SYMMETRY-BREAKING CONVERGENCE ANALYSIS OF CERTAIN TWO-LAYERED NEURAL NETWORKS WITH RELU NONLINEARITY

Yuandong Tian
Facebook AI Research
yuandong@fb.com

ABSTRACT

In this paper, we use dynamical system to analyze the nonlinear weight dynamics of two-layered bias-free networks in the form of $g(\mathbf{x};\mathbf{w}) = \sum_{j=1}^{K} \sigma(\mathbf{w}_j^\intercal \mathbf{x})$, where $\sigma(\cdot)$ is ReLU nonlinearity. We assume that the input $\mathbf{x}$ follow Gaussian distribution. The network is trained using gradient descent to mimic the output of a teacher network of the same size with fixed parameters $\mathbf{w}^*$ using l_2 loss. We first show that when $K = 1$, the nonlinear dynamics can be written in close form, and converges to $\mathbf{w}^*$ with at least $(1-\epsilon)/2$ probability, if random weight initializations of proper standard derivation ($\sim 1/\sqrt{d}$) is used, verifying empirical practice [Glorot & Bengio (2010); He et al. (2015); LeCun et al. (2012)]. For networks with many ReLU nodes ($K \geq 2$), we apply our close form dynamics and prove that when the teacher parameters $\{\mathbf{w}_j^*\}_{j=1}^{K}$ forms orthonormal bases, (1) a symmetric weight initialization yields a convergence to a saddle point and (2) a certain symmetry-breaking weight initialization yields global convergence to $\mathbf{w}^*$ without local minima. To our knowledge, this is the first proof that shows global convergence in nonlinear neural network without unrealistic assumptions on the independence of ReLU activations. In addition, we also give a concise gradient update formulation for a multilayer ReLU network when it follows a teacher of the same size with l_2 loss. Simulations verify our theoretical analysis.

1 INTRODUCTION

Deep learning has made substantial progress in many applications, including Computer Vision [He et al. (2016); Simonyan & Zisserman (2015); Szegedy et al. (2015); Krizhevsky et al. (2012)], Natural Language Processing [Sutskever et al. (2014)] and Speech Recognition [Hinton et al. (2012)]. However, till now, how and why it works remains elusive due to a lack of theoretical understanding. First, how simple approaches like gradient descent can solve a very complicated non-convex optimization effectively. Second, how the deep models, especially deep convolutional models, achieve generalization power despite massive parameters.

In this paper, we focus on the first problem and use dynamical system to analyze the nonlinear gradient descent dynamics of certain two-layered nonlinear network in the following form:

$$g(\mathbf{x};\mathbf{w}) = \sum_{j=1}^{K} \sigma(\mathbf{w}_j^\intercal \mathbf{x}) \tag{1}$$

where $\sigma(x) = \max(x, 0)$ is the ReLU nonlinearity. We consider the following setting: a student network learns the parameters that minimize the l_2 distance between its prediction and the supervision provided by the teacher network of the same size with a fixed set of parameters $\mathbf{w}^*$. We assume all inputs $\mathbf{x}$ to follow Gaussian distribution and thus the network is bias-free. Eqn. 1 is highly nonconvex and could contain exponential number of symmetrically equivalent solutions.

To analyze this, we first derive novel and concise gradient update rules for multilayer ReLU networks (See Lemma 2.1) in the teacher-student setting under l_2 loss. Then for $K = 1$, we prove that the nonlinear gradient dynamics of Eqn. 1 has a close form and converges to $\mathbf{w}^*$ with at least $(1 -$

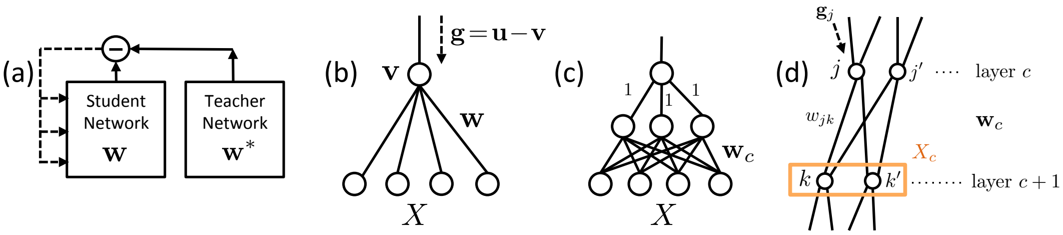

Figure 1: **(a)** We consider the student and teacher network as nonlinear neural networks with ReLU nonlinearity. The student network updates its weight $\mathbf{w}$ from the output of the teacher, whose weights $\mathbf{w}^*$ are fixed. **(b)-(c)** The network structure we consider in $K = 1$ and $K \geq 2$ cases. **(d)** Notations used in multilayer ReLU gradient update rule (Sec. 2.2)

