Towards Understanding the Importance of Shortcut Connections in Residual Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Residual Network (ResNet) is undoubtedly a milestone in deep learning. ResNet is equipped with shortcut connections between layers, and exhibits efficient training using simple first order algorithms. Despite of the great empirical success, the reason behind is far from being well understood. In this paper, we study a two-layer non-overlapping convolutional ResNet. Training such a network requires solving a non-convex optimization problem with a spurious local optimum. We show, however, that gradient descent combined with proper normalization, avoids being trapped by the spurious local optimum, and converges to a global optimum in polynomial time, when the weight of the first layer is initialized at 0, and that of the second layer is initialized arbitrarily in a ball. Numerical experiments are provided to support our theory.

1 Introduction

Neural Networks have revolutionized a variety of real world applications in the past few years, such as computer vision (Krizhevsky et al., 2012; Goodfellow et al., 2014; Long et al., 2015), natural language processing (Graves et al., 2013; Bahdanau et al., 2014; Young et al., 2018), etc. Among different types of networks, Residual Network (ResNet, He et al. (2016a)) is undoubted a milestone. ResNet is equipped with shortcut connections, which skip layers in the forward step of an input. Similar idea also appears in the Highway Networks (Srivastava et al., 2015), and further inspires densely connected convolutional networks (Huang et al., 2017).

ResNet owes its great success to a surprisingly efficient training compared to the widely used feedforward Convolutional Neural Networks (CNN, Krizhevsky et al. (2012)). Feedforward CNNs are seldomly used with more than 30 layers in the existing literature. There are experimental results suggest that very deep feedforward CNNs are significantly slow to train, and yield worse performance than their shallow counterparts (He et al., 2016a). However, simple first order algorithms such as stochastic gradient descent and its variants are able to train ResNet with hundreds of layers, and achieve better performance than the state-of-the-art. For example, ResNet-152 (He et al., 2016a), consisting of 152 layers, achieves a 19.38% top-1 error on ImageNet. He et al. (2016b) also demonstrated a more aggressive ResNet-1001 on the CIFAR-10 data set with 1000 layers. It achieves a 4.92% error — better than shallower ResNets such as ResNet-110.

Despite the great success and popularity of ResNet, the reason why it can be efficiently trained is still largely unknown. One line of research empirically studies ResNet and provides intriguing observations. Veit et al. (2016), for example, suggest that ResNet can be viewed as a collection of weakly dependent smaller networks of varying sizes. More interestingly, they reveal that these smaller networks alleviate the vanishing gradient problem. Balduzzi et al. (2017) further elaborate on the vanishing gradient problem. They show that the gradient in ResNet only decays sublinearly in contrast to the exponential decay in feedforward neural networks. Recently, Li et al. (2018) visualize the landscape of neural networks, and show that the shortcut connection yields a smoother

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

optimization landscape. In spite of these empirical evidences, rigorous theoretical justifications are seriously lacking.

Another line of research theoretically investigates ResNet with simple network architectures. Hardt and Ma (2016) show that linear ResNet has no spurious local optima (local optima that yield larger objective values than the global optima). Later, Li and Yuan (2017) study using Stochastic Gradient Descent (SGD) to train a two-layer ResNet with only one unknown layer. They show that the optimization landscape has no spurious local optima and saddle points. They also characterize the local convergence of SGD around the global optimum. These results, however, are often considered to be overoptimistic, due to the oversimplified assumptions.

To better understand ResNet, we study a two-layer non-overlapping convolutional neural network, whose optimization landscape contains a spurious local optimum. Such a network was first studied in Du et al. (2017). Specifically, we consider

$$g(v, a, Z) = a^\top \sigma\left(Z^\top v\right), \tag{1}$$

where $Z \in \mathbb{R}^{p\times k}$ is an input, $a \in \mathbb{R}^k, v \in \mathbb{R}^p$ are the output weight and the convolutional weight, respectively, and σ is the element-wise ReLU activation. Since the ReLU activation is positive homogeneous, the weights a and v can arbitrarily scale with each other. Thus, we impose the assumption $\|v\|_2 = 1$ to make the neural network identifiable. We further decompose $v = \mathbb{1}/\sqrt{p} + w$ with $\mathbb{1}$ being a vector of 1's in $\mathbb{R}^p$, and rewrite (1) as

$$f(w, a, Z) = a^\top \sigma\left(Z^\top(\mathbb{1}/\sqrt{p} + w)\right), \tag{2}$$

Here $\mathbb{1}/\sqrt{p}$ represents the average pooling shortcut connection, which allows a direct interaction between the input Z and the output weight a.

We investigate the convergence of training ResNet by considering a realizable case. Specifically, the training data is generated from a teacher network with true parameters a^*, v^* with $\|v^*\|_2 = 1$. We aim to recover the teacher neural network using a student network defined in (2) by solving an optimization problem:

$$(\widehat{w}, \widehat{a}) = \operatorname*{argmin}_{w,a} \frac{1}{2}\mathbb{E}_Z \left[f(w, a, Z) - g(v^*, a^*, Z)\right]^2, \tag{3}$$

where Z is independent Gaussian input. Although largely simplified, (3) is nonconvex and possesses a nuisance — There exists a spurious local optimum (see an explicit characterization in Section 2). Early work, Du et al. (2017), show that when the student network has the same architecture as the teacher network, GD with random initialization can be trapped in a spurious local optimum with a constant probability[1]. A natural question here is

Does the shortcut connection eases the training?

This paper suggests a positive answer: When initialized with $w = 0$ and a arbitrarily in a ball, GD with proper normalization converges to a global optimum of (3) in polynomial time, under the assumption that $(v^*)^\top(\mathbb{1}/\sqrt{p})$ is close to 1. Such an assumption requires that there exists a w^* of relatively small magnitude, such that $v^* = \mathbb{1}/\sqrt{p} + w^*$. This assumption is supported by both empirical and theoretical evidences. Specifically, the experiments in Li et al. (2016) and Yu et al. (2018), show that the weight in well-trained deep ResNet has a small magnitude, and the weight for each layer has vanishing norm as the depth tends to infinity. Hardt and Ma (2016) suggest that, when using linear ResNet to approximate linear transformations, the norm of the weight in each layer scales as $O(1/D)$ with D being the depth. Bartlett et al. (2018) further show that deep nonlinear ResNet, with the norm of the weight of order $O(\log D/D)$, is sufficient to express differentiable functions under certain regularity conditions. These results motivate us to assume w^* is relatively small.

Our analysis shows that the convergence of GD exhibits 2 stages. Specifically, our initialization guarantees w is sufficiently away from the spurious local optimum. In the first stage, with proper step sizes, we show that the shortcut connection helps the algorithm avoid being attracted by the spurious local optima. Meanwhile, the shortcut connection guides the algorithm to evolve towards a global optimum. In the second stage, the algorithm enters the basin of attraction of the global optimum. With properly chosen step sizes, w and a jointly converge to the global optimum.

Our analysis thus explains why ResNet benefits training, when the weights are simply initialized at zero (Li et al., 2016), or using the Fixup initialization in Zhang et al. (2019). We remark that our

[1]The probability is bounded between $1/4$ and $3/4$. Numerical experiments show that this probability can be as bad as $1/2$ with the worst configuration of a, v.

choice of step sizes is also related to learning rate warmup (Goyal et al., 2017), and other learning rate schemes for more efficient training of neural networks (Smith, 2017; Smith and Topin, 2018). We refer readers to Section 5 for a more detailed discussion.

Notations: Given a vector $v = (v_1, \ldots, v_m)^\top \in \mathbb{R}^m$, we denote the Euclidean norm $\|v\|_2^2 = v^\top v$. Given two vectors $u, v \in \mathbb{R}^d$, we denote the angle between them as $\angle(u, v) = \arccos \frac{u^\top v}{\|u\|_2 \|v\|_2}$, and the inner product as $\langle u, v \rangle = u^\top v$. We denote $\mathbb{1} \in \mathbb{R}^d$ as the vector of all the entries being 1. We also denote $\mathbb{B}_0(r) \in \mathbb{R}^d$ as the Euclidean ball centered at 0 with radius r.

2 Model and Algorithm

Model. We consider the realizable setting where the label is generated from a noiseless teacher network in the following form

$$g(v^*, a^*, Z) = \sum_{j=1}^k a_j^* \sigma\left(Z_j^\top v^*\right).$$

Here v^*, a^*, Z_j's are the true convolutional weight, true output weight, and input. σ denotes the element-wise ReLU activation. Our student network is defined in (2). For notational convenience, we expand the second layer and rewrite (2) as

$$f(w, a, Z) = \sum_{j=1}^k a_j \sigma\left(Z_j^\top (\mathbb{1}/\sqrt{p} + w)\right), \tag{4}$$

where $w \in \mathbb{R}^p$, $a_j \in \mathbb{R}$, and $Z_j \in \mathbb{R}^p$ for all $j = 1, 2, \ldots, k$. We assume the input data Z_j's are identically independently sampled from $\mathcal{N}(0, I)$. Note that the above network is not identifiable, because of the positive homogeneity of the ReLU function, that is $\mathbb{1}\sqrt{p} + w$ and a can scale with each other by any positive constant without changing the output value. Thus, to achieve identifiability, instead of (4), we propose to train the following student network,

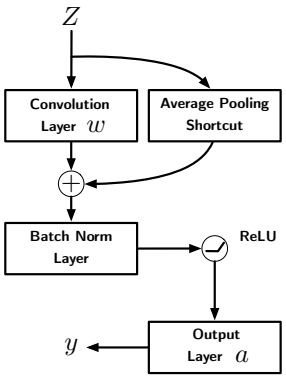

Figure 1: The non-overlapping two layer residual network with normalization layer.

$$f(w, a, Z) = \sum_{j=1}^k a_j \sigma\left(Z_j^\top \frac{\mathbb{1}/\sqrt{p} + w}{\|\mathbb{1}/\sqrt{p} + w\|_2}\right). \tag{5}$$

An illustration of (5) is provided in Figure 1. We then recover (v^*, a^*) of our teacher network by solving a nonconvex optimization problem

$$\min_{w,a} \mathcal{L}(w, a) = \frac{1}{2}\mathbb{E}_Z[g(v^*, a^*, Z) - f(w, a, Z)]^2. \tag{6}$$

Recall that we assume $\|v^*\|_2 = 1$. One can easily verify that (6) has global optima and spurious local optima. The characterization is analogous to Du et al. (2017), although the objective is different.

