf{1}^\top\mathbf{a}^t + \frac{\eta\left(k + g(\phi^t) - 1\right)}{2\pi}\mathbf{1}^\top\mathbf{a}^*.
\end{aligned}$$

Assume $\mathbf{1}^\top\mathbf{a}^* > 0$ (the other case is similar). By Lemma 4.4 we know $\mathbf{1}^\top\mathbf{a}^t < \mathbf{1}^\top\mathbf{a}^*$ for all t. Consider

$$\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^{t+1} = \left(1 - \frac{\eta(k+\pi-1)}{2\pi}\right)\left(\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^*\right) + \frac{\eta\left(\pi - g(\phi^t)\right)}{2\pi}\mathbf{1}^\top\mathbf{a}^*.$$

Therefore we have

$$\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^{t+1} - \frac{\left(\pi - g\left(\phi^t\right)\right)\mathbf{1}^\top\mathbf{a}^*}{k+\pi-1} = \left(1 - \frac{\eta(k+\pi-1)}{2\pi}\right)\left(\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^* - \frac{\left(\pi - g\left(\phi^t\right)\right)\mathbf{1}^\top\mathbf{a}^*}{k+\pi-1}\right).$$

After $T = O\left(\frac{1}{\eta k}\log\left(\frac{|\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^0|}{\epsilon\|\mathbf{a}^*\|_2}\right)\right)$ iterations, we have $\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^t - \frac{(\pi - g(\phi^t))\mathbf{1}^\top\mathbf{a}^*}{k+\pi-1} \leq \epsilon\|\mathbf{a}^*\|_2$, which implies$\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^t \leq 2\epsilon\|\mathbf{a}^*\|_2$. □

Lemma C.5. *If* $\pi - g(\phi^0) \leq \epsilon\frac{\|\mathbf{a}^*\|_2}{|\mathbf{1}^\top\mathbf{a}^*|}$ *and* $\left|\mathbf{1}^\top\mathbf{a}^* - \mathbf{1}^\top\mathbf{a}^0\right| \leq \frac{\epsilon}{k}\|\mathbf{a}^*\|_2$, *then after* $T = O\left(\frac{1}{\eta}\log\left(\frac{\|\mathbf{a}^* - \mathbf{a}^0\|_2}{\epsilon\|\mathbf{a}^*\|_2}\right)\right)$ *iterations,* $\left\|\mathbf{a}^* - \mathbf{a}^0\right\|_2 \leq C\epsilon\|\mathbf{a}^*\|_2$ *for some absolute constant* C.

Proof. We first consider the inner product

$$\begin{aligned}
&\langle \frac{\partial \ell\left(\mathbf{v}^t, \mathbf{a}^t\right)}{\mathbf{a}^t}, \mathbf{a}^t - \mathbf{a}^* \rangle \\
=& \frac{\pi - 1}{2\pi} \left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2^2 - \frac{g(\phi^t) - \pi}{2\pi} \left(\mathbf{a}^*\right)^\top \left(\mathbf{a}^t - \mathbf{a}^*\right) + \left(\mathbf{a}^t - \mathbf{a}^*\right) \mathbf{1}\mathbf{1}^\top \left(\mathbf{a}^\top - \mathbf{a}^*\right) \\
\geq& \frac{\pi - 1}{2\pi} \left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2^2 - \frac{g(\phi^t) - \pi}{2\pi} \left\|\mathbf{a}^*\right\|_2 \left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2 .
\end{aligned}$$