$\epsilon)/2$ probability, if initialized randomly with standard derivation on the order of $1/\sqrt{d}$, verifying commonly used initialization techniques [Glorot & Bengio (2010); He et al. (2015); LeCun et al. (2012)],. When $K \geq 2$, we prove that when the teacher parameters $\{\mathbf{w}_j\}_{j=1}^K$ form orthonormal bases, **(1)** a symmetric initialization of a student network gets stuck at a saddle point and **(2)** under a certain symmetric breaking weight initialization, the dynamics converges to $\mathbf{w}^*$, without getting stuck into any local minima. Note that in both cases, the initialization can be arbitrarily close to the origin for a fixed $\|\mathbf{w}^*\|$, showing that such a convergence behavior is beyond the local convex structure at $\mathbf{w}^*$. To our knowledge, this is the first proof of its kind.

Previous works also use dynamical system to analyze deep neural networks. [Saxe et al. (2013)] analyzes the dynamics of multilayer linear network, and [Kawaguchi (2016)] shows every local minima is global for multilinear network. Very little theoretical work has been done to analyze the dynamics of *nonlinear* networks, especially deep ones. [Mei et al. (2016)] shows the global convergence when $K = 1$ with activation function $\sigma(x)$ when its derivatives σ', σ'', σ''' are bounded and $\sigma' > 0$. Similar to our approach, [Saad & Solla (1996)] also uses the student-teacher setting and analyzes the dynamics of student network when the teacher's parameters $\mathbf{w}^*$ forms a orthonomal bases; however, it uses $\sigma(x) = \text{erf}(x)$ as the nonlinearity and only analyzes the local behaviors of the two critical points (the saddle point in symmetric initializations, and $\mathbf{w}^*$). In contrast, we prove the global convergence behavior in certain symmetry-breaking cases.

Many previous works analyze nonlinear network based on the assumption of *independent activations*: the activations of ReLU (or other nonlinear) nodes are independent of the input and/or mutually independent. For example, [Choromanska et al. (2015a;b)] relate the nonlinear ReLU network with spin-glass models when several assumptions hold, including the assumption of independent activations (**A1p** and **A5u**). [Kawaguchi (2016)] proves that every local minimum in nonlinear network is global based on similar assumptions. [Soudry & Carmon (2016)] shows the global optimality of the local minimum in a two-layered ReLU network, by assuming small sample size and applying independent multiplicative Bernoulli noise on the activations. In practice, the activations are highly dependent due to their common input. Ignoring such dependency also misses important behaviors, and may lead to misleading conclusions. In this paper, *no assumption of independent activation is made*. For sigmoid activation, [Fukumizu & Amari (2000)] gives quite complicated conditions for a local minimum to be global when adding a new node to a two-layered network. [Janzamin et al. (2015)] gives guarantees on recovering the parameters of a 2-layered neural network learnt with tensor decomposition. In comparison, we analyze ReLU networks trained with gradient descent, which is a more popular setting in practice.

The paper is organized as follows. Sec. 2 introduces the basic formulation and some interesting novel properties of ReLU in multilayered ReLU networks. Sec. 3 and Sec. 4 then analyze the two-layered model Eqn. 1 for $K = 1$ and $K \geq 2$, respectively. Sec. 5 shows that simulation results are consistent with theoretical analysis. Finally Sec. 7 gives detailed proofs for all theorems.

2 Preliminary

2.1 Notation

Denote X as a N-by-d input data matrix and $\mathbf{w}^*$ is the parameter of the teacher network with desired N-by-1 output $\mathbf{u} = g(X; \mathbf{w}^*)$. Now suppose we have an estimator $\mathbf{w}$ and the estimated output $\mathbf{v} = g(X; \mathbf{w})$. We want to know with l_2 loss $E(\mathbf{w}) = \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|^2 = \frac{1}{2}\|\mathbf{u} - g(X; \mathbf{w})\|^2$, whether gradient descent will converge to the desired solution $\mathbf{w}^*$.

The gradient descent update is $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta\Delta\mathbf{w}^{(t)}$, where $\Delta\mathbf{w}^{(t)} \equiv -\nabla E(\mathbf{w}^{(t)})$. If we let $\eta \to 0$, then the update rule becomes a first-order differential equation $\mathrm{d}\mathbf{w}/\mathrm{d}t = -\nabla E(\mathbf{w})$, or more concisely, $\dot{\mathbf{w}} = -\nabla E(\mathbf{w})$. In this case, $\dot{E} = \nabla E(\mathbf{w})^\intercal \dot{\mathbf{w}} = -\|\nabla E(\mathbf{w})\|^2 \le 0$, i.e., the function value E is nonincreasing over time. The key is to check whether there exist other critical points $\mathbf{w} \neq \mathbf{w}^*$ so that $\nabla E(\mathbf{w}) = 0$.