Proposition 1. *For any constant $\alpha > 0$, (w, a) is a global optimum of (6), if $\mathbb{1}/\sqrt{p} + w = \alpha v^*$ and $a = a^*$; (w, a) is a spurious local optimum of (6), if $\mathbb{1}/\sqrt{p} + w = -\alpha v^*$ and $a = (\mathbb{1}\mathbb{1}^\top + (\pi - 1)I)^{-1}(\mathbb{1}\mathbb{1}^\top - I)a^*$.*

The proof is adapted from Du et al. (2017), and the details are provided in Appendix B.1.

Now we formalize the assumption on v^* in Section 1, which is supported by the theoretical and empirical evidence in Li et al. (2016); Yu et al. (2018); Hardt and Ma (2016); Bartlett et al. (2018).

Assumption 1 (Shortcut Prior)**.** *There exists a w^* with $\|w^*\|_2 \leq 1$, such that $v^* = w^* + \mathbb{1}/\sqrt{p}$.*

Assumption 1 implies $(\mathbb{1}/\sqrt{p})^\top v^* \geq 1/2$. We remark that our analysis actually applies to any w^* satisfying $\|w^*\|_2 \leq c$ for any positive constant $c \in (0, \sqrt{2})$. Here we consider $\|w^*\|_2 \leq 1$ to ease the presentation. Throughout the rest of the paper, we assume this assumption holds true.

GD with Normalization. We solve the optimization problem (6) by gradient descent. Specifically, at the $(t+1)$-th iteration, we compute

$$\begin{aligned} \widetilde{w}_{t+1} &= w_t - \eta_w \nabla_w \mathcal{L}(w_t, a_t), \\ w_{t+1} &= \frac{\mathbb{1}/\sqrt{p} + \widetilde{w}_{t+1}}{\|\mathbb{1}/\sqrt{p} + \widetilde{w}_{t+1}\|_2} - \frac{\mathbb{1}}{\sqrt{p}}, \\ a_{t+1} &= a_t - \eta_a \nabla_a \mathcal{L}(w_t, a_t). \end{aligned} \tag{7}$$

Note that we normalize $\mathbb{1}/\sqrt{p} + w$ in (7), which essentially guarantees $\mathrm{Var}\left(Z_j^\top (\mathbb{1}/\sqrt{p} + w_{t+1})\right) = 1$. As Z_j is sampled from $N(0, I)$, we further have $\mathbb{E}\left(Z_j^\top (\mathbb{1}/\sqrt{p} + w_{t+1})\right) = 0$. The normalization

step in (7) can be viewed as a population version of the widely used batch normalization trick to accelerate the training of neural networks (Ioffe and Szegedy, 2015). Moreover, (6) has one unique optimal solution under such a normalization. Specifically, (w^*, a^*) is the unique global optimum, and $(\bar{w}, \bar{a})$ is the only spurious local optimum along the solution path, where $\bar{w} = -(\mathbb{1}/\sqrt{p}) - v^*$ and $\bar{a} = (\mathbb{1}\mathbb{1}^\top + (\pi - 1)I)^{-1}(\mathbb{1}\mathbb{1}^\top - I)a^*$.

We initialize our algorithm at (w_0, a_0) satisfying: $w_0 = 0$ and $a_0 \in \mathbb{B}_0(|\mathbb{1}^\top a^*|/\sqrt{k})$. We set a_0 with a magnitude of $O(1/\sqrt{k})$ to match common initialization techniques (Glorot and Bengio, 2010; LeCun et al., 2012; He et al., 2015). We highlight that our algorithm starts with an arbitrary initialization on a, which is different from random initialization. The step sizes η_a and η_w will be specified later in our analysis.

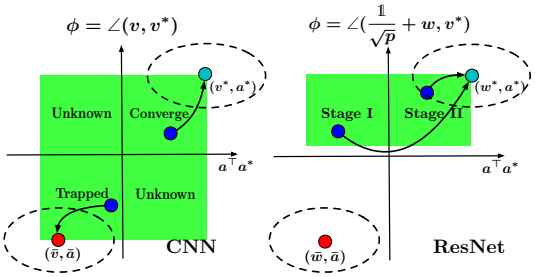

Figure 2: The left panel shows random initialization on feedforward CNN can be trapped in the spurious local optimum with probability at least $1/4$ (Du et al., 2017). The right panel demonstrates: 1). Under the shortcut prior, our initialization of (w, a) avoids starting near the spurious local optimum; 2). Convergence of GD exhibits two stages (I. improvement of a and avoid being attracted by $(\bar{w}, \bar{a})$ II. joint convergence).

3 Convergence Analysis

We characterize the algorithmic behavior of the gradient descent algorithm. Our analysis shows that under Assumption 1, the convergence of GD exhibits two stages. In the first stage, the algorithm avoids being trapped by the spurious local optimum. Given the algorithm is sufficiently away from the spurious local optima, the algorithm enters the basin of attraction of the global optimum and finally converge to it.

To present our main result, we begin with some notations. Denote

$$\phi_t = \angle(\mathbb{1}/\sqrt{p} + w_t, \mathbb{1}/\sqrt{p} + w^*)$$

as the angle between $\mathbb{1}/\sqrt{p} + w_t$ and the ground truth at the t-th iteration. Throughout the rest of the paper, we assume $\|a^*\|_2$ is a constant. The notation $\widetilde{O}(\cdot)$ hides $\mathrm{poly}(\|a^*\|_2)$, $\mathrm{poly}(\frac{1}{\|a^*\|_2})$, and $\mathrm{polylog}(\|a^*\|_2)$ factors. Then we state the convergence of GD in the following theorem.

Theorem 2 (Main Results)**.** *Let the GD algorithm defined in Section 2 be initialized with $w_0 = 0$ and arbitrary $a_0 \in \mathbb{B}_0(|\mathbb{1}^\top a^*|/\sqrt{k})$. Then the algorithm converges in two stages:*

Stage I: Avoid the spurious local optimum *(Theorem 4): We choose $\eta_a = O(1/k^2)$ and $\eta_w = \widetilde{O}(1/k^4)$. Then there exists $T_1 = \widetilde{O}(1/\eta_a)$, such that $m \le a_{T_1}^\top a^* \le M$ and $\phi_{T_1} \le \frac{5}{12}\pi$ hold for some constants $M > m > 0$.*

Stage II: Converge to the global optimum *(Theorem 13): After T_1 iterations, we restart the counter, and choose $\eta = \eta_a = \eta_w = \widetilde{O}(1/k^2)$. Then for any $\delta > 0$, any $t \ge T_2 = \widetilde{O}(\frac{1}{\eta}\log\frac{1}{\delta})$, we have $\|w_t - w^*\|_2^2 \le \delta$ and $\|a_t - a^*\|_2^2 \le 5\delta$.*

Note that the set $\{(w_t, a_t) \mid a_t^\top a^* \in [m, M], \phi_t \le 5\pi/12\}$ belongs to be the basin of attraction around the global optimum (Lemma 11), where certain regularity condition (partial dissipativity) guides the algorithm toward the global optimum. Hence, after the algorithm enters the second stage, we increase the step size η_w of w for a faster convergence. Figure 2 demonstrates the initialization of (w, a), and the convergence of GD both on CNN in Du et al. (2017) and our ResNet model.

We start our convergence analysis with the definition of partial dissipativity for $\mathcal{L}$.

Definition 3 (Partial Dissipativity)**.** *Given any $\delta \ge 0$ and a constant $c \ge 0$, $\nabla_w \mathcal{L}$ is (c, δ)-partially dissipative with respect to w^* in a set $\mathcal{K}_\delta$, if for every $(w, a) \in \mathcal{K}_\delta$, we have*

$$\langle -\nabla_w \mathcal{L}(w, a), w^* - w\rangle \ge c\|w - w^*\|_2^2 - \delta;$$

$\nabla_a \mathcal{L}$ is (c, δ)-partially dissipative with respect to a^* in a set $\mathcal{A}_\delta$, if for every $(w, a) \in \mathcal{A}_\delta$, we have

$$\langle -\nabla_a \mathcal{L}(w, a), a^* - a\rangle \ge c\|a - a^*\|_2^2 - \delta.$$

Moreover, If $\mathcal{K}_\delta \cap \mathcal{A}_\delta \neq \emptyset$, $\nabla\mathcal{L}$ is $(c, 2\delta)$-jointly dissipative with respect to (w^*, a^*) in $\mathcal{K}_\delta \cap \mathcal{A}_\delta$, i.e., for every $(w, a) \in \mathcal{K}_\delta \cap \mathcal{A}_\delta$, we have

$$\langle -\nabla_w \mathcal{L}(w, a), w^* - w\rangle + \langle -\nabla_a \mathcal{L}(w, a), a^* - a\rangle \ge c(\|w - w^*\|_2^2 + \|a - a^*\|_2^2) - 2\delta.$$

The concept of dissipativity is originally used in dynamical systems (Barrera and Jara, 2015), and is defined for general operators. It suffices to instantiate the concept to gradients here for our convergence analysis. The variational coherence studied in Zhou et al. (2017) and one point convexity studied in Li and Yuan (2017) can be viewed as special examples of partial dissipativity.

3.1 Stage I: Avoid the Spurious Local Optimum

We first show with properly chosen step sizes, GD algorithm can avoid being trapped by the spurious local optimum. We propose to update w, a using different step sizes. We formalize our result in the following theorem.