Next we consider the squared norm of gradient

$$\begin{aligned}
\left\|\frac{\partial \ell\left(\mathbf{v}, \mathbf{a}\right)}{\partial \mathbf{a}}\right\|_2^2 =& \frac{1}{4\pi^2} \left\|\left(\pi - 1\right)\left(\mathbf{a}^t - \mathbf{a}^*\right) + \left(\pi - g(\phi^t)\right)\mathbf{a}^* + \mathbf{1}\mathbf{1}^\top \left(\mathbf{a}^t - \mathbf{a}^*\right)\right\|_2^2 \\
\leq& \frac{3}{4\pi^2}\left(\left(\pi - 1\right)^2 \left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2^2 + \left(\pi - g(\phi^t)\right)^2 \left\|\mathbf{a}^*\right\|_2^2 + k^2 \left(\mathbf{1}^\top \mathbf{a}^t - \mathbf{1}^\top \mathbf{a}^*\right)^2\right).
\end{aligned}$$

Suppose $\left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2 \leq \epsilon \left\|\mathbf{a}^*\right\|_2$, then

$$\begin{aligned}
\langle \frac{\partial \ell\left(\mathbf{v}^t, \mathbf{a}^t\right)}{\mathbf{a}^t}, \mathbf{a}^t - \mathbf{a}^* \rangle \geq& \frac{\pi - 1}{2\pi} \left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2^2 - \frac{\epsilon^2}{2\pi} \left\|\mathbf{a}^*\right\|_2^2 \\
\left\|\frac{\partial \ell\left(\mathbf{v}, \mathbf{a}\right)}{\partial \mathbf{a}}\right\|_2^2 \leq& 3\epsilon^2 \left\|\mathbf{a}^*\right\|_2^2 .
\end{aligned}$$

Therefore we have

$$\begin{aligned}
&\left\|\mathbf{a}^{t+1} - \mathbf{a}^*\right\|_2^2 \leq \left(1 - \frac{\eta\left(\pi - 1\right)}{2\pi}\right)\left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2^2 + 4\eta\epsilon^2 \left\|\mathbf{a}\right\|^2 \\
\Rightarrow &\left\|\mathbf{a}^{t+1} - \mathbf{a}^*\right\|_2^2 - \frac{8\left(\pi - 1\right)\epsilon^2 \left\|\mathbf{a}^*\right\|_2^2}{\pi - 1} \leq \left(1 - \frac{\eta\left(\pi - 1\right)}{2\pi}\right)\left(\left\|\mathbf{a}^t - \mathbf{a}^*\right\|_2^2 - \frac{8\left(\pi - 1\right)\epsilon^2 \left\|\mathbf{a}^*\right\|_2^2}{\pi - 1}\right).
\end{aligned}$$

Thus after $O\left(\frac{1}{\eta}\left(\frac{1}{\epsilon}\right)\right)$ iterations, we must have $\left\|\mathbf{a}^{t+1} - \mathbf{a}^*\right\|_2^2 \leq C\epsilon \left\|\mathbf{a}^*\right\|_2$ for some large absolute constant C. Rescaling ϵ, we obtain the desired result. ☐

Lemma C.6. *If* $\pi - g(\phi) \leq \epsilon$ *and* $\left\|\mathbf{a} - \mathbf{a}^* \left\|\mathbf{w}^*\right\|_2\right\| \leq \epsilon \left\|\mathbf{a}^*\right\|_2 \left\|\mathbf{w}^*\right\|_2$, *then the population loss satisfies* $\ell\left(\mathbf{v}, \mathbf{a}\right) \leq C\epsilon \left\|\mathbf{a}^*\right\|_2^2 \left\|\mathbf{w}^*\right\|_2^2$ *for some constant* $C > 0$.