In our analysis, we assume entries of input X follow Gaussian distribution. In this situation, the gradient is a random variable and $\Delta\mathbf{w} = -\mathbb{E}\left[\nabla E(\mathbf{w})\right]$. The expected $\mathbb{E}\left[E(\mathbf{w})\right]$ is also nonincreasing no matter whether we follow the expected gradient or the gradient itself, because

$$\mathbb{E}\left[\dot{E}\right] = -\mathbb{E}\left[\nabla E(\mathbf{w})^\intercal \nabla E(\mathbf{w})\right] \le -\mathbb{E}\left[\nabla E(\mathbf{w})\right]^\intercal \mathbb{E}\left[\nabla E(\mathbf{w})\right] \le 0 \tag{2}$$

Therefore, we analyze the behavior of expected gradient $\mathbb{E}\left[\nabla E(\mathbf{w})\right]$ rather than $\nabla E(\mathbf{w})$.

2.2 Properties of ReLU

In this paper, we discover a few useful properties of ReLU that make our analysis much simpler. Denote $D = D(\mathbf{w}) = \mathrm{diag}(X\mathbf{w} > 0)$ as a N-by-N diagonal matrix. The l-th diagnonal element of D is a binary variable showing whether the neuron is on for sample l. Using this notation, we could write $\sigma(X\mathbf{w}) = DX\mathbf{w}$. Note that D only depends on the direction of $\mathbf{w}$ but not its magnitude.

Note that for ReLU, D is also "tranparent" on derivatives. For example, the Jacobian $J_\mathbf{w}[\sigma(X\mathbf{w})] = \sigma'(X\mathbf{w})X = DX$ at differentiable regions. This gives a very concise rule for gradient descent in ReLU network: suppose we have negative gradient inflow vector $\mathbf{g}$ (of dimension N-by-1) on the current ReLU node with weights $\mathbf{w}$, then we can simply write the update $\Delta\mathbf{w}$ as:

$$\Delta\mathbf{w} = J_\mathbf{w}[\sigma(X\mathbf{w})]^\intercal \mathbf{g} = X^\intercal D\mathbf{g} \tag{3}$$

This can be easily applied to multilayer ReLU network. Denote $j \in [c]$ if node j is in layer c, d_c as the width of layer c, and $\mathbf{u}_j$ and $\mathbf{v}_j$ as the output of teacher network and student network, respectively. A simple deduction yields the following lemma:

Lemma 2.1 *For neural network with ReLU nonlinearity and using l_2 loss to match with a teacher network of the same size, the negative gradient inflow $\mathbf{g}_j$ for node j at layer c has the following form:*

$$\mathbf{g}_j = L_j \sum_{j'} (L^*_{j'}\mathbf{u}_{j'} - L_{j'}\mathbf{v}_{j'}) \tag{4}$$

where L_j and L^*_j are N-by-N diagonal matrices. For any $k \in [c+1]$, $L_k = \sum_{j\in[c]} w_{jk} D_j L_j$ and similarly for L^*_k. For the first layer, $L = L^* = I$.

The intuition here is to start from $\mathbf{g} = \mathbf{u} - \mathbf{v}$ (true for l_2 loss) at the top layer and use induction. With this formulation, we could write the finite dynamics for $\mathbf{w}_c$ (all parameters in layer c). Denote the N-by-$d_{c+1}d_c$ matrix $R_c = [L_j D_j]_{j\in[c]} X_c$ and $R^*_c = [L^*_j D^*_j]_{j\in[c]} X^*_c$. Using gradient descent rules:

$$\begin{aligned}\Delta\mathbf{w}_j &= X_c^\intercal D_j \mathbf{g}_j = X_c^\intercal D_j L_j \left(\sum_{j'} L^*_{j'} D^*_{j'} X^*_c \mathbf{w}^*_{j'} - \sum_{j'} L_{j'} D_{j'} X_c \mathbf{w}_{j'}\right) &(5)\\ &= X_c^\intercal D_j L_j \left(R^*_c \mathbf{w}^*_c - R_c \mathbf{w}_c\right) &(6)\end{aligned}$$

Therefore we have:

$$\Delta\mathbf{w}_c = R_c^\intercal \left(R^*_c \mathbf{w}^*_c - R_c \mathbf{w}_c\right) \tag{7}$$

3 Single ReLU Case

Let's start with the simplest case where there is only one ReLU node, $K = 1$. At iteration t, following Eqn. 3, the gradient update rule is:

$$\Delta\mathbf{w}^{(t)} = X^\intercal D^{(t)} \mathbf{g}^{(t)} = X^\intercal D^{(t)} (D^* X \mathbf{w}^* - D^{(t)} X \mathbf{w}^{(t)}) \tag{8}$$

Note here how the notation of $D^{(t)}$ comes into play (and $D^{(t)} D^{(t)} = D^{(t)}$). Indeed, when the neuron is cut off at sample l, then $(D^{(t)})_{ll}$ is zero and will block the corresponding gradient component.