Theorem 4. *Initialize with arbitrary $a_0 \in \mathbb{B}_0(|\mathbb{1}^\top a^*|/\sqrt{k})$ and $w_0 = 0$. We choose step sizes*

$$\eta_a = \frac{\pi}{20(k+\pi-1)^2} = O\left(1/k^2\right), \quad \text{and} \quad \eta_w = C\|a^*\|_2^2\eta_a^2 = \widetilde{O}(\eta_a^2)$$

for some constant $C > 0$. Then, we have

$$\phi_t \le 5\pi/12 \quad \text{and} \quad 0 \le m \le a_t^\top a^* \le M, \tag{8}$$

for all $t \in [T_1, T]$, where

$$T_1 = \widetilde{O}(1/\eta_a), \ \ T = O(1/\eta_a^2), \ \ m = \|a^*\|_2^2/5, \ \ \text{and} \ \ M = 3\|a^*\|_2^2 + 2\left(\mathbb{1}^\top a^*\right)^2.$$

Proof Sketch. Due to the space limit, we only provide a proof sketch here. The detailed proof is deferred to Appendix B.2. We prove the two arguments in (8) in order. Before that, we first show our initialization scheme guarantees an important bound on a, as stated in the following lemma.

Lemma 5. *Given $w_0 = 0$ and $a_0 \in \mathbb{B}_0(|\mathbb{1}^\top a^*|/\sqrt{k})$, we choose $\eta_a \le \frac{2\pi}{k+\pi-1}$. Then for any $t > 0$,*

$$-3\left(\mathbb{1}^\top a^*\right)^2 \le \mathbb{1}^\top a^* \mathbb{1}^\top a_t - \left(\mathbb{1}^\top a^*\right)^2 \le 0. \tag{9}$$

Under the shortcut prior assumption 1 that w_0 is close to w^*, the update of w should be more conservative to provide enough accuracy for a to make progress. Based on Lemma 5, the next lemma shows that when η_w is small enough, ϕ_t stays acute ($\phi_t < \frac{\pi}{2}$), i.e., w is sufficiently away from $\bar{w} = -(\mathbb{1}/\sqrt{p}) - v^*$.

Lemma 6. *Given $w_0 = 0$ and $a_0 \in \mathbb{B}_0(|\mathbb{1}^\top a^*|/\sqrt{k})$, we choose $\eta_a < \frac{2\pi}{k+\pi-1}$ and $\eta_w = C\|a^*\|_2^2\eta_a^2 = \widetilde{O}(\eta_a^2)$ for some absolute constant $C > 0$. Then for all $t \le T = O(1/\eta_a^2)$,*

$$\phi_t \le 5\pi/12. \tag{10}$$

We want to remark that (9) and (10) are two of the key conditions that define the partially dissipative region of $\nabla_w \mathcal{L}$, as shown in the following lemma.

Lemma 7. *For any $(w, a) \in \mathcal{A}$, $\nabla_a \mathcal{L}$ satisfies*

$$\langle -\nabla_a \mathcal{L}(w, a), a^* - a\rangle \ge (1/10\pi)\|a - a^*\|_2^2, \tag{11}$$

where $\mathcal{A} = \Big\{(w, a) \ \Big| a^\top a^* \le \frac{1}{20}\|a^*\|_2^2$ or $\|a - a^*/2\|_2^2 \ge \|a^*\|_2^2, \ \|w + \mathbb{1}/\sqrt{p}\|_2 = 1, \ \phi \le \frac{5}{12}\pi, \ -3(\mathbb{1}^\top a^*)^2 \le \mathbb{1}^\top a^* \mathbb{1}^\top a - (\mathbb{1}^\top a^*)^2 \le 0\Big\}$.

Please refer to Appendix B.2.3 for a detailed proof. Note that with arbitrary initialization of a, $a^\top a^* \le \frac{1}{20}\|a^*\|_2^2$ or $\|a - a^*/2\|_2^2 \ge \|a^*\|_2^2$ possibly holds at a_0. In this case, (w_0, a_0) falls in $\mathcal{A}$, and (11) ensures the improvement of a.

Lemma 8. *Given $(w_0, a_0) \in \mathcal{A}$, we choose $\eta_a < \frac{\pi}{20(k+\pi-1)^2}$. Then there exists $\tau_{11} = O(1/\eta_a)$, such that $\frac{1}{20}\|a^*\|_2^2 \le a_{\tau_{11}}^\top a^* \le 2\|a^*\|_2^2$.*

One can easily verify that $a^\top a^* \le 2\|a^*\|_2^2$ holds for any $a \in \mathbb{B}_0(|\mathbb{1}^\top a^*|/\sqrt{k})$. Together with Lemma 8, we claim that even with arbitrary initialization, the iterates can always enter the region with $a^\top a^*$ positive and bounded in polynomial time. The next lemma shows that with proper chosen step sizes, $a^\top a^*$ stays positive and bounded.

Lemma 9. *Suppose $\frac{1}{20}\|a^*\|_2^2 \le a_0^\top a^* \le 2\|a^*\|_2^2$, $\phi_t \le \frac{5}{12}\pi$, and $-3\left(\mathbb{1}^\top a^*\right)^2 \le \mathbb{1}^\top a^* \mathbb{1}^\top a_t - \left(\mathbb{1}^\top a^*\right)^2 \le 0$ holds for all t. Choose $\eta_a < \frac{2\pi}{\pi-1}$, then we have for all $t \ge \tau_{12} = \widetilde{O}(1/\eta_a)$,*

$$\|a^*\|_2^2/5 \le a_t^\top a^* \le 3\|a^*\|_2^2 + 2\left(\mathbb{1}^\top a^*\right)^2.$$

Take $T_1 = \tau_{11} + \tau_{12}$, and we complete the proof. □

In Theorem 4, we choose a conservative η_w. This brings two benefits to the training process: 1). w stays away from $\bar{w}$. The update on w is quite limited, since η_w is small. Hence, w is kept sufficiently away from $\bar{w}$, even if w moves towards $\bar{w}$ in every iteration); 2). a continuously updates toward a^*.

Theorem 4 ensures that under the shortcut prior, GD with adaptive step sizes can successfully overcome the optimization challenge early in training, i.e., the iterate is sufficiently away from the spurious local optima at the end of Stage I. Meanwhile, (8) actually demonstrates that the algorithm enters the basin of attraction of the global optimum, and we next show the convergence of GD.

3.2 Stage II: Converge to the Global Optimum

Recall that in the previous stage, we use a conservative step size η_w to avoid being trapped by the spurious local optimum. However, the small step size η_w slows down the convergence of w in the basin of attraction of the global optimum. Now we choose larger step sizes to accelerate the convergence. The following theorem shows that, after Stage I, we can use a larger η_w, while the results in Theorem 4 still hold, i.e., the iterate stays in the basin of attraction of (w^*, a^*).

Theorem 10. *We restart the counter of time. Suppose $m \le a_0^\top a^* \le M$, and $\phi_0 \le \frac{5}{12}\pi$. We choose $\eta_w \le \frac{m}{M^2} = \widetilde{O}(\frac{1}{k^2})$ and $\eta_a < \frac{2\pi}{k+\pi-1}$. Then for all $t > 0$, we have*

$$\phi_t \le 5\pi/12 \quad \textit{and} \quad 0 \le m \le a_t^\top a^* \le M.$$

Proof Sketch. To prove the first argument, we need the partial dissipativity of $\nabla_w \mathcal{L}$.

Lemma 11. *For any $m > 0$, $\nabla_w \mathcal{L}$ satisfies*

$$\langle -\nabla_w \mathcal{L}(w, a), w^* - w \rangle \ge \frac{m}{8}\|w - w^*\|_2^2,$$

for any $(w, a) \in \mathcal{K}_m$, where

$$\mathcal{K}_m = \left\{(w, a) \;\middle|\; a^\top a^* \ge m,\ (w + \mathbb{1}/\sqrt{p})^\top v^* \ge 0,\ \|w + \mathbb{1}/\sqrt{p}\|_2 = 1\right\}.$$

This condition ensures that when $a^\top a^*$ is positive, w always makes positive progress towards w^*, or equivalently ϕ_t decreasing. We need not worry about ϕ_t getting obtuse, and thus a larger step size η_w can be adopted. The second argument can be proved following similar lines to Lemma 9. Please see Appendix B.3.2 for more details. □

Now we are ready to show the convergence of our GD algorithm. Note that Theorem 10 and Lemma 11 together show that the iterate stays in the partially dissipative region $\mathcal{K}_w$, which leads to the convergence of w. Moreover, as shown in the following lemma, when w is accurate enough, the partial gradient with respect to a enjoys partial dissipativity.

Lemma 12. *For any $\delta > 0$, $\nabla_a \mathcal{L}$ satisfies*

$$\langle -\nabla_a \mathcal{L}(w, a), a^* - a \rangle \ge \frac{\pi - 1}{2\pi}\|a - a^*\|_2^2 - \frac{1}{5}\delta,$$

for any $(w, a) \in \mathcal{A}_{m,M,\delta}$, where

$$\mathcal{A}_{m,M,\delta} = \left\{(w, a) \;\middle|\; a^\top a^* \in [m, M],\ \|w - w^*\|_2^2 \le \delta,\ \|w + \mathbb{1}/\sqrt{p}\|_2 = 1\right\}.$$

As a direct result, a converges to a^*. The next theorem formalize the above discussion.

Theorem 13 (Convergence)**.** *Suppose $\frac{1}{5}\|a^*\|_2^2 = m \le a_t^\top a^* \le M = 3\|a^*\|_2^2 + 2\left(\mathbb{1}^\top a^*\right)^2$ hold for all $t > 0$. For any $\delta > 0$, choose $\eta_a = \eta_w = \eta = \min\left\{\frac{m}{2M^2}, \frac{5\pi^2}{4(k+\pi-1)^2}\right\} = \widetilde{O}(\frac{1}{k^2})$, then we have*

$$\|w_t - w^*\|_2^2 \le \delta \ \textit{ and } \ \|a_t - a^*\|_2^2 \le 5\delta$$

for any $t \ge T_2 = \widetilde{O}(\frac{1}{\eta}\log\frac{1}{\delta})$.