Proof. The result follows by plugging in the assumptions in Theorem 2.1. ☐

D Proofs of Initialization Scheme

Proof of Theorem 3.2. The proof of the first part of Theorem 3.2 just uses the symmetry of unit sphere and ball and the second part is a direct application of Lemma 2.5 of [Hardt and Price, 2014]. Lastly, since $\mathbf{a}^0 \sim \mathcal{B}\left(\mathbf{0}\frac{\left|\mathbf{1}^\top \mathbf{a}^*\right|}{\sqrt{k}}\right)$, we have $\mathbf{1}^\top \mathbf{a}^0 \leq \left\|\mathbf{a}^0\right\|_1 \leq \sqrt{k}\left\|\mathbf{a}^0\right\|_2 \leq \left|\mathbf{1}^\top \mathbf{a}^*\right| \left\|\mathbf{firstlayer}^*\right\|_2$ where the second inequality is due to Hölder's inequality. ☐

E Proofs of Converging to Spurious Local Minimum

Proof of Theorem 3.3. The main idea is similar to Theorem 3.1 but here we show $\mathbf{w} \to -\mathbf{w}^*$ (without loss of generality, we assume $\|\mathbf{w}^*\|_2 = 1$). Different from Theorem 3.1, here we need to prove the invariance $\mathbf{a}^\top \mathbf{a}^* < 0$, which implies our desired result. We prove by induction, suppose $\left(\mathbf{a}^t\right)^\top \mathbf{a}^* > 0$, $\left|\mathbf{1}^\top \mathbf{a}^t\right| \leq \left|\mathbf{1}^\top \mathbf{a}^*\right|$, $g\left(\phi^0\right) \leq \frac{-2\left(\mathbf{1}^\top \mathbf{a}\right)^2}{\|\mathbf{a}^*\|_2^2} + 1$ and $\eta < \frac{k+\pi-1}{2\pi}$. Note $\left|\mathbf{1}^\top \mathbf{a}^t\right| \leq \left|\mathbf{1}^\top \mathbf{a}^*\right|$ are satisfied by Lemma 4.4 and $g\left(\phi^0\right) \leq \frac{-2\left(\mathbf{1}^\top \mathbf{a}\right)^2}{\|\mathbf{a}^*\|_2^2} + 1$ by our initialization condition and induction hypothesis that implies ϕ^t is increasing. Recall the dynamics of $\left(\mathbf{a}^t\right)^\top \mathbf{a}^*$.

$$\begin{aligned}\left(\mathbf{a}^{t+1}\right)^\top \mathbf{a}^* &= \left(1 - \frac{\eta\left(\pi - 1\right)}{2\pi}\right)\left(\mathbf{a}^t\right)^\top \mathbf{a}^* + \frac{\eta\left(g\left(\phi^t\right) - 1\right)}{2\pi}\|\mathbf{a}^*\|_2^2 + \frac{\eta}{2\pi}\left(\left(\mathbf{1}^\top \mathbf{a}^*\right)^2 - \left(\mathbf{1}^\top \mathbf{a}^t\right)\left(\mathbf{1}^\top \mathbf{a}^*\right)\right)\\ &\leq \frac{\eta\left(\left(g(\phi^t) - 1\right)\|\mathbf{a}^*\|_2 + 2\left(\mathbf{1}^\top \mathbf{a}^*\right)^2\right)}{2\pi} < 0\end{aligned}$$

where the first inequality we used our induction hypothesis on inner product between $\mathbf{a}^t$ and $\mathbf{a}^*$ and $\left|\mathbf{1}^\top \mathbf{a}^t\right| \leq \left|\mathbf{1}^\top \mathbf{a}^*\right|$ and the second inequality is by induction hypothesis on ϕ^t. Thus when gradient descent algorithm converges, according Lemma 4.1, $\theta\left(\mathbf{v}, \mathbf{w}^*\right) = \pi, \mathbf{a} = \left(\mathbf{1}\mathbf{1}^\top + \left(\pi - 1\right)\mathbf{I}\right)^{-1}\left(\mathbf{1}\mathbf{1}^\top - \mathbf{I}\right)\|\mathbf{w}^*\|_2 \mathbf{a}^*$. Plugging these into Theorem 2.1, with some routine algebra, we show $\ell\left(\mathbf{v}, \mathbf{a}\right) = \Omega\left(\|\mathbf{w}^*\|_2^2 \|\mathbf{a}^*\|_2^2\right)$. $\square$