Linear case. In this situation $D^{(t)} = D^* = I$ (no gating in either forward or backward propagation) and:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \frac{\eta}{N} X^\intercal X(\mathbf{w}^* - \mathbf{w}^{(t)}) \tag{9}$$

where η/N is the learning rate. When it is sufficiently small so that the spectral radius $\rho(I - \frac{\eta}{N} X^\intercal X) < 1$, $\mathbf{w}^{(t+1)}$ will converge to $\mathbf{w}^*$ when $t \to +\infty$. Note that this convergence is guaranteed for any initial condition $\mathbf{w}^{(1)}$, if $X^\intercal X$ is full rank with suitable η. This is consistent with its convex nature. If entries of X follow i.i.d Gaussian distribution, then $\mathbb{E}\left[\frac{1}{N} X^\intercal X\right] = I$ and the condition satisfies.

Nonlinear (ReLU) case. In this case, $\Delta\mathbf{w} = X^\intercal D(D^* X\mathbf{w}^* - DX\mathbf{w})$ in which D is a function of $\mathbf{w}$. Intuitively, this term goes to zero when $\mathbf{w} \to \mathbf{w}^*$, and should be approximated to be $\frac{N}{2}(\mathbf{w}^* - \mathbf{w})$ in the i.i.d Gaussian case, since roughly half of the samples are blocked. However, once we make such approximation, we lost the nonlinear behavior of the network and would draw the wrong conclusion of global convergence.

Then how should we analyze it? Notice that in $\Delta\mathbf{w}$, both of the two terms have the form $F(\mathbf{e}, \mathbf{w}) = X^\intercal D(\mathbf{e}) D(\mathbf{w}) X\mathbf{w}$. Using this form, $\mathbb{E}[\Delta\mathbf{w}] = \mathbb{E}[F(\mathbf{w}/\|\mathbf{w}\|, \mathbf{w}^*)] - \mathbb{E}[F(\mathbf{w}/\|\mathbf{w}\|, \mathbf{w})]$. Here $\mathbf{e}$ is a unit vector called the "projected" weight. In the following, we will show that $\mathbb{E}[F(\mathbf{e}, \mathbf{w})]$ has the following close form under i.i.d Gaussian assumption on X:

Lemma 3.1 *Denote $F(\mathbf{e}, \mathbf{w}) = X^\intercal D(\mathbf{e}) D(\mathbf{w}) X\mathbf{w}$ where $\mathbf{e}$ is a unit vector, $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^\intercal$ is N-by-d sample matrix and $D(\mathbf{w}) = \mathrm{diag}(X\mathbf{w} > 0)$ is a binary diagonal matrix. If $\mathbf{x}_i \sim N(0, I)$ and are i.i.d (and thus bias-free), then:*

$$\mathbb{E}[F(\mathbf{e}, \mathbf{w})] = \frac{N}{2\pi}\left[(\pi - \theta)\mathbf{w} + \|\mathbf{w}\| \sin\theta \mathbf{e}\right] \tag{10}$$

where $\theta = \angle(\mathbf{e}, \mathbf{w}) \in [0, \pi]$ is the angle between $\mathbf{e}$ and $\mathbf{w}$.

Note that the expectation analysis smooths out the non-differentiable property of ReLU, leaving only one singularity at $\mathbf{e} = 0$. The intuition is that expectation analysis involves an integration over the data distribution. With simple algebraic manipulation, $\mathbb{E}[\Delta\mathbf{w}]$ takes the following closed form:

$$\mathbb{E}[\Delta\mathbf{w}] = \frac{N}{2}(\mathbf{w}^* - \mathbf{w}) + \frac{N}{2\pi}(\alpha \sin\theta \mathbf{w} - \theta \mathbf{w}^*) \tag{11}$$

where $\alpha = \|\mathbf{w}^*\|/\|\mathbf{w}\|$ and $\theta \in [0, \pi]$ is the angle between $\mathbf{w}$ and $\mathbf{w}^*$. The first term is expected while the last two terms show the nonlinear behavior. Using Lyapunov's method, we show that the dynamics (if treated continuously) converges to $\mathbf{w}^*$ when $\mathbf{w}^{(1)} \in \Omega = \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}^*\| < \|\mathbf{w}^*\|\}$:

Lemma 3.2 *When $\mathbf{w}^{(1)} \in \Omega = \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}^*\| < \|\mathbf{w}^*\|\}$, following the dynamics of Eqn. 11, the Lyapunov function $V(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{w}^*\|^2$ has $\dot{V} < 0$ and the system is asymptotically stable and thus $\mathbf{w}^{(t)} \to \mathbf{w}^*$ when $t \to +\infty$.*