Proof Sketch. The detailed proof is provided in Appendix B.4. Our proof relies on the partial dissipativity of $\nabla_w \mathcal{L}$ (Lemma 11) and that of $\nabla_a \mathcal{L}$ (Lemma 12).

Note that the partial dissipative region $\mathcal{A}_{m,M,\delta}$, depends on the precision of w. Thus, we first show the convergence of w.

Lemma 14 (Convergence of w_t)**.** *Suppose* $\frac{1}{5}\|a^*\|_2^2 = m \le a_t^\top a^* \le M = 3\|a^*\|_2^2 + 4\left(\mathbb{1}^\top a^*\right)^2$ *hold for all* $t > 0$*. For any* $\delta > 0$*, choose* $\eta \le \frac{m}{2M^2} = \widetilde{O}(\frac{1}{k^2})$*, then we have*

$$\|w_t - w^*\|_2^2 \le \delta$$

for any $t \ge \tau_{21} = \frac{4}{m\eta}\log\frac{4}{\delta} = \widetilde{O}(\frac{1}{\eta}\log\frac{1}{\delta})$.

Lemma 14 implies that after τ_{21} iterations, the algorithm enters $\mathcal{A}_{m,M,\delta}$. Then we show the convergence property of a in next lemma.

Lemma 15 (Convergence of a_t)**.** *Suppose* $m \le a_t^\top a^* \le M$ *and* $\|w_t - w^*\|_2^2 \le \delta$ *holds for all* t*. We choose* $\eta \le \frac{5\pi^2}{4(k+\pi-1)^2} = O(\frac{1}{k^2})$*. Then for all* $t \ge \tau_{22} = \frac{4}{\eta}\log\frac{\|a_0 - a^*\|_2^2}{\delta} = \widetilde{O}(\frac{1}{\eta}\log\frac{1}{\delta})$*, we have*

$$\|a_t - a^*\|_2^2 \le 5\delta.$$

Combine the above two lemmas together, take $T_2 = \tau_{21} + \tau_{22}$, and we complete the proof. □

Theorem 13 shows that with larger η_w than in Stage I, GD converges to the global optimum in polynomial time. Compared to the convergence with constant probability for CNN (Du et al., 2017), Assumption 1 assures convergence even under arbitrary initialization of a. This partially justifies the importance of shortcut in ResNet.

4 Numerical Experiment

We present numerical experiments to illustrate the convergence of the GD algorithm. We first demonstrate that with the shortcut prior, our choice of step sizes and the initialization guarantee the convergence of GD. We consider the training of a two-layer non-overlapping convolutional ResNet by solving (6). Specifically, we set $p = 8$ and $k \in \{16, 25, 36, 49, 64, 81, 100\}$. The teacher network is set with parameters a^* satisfying $\mathbb{1}^\top a^* = \frac{1}{4}\|a^*\|_2^2$, and v^* satisfying $v_1^* = \cos(7\pi/10)$, $v_2^* = \sin(7\pi/10)$, and $v_j^* = 0$ for $j = 3, \ldots, p$.[2] More detailed experimental setting is provided in Appendix C. We initialize with $w_0 = 0$ and a_0 uniformly distributed over $\mathbb{B}_0(|\mathbb{1}^\top a^*|/\sqrt{k})$. We adopt the following learning rate scheme with Step Size Warmup (SSW) suggested in Section 3: We first choose step sizes $\eta_a = 1/k^2$ and $\eta_w = \eta_a^2$, and run for 1000 iterations. Then, we choose $\eta_a = \eta_w = 1/k^2$. We also consider learning the same teacher network using step sizes $\eta_w = \eta_a = 1/k^2$ throughout, i.e., without step size warmup.

We further demonstrate learning the aforementioned teacher network using a student network of the same architecture. Specifically, we keep a^*, v^* unchanged. We use the GD in Du et al. (2017) with step size $\eta = 0.1$, and initialize v_0 uniformly distributed over the unit sphere and a uniformly distributed over $\mathbb{B}_0(|\mathbb{1}^\top a^*|/\sqrt{k})$.

For each combination of k and a^*, we repeat 5000 simulations for aforementioned three settings, and report the success rate of converging to the global optimum in Table 1. As can be seen, our GD on ResNet is capable of avoiding the spurious local optimum, and converges to the global optimum in all 5000 simulations. However, GD without SSW can be trapped in the spurious local optimum. The failure probability diminishes as the dimension increase. Learning the teacher network using a two-layer CNN student network (Du et al., 2017) can also be trapped in the spurious local optimum.

Table 1: *Success rates of converging to the global optimum for GD training ResNet with and without SSW and CNN with varying* k *and and* $p = 8$.

k	16	25	36	49	64	81	100
ResNet w/ SSW	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
ResNet w/o SSW	0.7042	0.7354	0.7776	0.7848	0.8220	0.8388	0.8426
CNN	0.5348	0.5528	0.5312	0.5426	0.5192	0.5368	0.5374

We then demonstrate the algorithmic behavior of our GD. We set $k = 25$ for the teacher network, and other parameters the same as in the previous experiment. We initialize $w_0 = 0$ and $a_0 \in \mathbb{B}_0(|\mathbb{1}^\top a^*|/\sqrt{k})$. We start with $\eta_a = 1/k^2$ and $\eta_w = \eta_a^2$. After 1000 iterations, we set the step sizes $\eta_a = \eta_w = 1/k^2$. The algorithm is terminated when $\|a_t - a^*\|_2^2 + \|w_t - w^*\|_2^2 \le 10^{-6}$. We also demonstrate the GD algorithm without SSW at the same initialization. The step sizes are $\eta_a = \eta_w = 1/k^2$ throughout the training.

[2] v^* essentially satisfies $\angle(v^*, \mathbb{1}/\sqrt{p}) = 0.45\pi$.

One solution path of GD with SSW is shown in the first column of Figure 3. As can be seen, the algorithm has a phase transition. In the first stage, we observe that w_t makes very slow progress due to the small step size η_w. While $a_t^\top a^*$ gradually increases. This implies the algorithm avoids being attracted by the spurious local optimum. In the second stage, w_t and a_t both continuously evolve towards the global optimum.

The second row of Figure 3 illustrates the trajectory of GD without SSW being trapped by the spurious local optimum. Specifically, (w_t, a_t) converges to $(\bar{w}, \bar{a})$ as we observe that ϕ_t converges to π, and $\|w_t - w^*\|_2^2$ converges to $4\|v^*\|_2^2$.

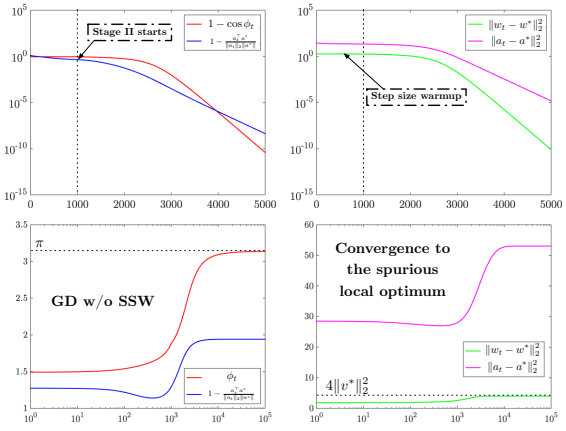

Figure 3: Algorithmic behavior of GD on ResNet. The horizontal axis corresponds to the number of iterations.

5 Discussions

Deep ResNet. Our two-layer network model is largely simplified compared with deep and wide ResNets in practice, where the role of the shortcut connection is more complicated. It is worth mentioning that the empirical results in Veit et al. (2016) show that ResNet can be viewed as an ensemble of smaller networks, and most of the smaller networks are shallow due to the shortcut connection. They also suggest that the training is dominated by the shallow smaller networks. We are interested in investigating whether these shallow smaller networks possesses similar benign properties to ease the training as our two-layer model.

Moreover, our student network and the teacher network have the same degree of freedom. We have not considered deeper and wider student networks. It is also worth an investigation that what is the role of shortcut connections in deeper and wider networks.

From GD to SGD. A straightforward extension is to investigate the convergence of SGD with mini-batch. We remark that when the batch size is large, the effect of the noise on gradient is limited and SGD mimics the behavior of GD. When the batch size is small, the noise on gradient plays a significant role in training, which is technically more challenging.

Related Work. Li and Yuan (2017) study ResNet-type two-layer neural networks with the output weight known ($a = \mathbb{1}$), which is equivalent to assuming $a_t^\top a^* > 0$ for all t in our analysis. Thus, their analysis does not have Stage I ($a_0^\top a^* < 0$). Moreover, since they do not need to optimize a, they only need to handle the partial dissipativity of $\nabla \mathcal{L}_w$ with $\delta = 0$ (one-point convexity). In our analysis, however, we also need to handle the the partial dissipativity of $\nabla \mathcal{L}_a$ with $\delta \neq 0$, which makes our proof more involved.

Initialization. Our analysis shows that GD converges to the global optimum, when w is initialized at zero. Empirical results in Li et al. (2016) and Zhang et al. (2019) also suggest that deep ResNet works well, when the weights are simply initialized at zero or using the Fixup initialization. We are interested in building a connection between training a two-layer ResNet and its deep counterpart.

Step Size Warmup. Our choice of step size η_w is related to the learning rate warmup and layerwise learning rate in the existing literature. Specifically, Goyal et al. (2017) presents an effective learning rate scheme for training ResNet on ImageNet for less than 1 hour. They start with a small step size, gradually increase (linear scale) it, and finally shrink it for convergence. Our analysis suggests that in the first stage, we need smaller η_w to avoid being attracted by the spurious local optimum. This is essentially consistent with Goyal et al. (2017). Note that we are considering GD (no noise), hence, we do not need to shrink the step size in the final stage. While Goyal et al. (2017) need to shrink the step size to control the noise in SGD. Similar learning rate schemes are proposed by Smith (2017).