See Appendix for the proof. The intuition is to represent V as a 2-by-2 bilinear form of vector $[\|\mathbf{w}\|, \|\mathbf{w}^*\|]$, and the bilinear coefficient matrix is positive definite. One question arises: will the same approach show the dynamics converges when the initial conditions lie outside the region Ω, in particular for any region that includes the origin? The answer is probably no. Note that $\mathbf{w} = 0$ is a singularity in which $\Delta\mathbf{w}$ is not continuous (if approaching from different directions towards $\mathbf{w} = 0$, $\Delta\mathbf{w}$ is different). It is due to the fact that ReLU function is not differentiable at the origin. We could remove this singularity by "smoothing out" ReLU around the origin. This will yield $\Delta\mathbf{w} \to 0$ when $\mathbf{w} \to 0$. In this case, $\dot{V}(0) = 0$ so Lyapunov method could only tell that the dynamics is stable but not convergent. Note that for ReLU activation, $\sigma'(x) = 0$ for certain negative x even after a local smoothing, so the global convergence claim in [Mei et al. (2016)] for l_2 loss does not apply.

Random Initialization. Then we study how to sample $\mathbf{w}^{(1)}$ so that $\mathbf{w}^{(1)} \in \Omega$. We would like to sample within Ω, but we don't know where is $\mathbf{w}^*$. Sampling around origin with big radius $r \geq 2\|\mathbf{w}^*\|$ is inefficient in particular in high-dimensional space. This is because when the sample is uniform, the probability of hitting the ball is proportional to $(r/\|\mathbf{w}^*\|)^d \leq 2^{-d}$, which is exponentially small.

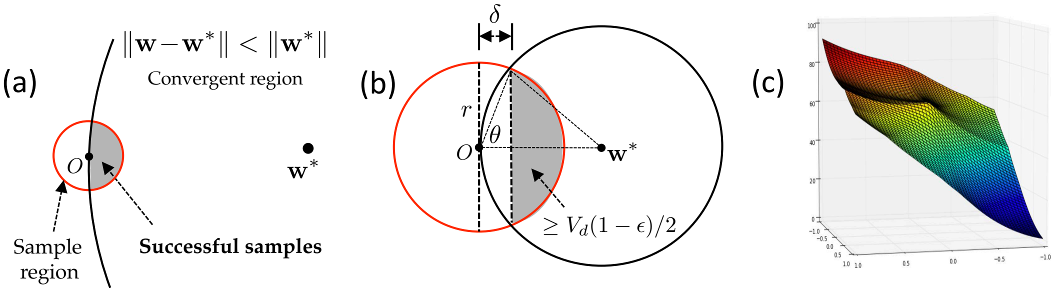

Figure 2: **(a)** Sampling strategy to maximize the probability of convergence. **(b)** Relationship between sampling range r and desired probability of success $(1-\epsilon)/2$. **(c)** Geometry of $K = 1$ 2D case. There is a singularity at the origin. Initialization with random weights around the origin has decent probability to converge to $\mathbf{w}^*$.

A better idea is to sample around the origin with very small radius (but not at $\mathbf{w} = 0$), so that the convergent hypersphere behaves like a hyperplane near the origin, and thus almost half of the samples is useful (Fig. 2(a)), as shown in the following theorem:

Theorem 3.3 *The dynamics in Eqn. 11 converges to* $\mathbf{w}^*$ *with probability at least* $(1-\epsilon)/2$*, if the initial value* $\mathbf{w}^{(1)}$ *is sampled uniformly from* $B_r = \{\mathbf{w} : \|\mathbf{w}\| \leq r\}$ *with* $r \leq \epsilon\sqrt{\frac{2\pi}{d+1}}\|\mathbf{w}^*\|$.

The intution here is to lower-bound the probability of the shaded area (Fig. 2(b)). From the proof, the conclusion could be made stronger to show $r \sim 1/\sqrt{d}$, consistent with common initialization techniques [Glorot & Bengio (2010); He et al. (2015); LeCun et al. (2012)]. Fig. 2(c) shows an example in the 2D case, in which there is a singularity at the origin, and sampling towards $\mathbf{w}^*$ yields the convergence. This is consistent with the analysis above.

4 Multiple ReLUs Case

Now we are ready to analyze the network $g(\mathbf{x}) = \sum_{j=1}^K \sigma(\mathbf{w}_j^\intercal \mathbf{x})$ for $K \geq 2$ (Fig. 1(c)). Theoretical analysis of such networks is also the main topic in many previous works [Saad & Solla (1996); Soudry & Carmon (2016); Fukumizu & Amari (2000)]. In this case, $L_j = L_j^* = I$ for $1 \leq j \leq K$. Then we have the following nonlinear dynamics from Eqn. 7:

$$\Delta\mathbf{w}_j = \sum_{j'=1}^K f(\mathbf{w}_j, \mathbf{w}_{j'}, \mathbf{w}^*_{j'}) \tag{12}$$