On the other hand, we incorporate the shortcut prior, and adopt a smaller step size for the inner layer, and a larger step size for the outer layer. Such a choice of step size is shown to be helpful in both deep learning and transfer learning (Singh et al., 2015; Howard and Ruder, 2018), where it is referred to as differential learning rates or discriminative fine-tuning. It is interesting to build a connection between our theoretical discoveries and these empirical observations.

References

BAHDANAU, D., CHO, K. and BENGIO, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

BALDUZZI, D., FREAN, M., LEARY, L., LEWIS, J., MA, K. W.-D. and MCWILLIAMS, B. (2017). The shattered gradients problem: If resnets are the answer, then what is the question? In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

BARRERA, G. and JARA, M. (2015). Thermalisation for stochastic small random perturbations of hyperbolic dynamical systems. *arXiv preprint arXiv:1510.09207* .

BARTLETT, P. L., EVANS, S. N. and LONG, P. M. (2018). Representing smooth functions as compositions of near-identity functions with implications for deep network optimization. *arXiv preprint arXiv:1804.05012* .

DU, S. S., LEE, J. D., TIAN, Y., POCZOS, B. and SINGH, A. (2017). Gradient descent learns one-hidden-layer cnn: Don't be afraid of spurious local minima. *arXiv preprint arXiv:1712.00779* .

GLOROT, X. and BENGIO, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*.

GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*.

GOYAL, P., DOLLÁR, P., GIRSHICK, R., NOORDHUIS, P., WESOLOWSKI, L., KYROLA, A., TULLOCH, A., JIA, Y. and HE, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677* .

GRAVES, A., MOHAMED, A.-R. and HINTON, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE.

HARDT, M. and MA, T. (2016). Identity matters in deep learning. *arXiv preprint arXiv:1611.04231* .

HE, K., ZHANG, X., REN, S. and SUN, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*.

HE, K., ZHANG, X., REN, S. and SUN, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

HE, K., ZHANG, X., REN, S. and SUN, J. (2016b). Identity mappings in deep residual networks. *arXiv preprint arXiv:1603.05027* .

HOWARD, J. and RUDER, S. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146* .

HUANG, G., LIU, Z., VAN DER MAATEN, L. and WEINBERGER, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

IOFFE, S. and SZEGEDY, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* .

KRIZHEVSKY, A., SUTSKEVER, I. and HINTON, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

LECUN, Y. A., BOTTOU, L., ORR, G. B. and MÜLLER, K.-R. (2012). Efficient backprop. In *Neural networks: Tricks of the trade*. Springer, 9–48.

LI, H., XU, Z., TAYLOR, G., STUDER, C. and GOLDSTEIN, T. (2018). Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*.

LI, S., JIAO, J., HAN, Y. and WEISSMAN, T. (2016). Demystifying resnet. *arXiv preprint arXiv:1611.01186* .

LI, Y. and YUAN, Y. (2017). Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*.

LONG, J., SHELHAMER, E. and DARRELL, T. (2015). Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

SINGH, B., DE, S., ZHANG, Y., GOLDSTEIN, T. and TAYLOR, G. (2015). Layer-specific adaptive learning rates for deep networks. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. IEEE.

SMITH, L. N. (2017). Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE.

SMITH, L. N. and TOPIN, N. (2018). Super-convergence: Very fast training of residual networks using large learning rates .

SRIVASTAVA, R. K., GREFF, K. and SCHMIDHUBER, J. (2015). Training very deep networks. In *Advances in neural information processing systems*.

VEIT, A., WILBER, M. J. and BELONGIE, S. (2016). Residual networks behave like ensembles of relatively shallow networks. In *Advances in Neural Information Processing Systems*.

YOUNG, T., HAZARIKA, D., PORIA, S. and CAMBRIA, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine* **13** 55–75.

YU, X., YU, Z. and RAMALINGAM, S. (2018). Learning strict identity mappings in deep residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

ZHANG, H., DAUPHIN, Y. N. and MA, T. (2019). Fixup initialization: Residual learning without normalization. *arXiv preprint arXiv:1901.09321* .

ZHOU, Z., MERTIKOPOULOS, P., BAMBOS, N., BOYD, S. and GLYNN, P. W. (2017). Stochastic mirror descent in variationally coherent optimization problems. In *Advances in Neural Information Processing Systems*.

Supplementary Material for Understanding the Importance of Shortcut Connections in ResNet

A Preliminaries

We first provide the explicit forms of the loss function and its gradients with respect to w and a.

Proposition 16. *Let $\phi = \angle(\mathbb{1}/\sqrt{p} + w, v^*)$. When $\|\mathbb{1}/\sqrt{p} + w\|_2 = 1$, the loss function $\mathcal{L}(w, a)$ and the gradient w.r.t (w, a), i.e., $\nabla_a \mathcal{L}(w, a)$ and $\nabla_w \mathcal{L}(w, a)$ have the following analytic forms.*

$$\mathcal{L}(w,a) = \frac{1}{2}[\frac{(\pi-1)}{2\pi}\|a^*\|_2^2 + \frac{(\pi-1)}{2\pi}\|a\|_2^2 - \frac{1}{\pi}(g(\phi)-1)\,a^\top a^* + \frac{1}{2\pi}\left(\mathbb{1}^\top a^*\right)^2 + \frac{1}{2\pi}\left(\mathbb{1}^\top a\right)^2 - \frac{1}{\pi}\mathbb{1}^\top a^* a^\top \mathbb{1}],$$

$$\nabla_a \mathcal{L}(a,w) = \frac{1}{2\pi}(\mathbb{1}\mathbb{1}^\top + (\pi-1))a - \frac{1}{2\pi}(\mathbb{1}\mathbb{1}^\top + (g(\phi)-1))a^*,$$

$$\nabla_w \mathcal{L}(w,a) = -\frac{a^\top a^*(\pi-\phi)}{2\pi}\left(I - (\mathbb{1}/\sqrt{p}+w)(\mathbb{1}/\sqrt{p}+w)^\top\right)v^*,$$

where $g(\phi) = (\pi-\phi)\cos(\phi) + \sin(\phi)$.

This proposition is a simple extension of Theorem 3.1 in Du et al. (2017). Here, we omit the proof.

For notational simplicity, we denote $v_t = \mathbb{1}/\sqrt{p} + w_t$ in the future proof.

B Proof of Theoretical Results

B.1 Proof of Proposition B.1

Proof. Recall that Du et al. (2017) proves that $(\bar{v}, \bar{a}) = (-v^*, (\mathbb{1}\mathbb{1}^\top + (\pi-1)I)^{-1}(\mathbb{1}\mathbb{1}^\top - I)a^*)$ is the spurious local optimum of the CNN counterpart to our ResNet. Substitute $\bar{v}$ by $\frac{\mathbb{1}/\sqrt{p}+w}{\|\mathbb{1}/\sqrt{p}+w\|_2}$ and we prove the result. □

B.2 Proof of Theorem 4

B.2.1 Proof of Lemma 5

Proof. By simple manipulication, we know that the initialization of a satisfies $-2\left(\mathbb{1}^\top a^*\right)^2 \le \mathbb{1}^\top a^* \mathbb{1}^\top a_0 - \left(\mathbb{1}^\top a^*\right)^2 \le 0$. We first prove the right side of the inequality. Expand a_t as $a_{t-1} - \eta_a \nabla_a \mathcal{L}(w_{t-1}, a_{t-1})$, and we have

$$\begin{aligned}\mathbb{1}^\top a^* \mathbb{1}^\top a_t &= \left(1 - \frac{\eta_a(k+\pi-1)}{2\pi}\right)\mathbb{1}^\top a^* \mathbb{1}^\top a_{t-1} + \frac{\eta_a(k+g(\phi_{t-1})-1)}{2\pi}\left(\mathbb{1}^\top a^*\right)^2 \\ &\le \left(1 - \frac{\eta_a(k+\pi-1)}{2\pi}\right)\mathbb{1}^\top a^* \mathbb{1}^\top a_{t-1} + \frac{\eta_a(k+\pi-1)}{2\pi}\left(\mathbb{1}^\top a^*\right)^2.\end{aligned}$$

Subtract $\left(\mathbb{1}^\top a^*\right)^2$ from both sides, then we get

$$\begin{aligned}\mathbb{1}^\top a^* \mathbb{1}^\top a_t - \left(\mathbb{1}^\top a^*\right)^2 &\le \left(1 - \frac{\eta_a(k+\pi-1)}{2\pi}\right)(\mathbb{1}^\top a^* \mathbb{1}^\top a_{t-1} - \left(\mathbb{1}^\top a^*\right)^2) \\ &\le \left(1 - \frac{\eta_a(k+\pi-1)}{2\pi}\right)^t (\mathbb{1}^\top a^* \mathbb{1}^\top a_0 - \left(\mathbb{1}^\top a^*\right)^2) \le 0,\end{aligned}$$

for any $t \ge 1$. The right side inequality is proved.

The proof of the left side follows similar lines. Since $g(\phi) \geq 0$, we have

$$\begin{aligned}\mathbb{1}^\top a^* \mathbb{1}^\top a_t &= \left(1 - \frac{\eta_a\left(k+\pi-1\right)}{2\pi}\right)\mathbb{1}^\top a^* \mathbb{1}^\top a_{t-1} + \frac{\eta_a\left(k+g\left(\phi_{t-1}\right)-1\right)}{2\pi}\left(\mathbb{1}^\top a^*\right)^2 \\ &\geq \left(1 - \frac{\eta_a\left(k+\pi-1\right)}{2\pi}\right)\mathbb{1}^\top a^* \mathbb{1}^\top a_{t-1} + \eta_a\frac{k-1}{2\pi}\left(\mathbb{1}^\top a^*\right)^2,\end{aligned}$$

which is equivalent to the following inequality.

$$\begin{aligned}\mathbb{1}^\top a^* \mathbb{1}^\top a_t - \left(\mathbb{1}^\top a^*\right)^2 &\geq \left(1 - \frac{\eta_a\left(k+\pi-1\right)}{2\pi}\right)\left(\mathbb{1}^\top a^* \mathbb{1}^\top a_{t-1} - \left(\mathbb{1}^\top a^*\right)^2\right) - \frac{\eta_a}{2}\left(\mathbb{1}^\top a^*\right)^2. \\ &\geq \left(1 - \frac{\eta_a\left(k+\pi-1\right)}{2\pi}\right)^t\left(\mathbb{1}^\top a^* \mathbb{1}^\top a_0 - \left(\mathbb{1}^\top a^*\right)^2\right) - \frac{1}{1-\left(1-\frac{\eta_a(k+\pi-1)}{2\pi}\right)}\frac{\eta_a}{2}\left(\mathbb{1}^\top a^*\right)^2 \\ &\geq \left(1 - \frac{\eta_a\left(k+\pi-1\right)}{2\pi}\right)^t\left(\mathbb{1}^\top a^* \mathbb{1}^\top a_0 - \left(\mathbb{1}^\top a^*\right)^2\right) - \frac{\pi}{k+\pi-1}\left(\mathbb{1}^\top a^*\right)^2 \\ &\geq \left(1 - \frac{\eta_a\left(k+\pi-1\right)}{2\pi}\right)^t\left(-2\left(\mathbb{1}^\top a^*\right)^2\right) - \frac{\pi}{k+\pi-1}\left(\mathbb{1}^\top a^*\right)^2 \\ &\geq -3\left(\mathbb{1}^\top a^*\right)^2.\end{aligned}$$

Then we prove the lemma. □

B.2.2 Proof of Lemma 6

Proof. For each iteration, the distance of w_t moving towards $\bar{w}$ is upper bounded by the product of the step size η_w and the norm of the gradient $\nabla_w \mathcal{L}\left(w, a\right)$. We first bound the norm of the gradient. From the analytic form of $\nabla_w \mathcal{L}\left(w, a\right)$, we need to bound $a^\top a^*$. We first have the following lower bound.

$$\begin{aligned}a_{t+1}a^* &= \left(1 - \frac{\eta_a\left(\pi-1\right)}{2\pi}\right)a_t^\top a^* + \frac{\eta_a\left(g\left(\phi_t\right)-1\right)}{2\pi}\|a^*\|_2^2 + \frac{\eta_a}{2\pi}\left(\left(\mathbb{1}^\top a^*\right)^2 - \mathbb{1}^\top a^* \mathbb{1}^\top a_t\right) \\ &\geq \left(1 - \frac{\eta_a\left(\pi-1\right)}{2\pi}\right)a_t^\top a^* - \eta_a\frac{2}{\pi}\|a^*\|_2^2,\end{aligned}$$

which is equivalent to

$$\begin{aligned}a_{t+1}a^* + \frac{4}{\pi-1}\|a^*\|_2^2 &\geq \left(1 - \frac{\eta_a\left(\pi-1\right)}{2\pi}\right)\left(a_t^\top a^* + \frac{4}{\pi-1}\|a^*\|_2^2\right) \\ &\geq \left(1 - \frac{\eta_a\left(\pi-1\right)}{2\pi}\right)^{t+1}\left(a_0^\top a^* + \frac{4}{\pi-1}\|a^*\|_2^2\right).\end{aligned}$$

Since $a_0^\top a^* \geq -\|a^*\|_2^2$, we have $a_0^\top a^* + \frac{4}{\pi-1}\|a^*\|_2^2 \geq 0$. Thus, when $\eta_a < \frac{2\pi}{\pi-1}$,

$$a_{t+1}a^* \geq -\frac{4}{\pi-1}\|a^*\|_2^2 \geq -2\|a^*\|_2^2.$$

When $a_{t+1}a^* < 2\|a^*\|_2^2$, the following inequality holds true.

$$\begin{aligned}\|\nabla_w L\left(w_t, a_t\right)\|_2^2 &= \frac{\left(a_t^\top a^*\right)^2\left(\pi-\phi_t\right)^2}{4\pi^2}v^{*\top}\left(I - v_t v_t^\top\right)v^* \\ &\leq \|a^*\|_2^4(I - v_t^\top v^*)(I + v_t^\top v^*) \leq \|a^*\|_2^4\|v_t - v^*\|_2^2.\end{aligned}$$

We next prove that when η_w is small enough, $\phi_t < \pi/2$ holds for all $t \leq T = O(1/\eta_a^2)$. We first have the following inequality.

$$1 \leq \|\widetilde{v}_{t+1}\|_2 = \sqrt{\|v_t\|_2^2 + \|\eta_w \nabla_w L\left(w_t, a_t\right)\|_2^2} \leq 1 + \|\eta_w \nabla_w L\left(w_t, a_t\right)\|_2.$$

Under Assumption 1, we know that $\phi_0 < \pi/3$. Then we can bound the norm of the difference between iterates w_{t+1} and w^* as follows.

$$\begin{aligned}\|v_{t+1} - v^*\|_2 &= \|\widetilde{v}_{t+1}/\|\widetilde{v}_{t+1}\|_2 - v^*\|_2 \leq \frac{1}{\|\widetilde{v}_{t+1}\|_2}\|\widetilde{v}_{t+1} - v^*\|_2 + 1 - \frac{1}{\|\widetilde{v}_{t+1}\|_2} \\ &\leq \|\widetilde{v}_{t+1} - v^*\|_2 + 1 - \frac{1}{1 + \|\eta_w \nabla_w L\left(w_t, a_t\right)\|_2}.\end{aligned}$$

Plug in the upper bound of the norm of $\nabla_w L\left(w_t, a_t\right)$, and we obtain

$$\begin{aligned}
\|v_{t+1} - v^*\|_2 &\leq \|\widetilde{v}_{t+1} - v^*\|_2 + 1 - \frac{1}{1 + \eta_w \|a^*\|_2^2 \|v_t - v^*\|_2} \\
&= \|v_t - v^* - \eta_w \nabla_w \mathcal{L}(a_t, w_t)\|_2^2 + \frac{\eta_w \|a^*\|_2^2 \|v_t - v^*\|_2}{1 + \eta_w \|a^*\|_2^2 \|v_t - v^*\|_2} \\
&\leq \|v_t - v^*\|_2 + \eta_w \|\nabla_w \mathcal{L}(a_t, w_t)\|_2 + \eta_w \|a^*\|_2^2 \|v_t - v^*\|_2 \\
&\leq \|v_t - v^*\|_2 + \eta_w \|a^*\|_2^2 \|v_t - v^*\|_2 + \eta_w \|a^*\|_2^2 \|v_t - v^*\|_2 \\
&= (1 + 2\eta_w \|a^*\|_2^2) \|v_t - v^*\|_2 \leq (1 + 2\eta_w \|a^*\|_2^2)^t \|v_0 - v^*\|_2 \\
&\leq \exp(2t\eta_w \|a^*\|_2^2) \|v_0 - v^*\|_2 \\
&\leq \exp(2t\eta_w \|a^*\|_2^2) \leq 2 - 2\cos\left(\frac{5}{12}\pi\right),
\end{aligned}$$

for all $t \leq T = O(1/\eta_a^2)$, when $\eta_w = C_1 \|a^*\|_2^2 \eta_a^2 = \widetilde{O}(\eta_a^2)$ for some constant $C_1 > 0$. Thus $\phi_t \leq \frac{5}{12}\pi$ for all $t \leq T = O(1/\eta_a^2)$. □

B.2.3 Proof of Lemma 7

Proof. For any $C_3 \in (0, 1)$, if we have $a^\top a^* \leq C_3 \|a^*\|_2^2$, the norm of the difference between a and a^* satisfies the following inequality.

$$\|a - a^*\|_2^2 \geq (1 - 2C_3) \|a^*\|_2^2.$$

Let $C_2 = g(\frac{5}{12}\pi) - 1 = 0.4402$. SInce $\phi \leq \frac{5}{12}\pi$, and g is strictly decreasing, we know that $g(\phi) \geq C_2$. Using the above two inequalities, we can lower bound the inner product between the negative gradient and the difference between a and a^* as follows.

$$\begin{aligned}
\left\langle -\nabla_a L\left(w + \xi, a + \epsilon\right), a^* - a \right\rangle &= \frac{1}{2\pi} \left(\mathbb{1}^\top a - \mathbb{1}^\top a^*\right)^2 + \frac{1}{2\pi} \left((\pi - 1) a - \left(g\left(\phi\right) - 1\right) a^*\right)^\top \left(a - a^*\right) \\
&= \frac{1}{2\pi} \left(\mathbb{1}^\top a - \mathbb{1}^\top a^*\right)^2 + \frac{1}{2\pi} \left(\pi - g\left(\phi\right)\right) a^\top \left(a - a^*\right) + \frac{g\left(\phi\right) - 1}{2\pi} \|a - a^*\|_2 \\
&\geq -\frac{1}{2\pi} \left(\pi - g\left(\phi\right)\right) a^\top a^* + \frac{g\left(\phi\right) - 1}{2\pi} \|a - a^*\|_2^2 \\
&\geq -\frac{1}{2\pi} \left(\pi - g\left(\phi\right)\right) a^\top a^* + \frac{g\left(\phi\right) - 1}{4\pi} \|a - a^*\|_2^2 + \frac{g\left(\phi\right) - 1}{4\pi} \|a - a^*\|_2^2 \\
&\geq -\frac{C_3}{2} \|a^*\|_2^2 + \frac{C_2}{4\pi} \left(1 - 2C_3\right) \|a^*\|_2^2 + \frac{C_2}{4\pi} \|a - a^*\|_2^2 \\
&\geq \frac{C_2}{4\pi} \|a - a^*\|_2^2 \geq \frac{1}{10\pi} \|a - a^*\|_2^2,
\end{aligned}$$

when $C_3 \leq \frac{C_2}{2(C_2 + \pi)}$. Take $C_3 = \frac{1}{20}$, and we prove the result. □