where $f = F(\mathbf{w}_j/\|\mathbf{w}_j\|, \mathbf{w}^*_{j'}) - F(\mathbf{w}_j/\|\mathbf{w}_j\|, \mathbf{w}_{j'})$. Therefore, using Eqn. 10, its expectation is:

$$\frac{2\pi}{N}\mathbb{E}\left[f(\mathbf{w}_j, \mathbf{w}_{j'}, \mathbf{w}^*_{j'})\right] = (\pi - \theta_j^{*j'})\mathbf{w}^*_{j'} - (\pi - \theta_j^{j'})\mathbf{w}_{j'} + \left(\frac{\|\mathbf{w}^*_{j'}\|}{\|\mathbf{w}_j\|}\sin\theta_j^{*j'} - \frac{\|\mathbf{w}_{j'}\|}{\|\mathbf{w}_j\|}\sin\theta_j^{j'}\right)\mathbf{w}_j \tag{13}$$

where $\theta_j^{*j'} \equiv \angle(\mathbf{w}_j, \mathbf{w}^*_{j'})$ and $\theta_j^{j'} \equiv \angle(\mathbf{w}_j, \mathbf{w}_{j'})$.

Eqn. 12 (and its expected version) gives very complicated nonlinear dynamics and could be hard to solve in general. Unlike $K = 1$, a similar approach with Lyaponov function does not yield a decisive conclusion. However, if we consider the symmetric case: $\mathbf{w}_j = P_j\mathbf{w}$ and $\mathbf{w}^*_j = P_j\mathbf{w}^*$ where P_j is a cyclic permutation matrix that maps index $j'+1$ to $(j'+j \bmod K)+1$ (and P_1 is the identity matrix), then RHS of the expected version of Eqn. 12 can be simplified as follows:

$$\begin{aligned}
\mathbb{E}\left[\Delta\mathbf{w}_j\right] &= \sum_{j'}\mathbb{E}\left[f(\mathbf{w}_j, \mathbf{w}_{j'}, \mathbf{w}^*_{j'})\right] = \sum_{j'}\mathbb{E}\left[f(P_j\mathbf{w}, P_{j'}\mathbf{w}, P_{j'}\mathbf{w}^*)\right] \\
&= \sum_{j''}\mathbb{E}\left[f(P_j\mathbf{w}, P_jP_{j''}\mathbf{w}, P_jP_{j''}\mathbf{w}^*)\right] \quad (\{P_j\}_{j=1}^K \text{ is a group}) \\
&= P_j\sum_{j''}\mathbb{E}\left[f(\mathbf{w}, P_{j''}\mathbf{w}, P_{j''}\mathbf{w}^*)\right] \quad (\|P\mathbf{w}_1\| = \|\mathbf{w}_1\|,\ \angle(P\mathbf{w}_1, P\mathbf{w}_2) = \angle(\mathbf{w}_1, \mathbf{w}_2)) \\
&= P_j\mathbb{E}\left[\Delta\mathbf{w}_1\right]
\end{aligned} \tag{14}$$

which means that if all $\mathbf{w}_j$ and $\mathbf{w}^*_j$ are symmetric under the action of cyclic group, so does their expected gradient. Therefore, the trajectory $\{\mathbf{w}^{(t)}\}$ keeps such cyclic structure. Instead of solving a system of K equations, we only need to solve one:

$$\mathbb{E}\left[\Delta\mathbf{w}\right] = \sum_{j=1}^{K} \mathbb{E}\left[f(\mathbf{w}, P_j\mathbf{w}, P_j\mathbf{w}^*)\right] \tag{15}$$

Surprisingly, there is another layer of symmetry in Eqn. 15 when $\{\mathbf{w}^*_j\}$ forms an orthonomal basis ($\mathbf{w}^{*\intercal}_{j'}\mathbf{w}^*_j = \delta_{jj'}$). In this case, if we start with $\mathbf{w}^{(1)} = x\mathbf{w}^* + y\sum_{j\neq 1} P_j\mathbf{w}^*$ then we could show that the trajectory keeps this structure and Eqn. 15 can be further reduced into the following 2D nonlinear dynamics:

$$\begin{aligned}\frac{2\pi}{N}\mathbb{E}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix} &= -\left\{[(\pi-\phi)(x-1+(K-1)y)]\begin{bmatrix}1\\ 1\end{bmatrix} + \begin{bmatrix}\theta\\ \phi^*-\phi\end{bmatrix} + \phi\begin{bmatrix}x-1\\ y\end{bmatrix}\right\}\\ &+ \quad [(K-1)(\alpha\sin\phi^* - \sin\phi) + \alpha\sin\theta]\begin{bmatrix}x\\ y\end{bmatrix}\end{aligned} \tag{16}$$