B.2.4 Proof of Lemma 8

Proof. We prove the result by contradiction. Specifically, we show that if $a_t \in \mathcal{A}$ always holds, there always exist some time τ such that $a_\tau \notin \mathcal{A}$, which is a contradiction. Formally, suppose $\forall \tau \leq t, a_\tau \in \mathcal{A}$, then we have

$$\begin{aligned}
\|a_{t+1} - a^*\|_2^2 = \|a_t - a^*\|_2^2 &- 2\langle -\eta_a \mathbb{E}_{\xi,\epsilon} \nabla_a L\left(w_t, a_t\right), a^* - a_t \rangle \qquad (12) \\
&+ \|\eta_a \nabla_a L\left(w_t, a_t\right)\|_2^2. \qquad (13)
\end{aligned}$$

The second term is lower bounded according to the partial dissipativity of $\nabla_a \mathcal{L}$. Thus, we only need to bound the norm of the gradient.

$$\begin{aligned}
\|\nabla_a L\left(w_t, a_t\right)\|_2^2 &= \|\nabla_a L\left(w_t, a_t\right) - \nabla_a L\left(w^*, a^*\right)\|_2^2 \\
&= \|\frac{1}{2\pi} \left(\mathbb{1}\mathbb{1}^\top + (\pi - 1) I\right) \left(a_t - a^*\right) - \frac{g\left(\phi\right) - \pi}{2\pi} a^*\|_2^2 \\
&\leq \frac{1}{2\pi^2} \|\left(\mathbb{1}\mathbb{1}^\top + (\pi - 1) I\right) \left(a_t - a^*\right)\|_2^2 + \frac{1}{2} \|a^*\|_2^2 \\
&\leq \frac{\left(k + \pi - 1\right)^2}{\pi^2} \left(\|a_t - a^*\|_2^2\right) + \frac{1}{2} \|a^*\|_2^2.
\end{aligned}$$

Plug the above bound into (12), then we have

$$\begin{aligned}\|a_{t+1}-a^*\|_2^2 &\le \left(1-\frac{\pi}{5}\eta_a+\eta_a^2\frac{(k+\pi-1)^2}{\pi^2}\right)\|a_t-a^*\|_2^2+\frac{\eta_a^2}{2}\|a^*\|_2^2\\ &\le (1-\lambda_1)\|a_t-a^*\|_2^2+b_1\\ &\le (1-\lambda_1)^{t+1}\|a_0-a^*\|_2^2+\frac{b_1}{\lambda_1},\end{aligned}$$

where $\lambda_1=\frac{\pi}{5}\eta_a-\eta_a^2\frac{(k+\pi-1)^2}{\pi^2}$ and $b_1=\frac{\eta_a^2}{2}\|a^*\|_2^2$. When $\eta_a<\frac{\pi}{20(k+\pi-1)^2}$, we have $\frac{b_1}{\lambda_1}\le\frac{\|a^*\|_2^2}{6}$. Thus, after $\tau_{11}=O(\frac{1}{\eta_a})$ iterations, we have

$$\|a_{\tau_{11}}-a^*\|_2^2<\frac{\|a^*\|_2^2}{4}.$$

On the other hand, $a_{\tau_{11}}\in\mathcal{A}$ implies that $\|a_{\tau_{11}}-a^*\|_2^2\ge\frac{1}{4}\|a^*\|_2^2$. Thus, after $\tau_{11}=O(\frac{1}{\eta_a})$ iterations, we have

$$\frac{1}{20}\|a^*\|_2^2\le a_t^\top a^* \text{ and } \|a_t-a^*/2\|_2^2\le\|a^*\|_2^2.$$

Moreover, $\|a_t-a^*/2\|_2^2\le\|a^*\|_2^2$ implies $a_t^\top a^*\le 2\|a^*\|_2^2$, and we prove the lemma. □

B.2.5 Proof of Lemma 9

Proof. We first prove the left side. Write $a_{t+1}=a_t-\eta_a\nabla_a\mathcal{L}(w,a)$ and we have

$$\begin{aligned}a_{t+1}^\top a^* &= \left(1-\frac{\eta_a(\pi-1)}{2\pi}\right)a_t^\top a^*+\frac{\eta_a(g(\phi_t)-1)}{2\pi}\|a^*\|_2^2+\frac{\eta_a}{2\pi}\left(\left(\mathbb{1}^\top a^*\right)^2-\mathbb{1}^\top a^*\mathbb{1}^\top a_t\right)\\ &\ge \left(1-\frac{\eta_a(\pi-1)}{2\pi}\right)a_t^\top a^*+\eta_a\frac{C_2}{2\pi}\|a^*\|_2^2.\end{aligned}$$

The last inequality holds since $g(\phi)\ge 1$ and $\left(\mathbb{1}^\top a^*\right)^2-\mathbb{1}^\top a^*\mathbb{1}^\top a_t\ge 0$. Subtract $\frac{C_2}{\pi-1}\|a^*\|_2^2$ from both sides and we have the following inequality

$$\begin{aligned}a_{t+1}^\top a^*-\frac{C_2}{\pi-1}\|a^*\|_2^2 &\ge \left(1-\frac{\eta_a(\pi-1)}{2\pi}\right)\left(a_t^\top a^*-\frac{C_2}{\pi-1}\|a^*\|_2^2\right)\\ &\ge \left(1-\frac{\eta_a(\pi-1)}{2\pi}\right)^t\left(a_0^\top a^*-\frac{C_2}{\pi-1}\|a^*\|_2^2\right).\end{aligned}$$

Thus, when $t\ge\tau_{12}=\widetilde{O}(1/\eta_a)>0$, we have $a_t^\top a^*\ge\frac{1}{5}\|a^*\|_2^2$.

For the right side, follows similar lines to the left side, we have

$$\begin{aligned}a_{t+1}^\top a^* &= \left(1-\frac{\eta_a(\pi-1)}{2\pi}\right)a_t^\top a^*+\frac{\eta_a(g(\phi_t)-1)}{2\pi}\|a^*\|_2^2+\frac{\eta_a}{2\pi}\left(\left(\mathbb{1}^\top a^*\right)^2-\mathbb{1}^\top a^*\mathbb{1}^\top a_t\right)\\ &\le \left(1-\frac{\eta_a(\pi-1)}{2\pi}\right)a_t^\top a^*+\eta_a\frac{\pi-1}{2\pi}\|a^*\|_2^2+\eta_a\frac{3}{2\pi}(\mathbb{1}^\top a^*)^2\\ &\le \left(1-\frac{\eta_a(\pi-1)}{2\pi}\right)^{t+1}a_0^\top a^*+\|a^*\|_2^2+\frac{3}{\pi-1}(\mathbb{1}^\top a^*)^2.\end{aligned}$$

Note that $a_0^\top a^*\le 2\|a^*\|_2^2$. Thus, for all t, $a_{t+1}^\top a^*\le 3\|a^*\|_2^2+2(\mathbb{1}^\top a^*)^2$. □

B.3 proof of Theorem 10

B.3.1 Proof of Lemma 11

Proof. Note that $\|v_t\|_2=\|v^*\|_2=1$, according to Proposition 16, the gradient with respect to w can be rewritten as

$$\nabla_w\mathcal{L}(w_t,a_t)=-\frac{a_t^\top a_t^*(\pi-\phi_t)}{2\pi}\left(I-v_tv_t^\top\right)v^*.$$

Then we have the following inequality.

$$\begin{aligned}\langle -\nabla_w \mathcal{L}(w_t, a_t), w^* - w_t\rangle &= \langle -\nabla_w \mathcal{L}(w_t, a_t), v^* - v_t\rangle \\ &= \frac{a_t^\top a_t^* (\pi - \phi_t)}{2\pi}\left(1 - (v_t^\top v^*)^2\right) \\ &\geq \frac{m}{4}(1 - v_t^\top v^*) \\ &= \frac{m}{8}\|w - w^*\|_2^2.\end{aligned}$$

□

B.3.2 proof of Theorem 10

Proof. First, we bound the norm of the gradient as follows

$$\|\nabla_w L\left(w, a\right)\|_2^2 = \frac{\left(a^\top a^*\right)^2 (\pi - \phi)^2}{4\pi^2} v^{*\top}\left(I - vv^\top\right) v^* \leq \frac{M^2}{4}(I - v^\top v^*)(I + v^\top v^*) \leq \frac{M^2}{4}\|v - v^*\|_2^2.$$

Next we show that $\|v_{t+1} - v^*\|_2^2 \leq \|\widetilde{v}_{t+1} - v^*\|_2^2$. We first have the following two inequalities.

$$\|\widetilde{v}_{t+1}\|_2^2 = \|v_t\|_2^2 + \|\eta_w \nabla_w L\left(w_t, a_t\right)\|_2^2 \geq 1.$$

$$\widetilde{v}_{t+1}^\top v^* = v_t^\top v^* + \eta_w \langle -\nabla_w \mathcal{L}(w_t + \xi, a_t + \epsilon), v^* - v_t\rangle \geq v_t^\top v^* > 0.$$

Thus, $0 < v_{t+1}^\top v^* \leq 1$. We then have

$$\begin{aligned}\|\widetilde{v}_{t+1} - v^*\|_2^2 &= 1 + \|\widetilde{v}_{t+1}\|_2^2 - 2\|\widetilde{v}_{t+1}\|_2 v_{t+1}^\top v^* \\ &\geq 1 + 1 - 2w_{t+1}^\top w^* = \|v_{t+1} - v^*\|_2^2.\end{aligned}$$

Then the distance between $\widetilde{w}_{t+1}$ and w^* is as follows.

$$\begin{aligned}\|v_{t+1} - v^*\|_2^2 \leq \|\widetilde{v}_{t+1} - v^*\|_2^2 &= \|w_t - \eta_w \nabla_w \mathcal{L}(a_t, w_t) - w^*\|_2^2 \\ &= \|v_t - v^*\|_2^2 + \|\eta_w \nabla_w \mathcal{L}(a_t, w_t)\|_2^2 - 2\langle -\nabla_w \mathcal{L}(w_t + \xi, a_t + \epsilon), v^* - v_t\rangle \\ &\leq (1 - \eta_w \frac{m}{4} + \eta_w^2 \frac{M^2}{4})\|v_t - v^*\|_2^2 \leq \|v_t - v^*\|_2^2,\end{aligned}$$

when $\eta_w \leq \frac{m}{M^2}$. Thus, $\phi_t \leq \phi_0 \leq \frac{5}{12}\pi$. We prove the first part.

We then prove the second part. Using the same expansion as in Lemma 9, we get

$$\begin{aligned}a_{t+1}^\top a^* &= \left(1 - \frac{\eta_a (\pi - 1)}{2\pi}\right) a_t^\top a^* + \frac{\eta_a \left(g\left(\phi_t\right) - 1\right)}{2\pi}\|a^*\|_2^2 + \frac{\eta_a}{2\pi}\left(\left(\mathbb{1}^\top a^*\right)^2 - \mathbb{1}^\top a^* \mathbb{1}^\top a_t\right) \\ &\geq \left(1 - \frac{\eta_a (\pi - 1)}{2\pi}\right) a_t^\top a^* + \eta_a \frac{C_2}{2\pi}\|a^*\|_2^2.\end{aligned}$$

Choose $\eta_a < \frac{2\pi}{\pi - 1}$, such that $1 - \frac{\eta_a(\pi-1)}{2\pi} < 1$. If $m \leq a_t^\top a^* \leq \frac{C_2}{\pi - 1}\|a^*\|_2^2$, the following inequality shows that $a_t^\top a^*$ increases over time.

$$a_{t+1}^\top a^* \geq \left(1 - \frac{\eta_a (\pi - 1)}{2\pi}\right) a_t^\top a^* + \eta_a \frac{C_2}{2\pi}\|a^*\|_2^2 \geq a_t^\top a^* \geq m.$$

If $a_t^\top a^* \geq \frac{C_2}{\pi - 1}\|a^*\|_2^2$, we show that this inequality holds for all t.

$$\begin{aligned}a_{t+1}^\top a^* &\geq \left(1 - \frac{\eta_a (\pi - 1)}{2\pi}\right) a_t^\top a^* + \eta_a \frac{C_2}{2\pi}\|a^*\|_2^2, \\ &\geq \left(1 - \frac{\eta_a (\pi - 1)}{2\pi}\right) \frac{C_2}{\pi - 1}\|a^*\|_2^2 + \eta_a \frac{C_2}{2\pi}\|a^*\|_2^2 = \frac{C_2}{\pi - 1}\|a^*\|_2^2\end{aligned}$$

Combine these two cases together, we have $a_{t+1}^\top a^* \geq \min\{m, \frac{C_2}{\pi - 1}\|a^*\|_2^2\} = m$. The other side follows similar lines in Lemma 9. Here, we omit the proof. □

B.4 Proof of Theorem 13

B.4.1 Proof of Lemma 12

Proof. Note that

$$\|w_t - w^*\|_2^2 \le \delta \iff \cos(\phi_t) \ge 1 - \frac{\delta}{2}.$$

Moreover, we can bound $g(\phi_t)$ as follows

$$\pi \ge g(\phi_t) = (\pi - \phi_t)\cos\phi_t + \sin\phi_t \ge \left(1 - \frac{\delta}{2}\right)\pi = \pi - \frac{\delta}{2}\pi.$$

Thus we have the partial dissipativity of $\nabla_a \mathcal{L}$.

$$\begin{aligned}
\langle -\nabla_a L(w,a), a^* - a\rangle &= \frac{1}{2\pi}\left(\mathbb{1}^\top a - \mathbb{1}^\top a^*\right)^2 + \frac{1}{2\pi}\left((\pi-1)a - (g(\phi)-1)a^*\right)^\top (a - a^*) \\
&= \frac{1}{2\pi}\left(\mathbb{1}^\top a - \mathbb{1}^\top a^*\right)^2 + \frac{1}{2\pi}(\pi - g(\phi))a^{*\top}(a - a^*) + \frac{\pi-1}{2\pi}\|a - a^*\|_2^2 \\
&\ge \frac{\pi-1}{2\pi}\|a - a^*\|_2^2 - \delta/5.
\end{aligned}$$

□

B.4.2 Proof of Lemma 14

Proof. First, we bound the norm of the gradient as follows

$$\|\nabla_w L(w,a)\|_2^2 = \frac{\left(a^\top a^*\right)^2(\pi-\phi)^2}{4\pi^2} v^{*\top}\left(I - vv^\top\right)v^* \le \frac{M^2}{4}(I - v^\top v^*)(I + v^\top v^*) \le \frac{M^2}{4}\|v - v^*\|_2^2$$

We next show that $\|w_{t+1} - w^*\|_2^2 \le \|\widetilde{w}_{t+1} - w^*\|_2^2$. We first have the following inequality.

$$\|\widetilde{w}_{t+1}\|_2^2 = \|w_t\|_2^2 + \|\eta\nabla_w L(w_t, a_t)\|_2^2 \ge 1.$$

Since we have $w_{t+1}^\top w^* \le 1$, we show that $\|\widetilde{v}_{t+1} - v^*\|_2^2 \le \|v_{t+1} - v^*\|_2^2$.

$$\begin{aligned}
\|\widetilde{v}_{t+1} - v^*\|_2^2 &= 1 + \|\widetilde{w}_{t+1}\|_2^2 - 2\|\widetilde{w}_{t+1}\|_2 w_{t+1}^\top w^* \\
&\ge 1 + 1 - 2w_{t+1}^\top w^* = \|v_{t+1} - v^*\|_2^2.
\end{aligned}$$

Then the distance between $\widetilde{w}_{t+1}$ and w^* is as follows.

$$\begin{aligned}
\|v_{t+1} - v^*\|_2^2 \le \|\widetilde{v}_{t+1} - v^*\|_2^2 &= \|w_t - \eta\nabla_w\mathcal{L}(a_t, w_t) - w^*\|_2^2 \\
&= \|w_t - w^*\|_2^2 + \|\eta\nabla_w\mathcal{L}(a_t, w_t)\|_2^2 - 2\langle -\nabla_w\mathcal{L}(w+\xi, a+\epsilon), w^* - w\rangle \\
&\le (1 - \eta\frac{m}{4} + \eta^2\frac{M^2}{4})\|v_t - v^*\|_2^2.
\end{aligned}$$

So we have for any t,

$$\|v_t - v^*\|_2^2 \le (1 - \eta\frac{m}{4} + \eta^2\frac{M^2}{4})^t\|v_0 - v^*\|_2^2.$$

Thus, choose $\eta \le \frac{m}{2M^2} = \widetilde{O}(\frac{1}{k^2})$, and after $t \ge \tau_{21} = \frac{4}{m\eta}\log\frac{4}{\delta}$ iterations, we have

$$\|v_t - v^*\|_2^2 \le \delta,$$

which is equivalent to

$$\|w_t - w^*\|_2^2 \le \delta.$$

□

B.4.3 Proof of Lemma 15

Proof. The proof follows similar lines to that of Lemma 14. By the partial dissipativity of $\mathcal{L}_a$, we have

$$\begin{aligned}
\|a_{t+1} - a^*\|_2^2 &= \|a_t - a^*\|_2^2 - 2\langle -\eta\mathbb{E}_{\xi,\epsilon}\nabla_a L(w_t, a_t), a^* - a_t\rangle \\
&\quad + \|\eta\nabla_a L(w_t, a_t)\|_2^2 \\
&\le \left(1 - \eta\frac{\pi-1}{\pi} + \eta^2\frac{(k+\pi-1)^2}{\pi^2}\right)\|a_t - a^*\|_2^2 + 2\eta^2\delta^2/25 + 2\eta\delta/5 \\
&\le (1 - \lambda_2)\|a_t - a^*\|_2^2 + \frac{4}{5}\eta\delta \\
&\le (1 - \lambda_2)^{t+1}\|a_0 - a^*\|_2^2 + \frac{b_2}{\lambda_2}.
\end{aligned}$$

where $\lambda_2 = \eta\frac{\pi-1}{\pi} - \eta^2\frac{(k+\pi-1)^2}{\pi^2}$ and $b_2 = \frac{4}{5}\eta\delta$. Take $\eta \leq \frac{5\pi^2}{4(k+\pi-1)^2}$, and then $\lambda_2 \geq \frac{\eta}{4}$. When $t \geq \tau_{22} = \frac{4}{\eta}\log\frac{\|a_0-a^*\|_2^2}{\delta} = \widetilde{O}(\frac{1}{\eta}\log\frac{1}{\delta})$, we have

$$\|a_t - a^*\|_2^2 \leq 5\delta.$$

□

C Experimental Settings

The output weight a^* in the teacher network is chosen as in Table 2.

k	$(a^*)^\top$
16	$[\underbrace{1,\ldots,1}_{9},\underbrace{-1,\ldots,-1}_{7}]$
25	$[\underbrace{1,\ldots,1}_{14},\underbrace{-1,\ldots,-1}_{11}]$
36	$[\underbrace{1,\ldots,1}_{19},\underbrace{-1,\ldots,-1}_{16},0]$
49	$[\underbrace{1,\ldots,1}_{26},\underbrace{-1,\ldots,-1}_{22},0]$
64	$[\underbrace{1,\ldots,1}_{34},\underbrace{-1,\ldots,-1}_{30}]$
81	$[\underbrace{1,\ldots,1}_{43},\underbrace{-1,\ldots,-1}_{38}]$
100	$[\underbrace{1,\ldots,1}_{52},\underbrace{-1,\ldots,-1}_{47},0]$

Table 2: Output weight a^*.

The trajectories in Figure 3 are obtained with a initialized at

$$\begin{aligned} a_0 =&[-0.1268, -0.1590, -0.1071, -0.1594, -0.4670, 0.1563, 0.1894, -0.2390, -0.0602, \\ &-0.5047, 0.0325, -0.0886, 0.1514, -0.0883, -0.0243, 0.1198, -0.2805, 0.0024, \\ &-0.0855, 0.0742, -0.0976, -0.1768, 0.1207, 0.0049, 0.1809]. \end{aligned}$$