Here the symmetrical factor ($\alpha \equiv \|\mathbf{w}^*_{j'}\|/\|\mathbf{w}_j\|, \theta \equiv \theta^{*j}_j, \phi \equiv \theta^{j'}_j, \phi^* \equiv \theta^{*j'}_j$) are defined as follows:

$$\alpha = (x^2 + (K-1)y^2)^{-1/2}, \quad \cos\theta = \alpha x, \quad \cos\phi^* = \alpha y, \quad \cos\phi = \alpha^2(2xy + (K-2)y^2) \tag{17}$$

For this 2D dynamics, we thus have the following theorem:

Theorem 4.1 *For any $K \geq 2$, the 2D dynamics (Eqn. 16) shows the following behaviors:*

(1) ***Symmetric case.*** *If the initial condition $x^{(1)} = y^{(1)} \in (0, 1]$, then the dynamics reduces to 1D and converges to a saddle point $x = y = \frac{1}{\pi K}(\sqrt{K-1} - \arccos(1/\sqrt{K}) + \pi)$.*

(2) ***Symmetry-Breaking.*** *If $(x^{(1)}, y^{(1)}) \in \Omega = \{x \in (0, 1], y \in [0, 1], x > y\}$, then dynamics always converges to $(x, y) = (1, 0)$.*

From $(x^{(t)}, y^{(t)})$ we could recover $\mathbf{w}^{(t)}_j = x^{(t)}\mathbf{w}^*_j + y^{(t)}\sum_{j'\neq j}\mathbf{w}^*_{j'}$. Obviously, a convergence of Eqn. 16 to $(1, 0)$ means Eqn. 12 converges to $\{\mathbf{w}^*_j\}$, i.e, the teacher parameters are recovered:

Corollary 4.2 *For a bias-free two-layered ReLU network $g(\mathbf{x}; \mathbf{w}) = \sum_j \sigma(\mathbf{w}^\intercal_j\mathbf{x})$ that takes Gaussian i.i.d inputs (Fig. 1), if the teacher's parameters $\{\mathbf{w}^*_j\}$ form orthogonal bases, then when the student parameters is initialized in the form of $\mathbf{w}^{(1)}_j = x^{(1)}\mathbf{w}^*_j + y^{(1)}\sum_{j'\neq j}\mathbf{w}^*_{j'}$ where $(x^{(1)}, y^{(1)}) \in \Omega = \{x \in (0, 1], y \in [0, 1], x > y\}$, then the dynamics (Eqn. 12) converges to $\{\mathbf{w}^*_j\}$ without being trapped into local minima.*

When symmetry is broken, since the closure of Ω includes the origin, there exists a path starting at arbitrarily small neighborhood of origin to $\mathbf{w}^*$, regardless of how large $\|\mathbf{w}^*\|$ is. In contrast to traditional convex analysis that only gives the local parameter-dependent convergence basin around $\mathbf{w}^*_j$, here we obtain a convergence basin that is parameter-independent. In comparison, [Saad & Solla (1996)] uses a different activation function ($\sigma(x) = \mathrm{erf}(x)$) and only analyzes local behaviors near the two fixed points (the symmetric saddle point and the teacher's weights $\mathbf{w}^*$), leaving symmetry breaking an empirical procedure. Here we show that it is possible to give global convergence analysis on certain symmetry breaking cases for two-layered ReLU network.

By symmetry, Corollary 4.1 immediately suggests that when $\mathbf{w}^{(1)} = y^{(1)}\sum_{j=1}^{K}\mathbf{w}^*_j + (x^{(1)} - y^{(1)})\mathbf{w}^*_{j'}$, then the dynamics will converge to $P_{j'}\mathbf{w}^*$. Since $x > y$ but can be arbitrarily close, a slighest preturbation on the symmetric solution $x = y$ leads to a different fixed point, which is a permutation of $\mathbf{w}^*$. This is very similar to Spontaneously Symmetric-Breaking (SSB) procedure in physics, in which a high energy state with full symmetry goes to a low energy state and only retains part of the symmetry. In this case, the energy is the objective function E, the high energy state is the initialization that is almost symmetrical but with small fluctuation, and the low energy state is the fixed point the dynamics converges into.

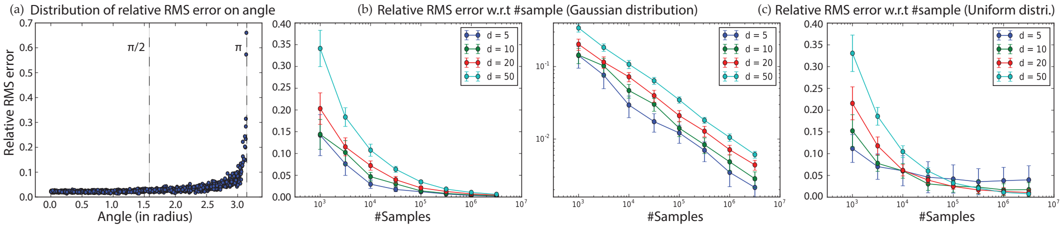

Figure 3: **(a)** Distribution of relative RMS error with respect to $\theta = \angle(\mathbf{w}, \mathbf{e})$. **(b)** Relative RMS error decreases with sample size, showing the asympototic behavior of the close form expression Eqn. 10. **(c)** Eqn. 10 also works well when the input data X are generated by other zero-mean distribution X, e.g., uniform distribution in $[-1/2, 1/2]$.

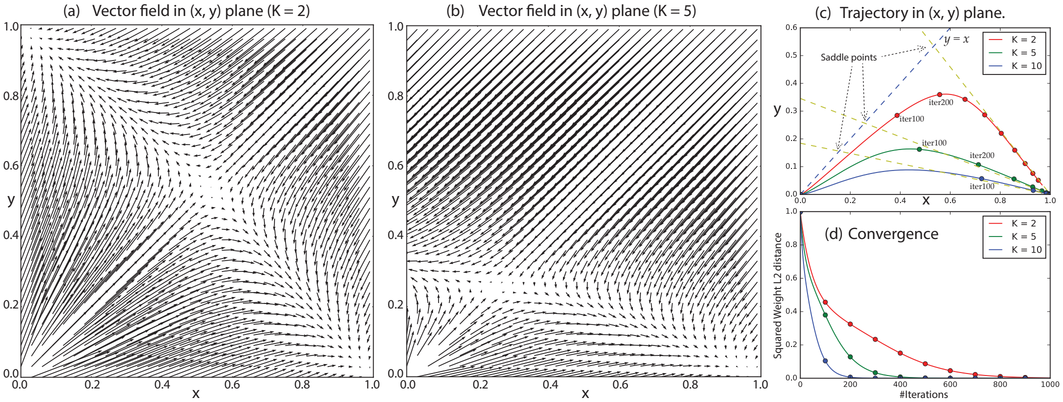

Figure 4: **(a)-(b)** Vector field in (x, y) plane following 2D dynamics (Eqn. 16) for $K = 2$ and $K = 5$. Saddle points are visible. The parameters of teacher's network are at $(1, 0)$. **(c)** Trajectory in (x, y) plane for $K = 2$, $K = 5$, and $K = 10$. All trajectories start from $(10^{-3}, 0)$. Even the starting point are aligned with $\mathbf{w}^*$, gradient descent dynamics takes detours. **(d)** Training curve. When K is larger the convergence is faster.

From the simulation shown in Fig. 4, we could see that gradient descent takes a detour to reach the desired solution $\mathbf{w}^*$, even when the initialization is aligned with $\mathbf{w}^*$. This is because in the first stage, all ReLU nodes receive the residue and try to explain the data in the same way (both x and y increases); when the "obvious" component has been explained away, then the residue changes its direction and pushes some ReLU nodes to explain other components as well (x increases but y decreases).

Empirically this path also converges to $\mathbf{w}^*$ under noise. We leave it a conjecture that the system converges in the presence of reasonably large noise. If this conjecture is true, then with high probability a random initialization stays in the convergence basin and converges to a *permutation* of $\mathbf{w}^*$. The reason is that a random initialization almost never gives ties. Without a tie, there exists one leading component which will dominate the convergence.

Conjecture 4.3 *When the initialization $\mathbf{w}^{(1)} = x^{(1)}\mathbf{w}_j^* + y^{(1)} \sum_{j' \neq j} \mathbf{w}_{j'}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is Gaussian noise and $(x^{(1)}, y^{(1)}) \in \Omega$, then the dynamics Eqn. 12 also converges to $\mathbf{w}^*$ without trapped into local minima.*

5 Simulation

5.1 Close form solution for one ReLU node

We verify our close form expression of $\mathbb{E}[F(\mathbf{e}, \mathbf{w})] = \mathbb{E}[X^\intercal D(\mathbf{e}) D(\mathbf{w}) X\mathbf{w}]$ (Eqn. 10) with simulation. We randomly pick $\mathbf{e}$ and $\mathbf{w}$ so that their angle $\angle(\mathbf{e}, \mathbf{w})$ is uniformly distributed in $[0, \pi]$. We prepare the input data X with standard Gaussian distribution and compare the close form solution $\mathbb{E}[F(\mathbf{e}, \mathbf{w})]$ with $F(\mathbf{e}, \mathbf{w})$, the actual data term in gradient descent without expectation. We use relative RMS error: $err = \|\mathbb{E}[F(\mathbf{e}, \mathbf{w})] - F(\mathbf{e}, \mathbf{w})\| / \|F(\mathbf{e}, \mathbf{w})\|$. As shown in Fig. 3(a), The error distribution on angles shows the properties of the close-form solution. For small θ, $D(\mathbf{w})$ and

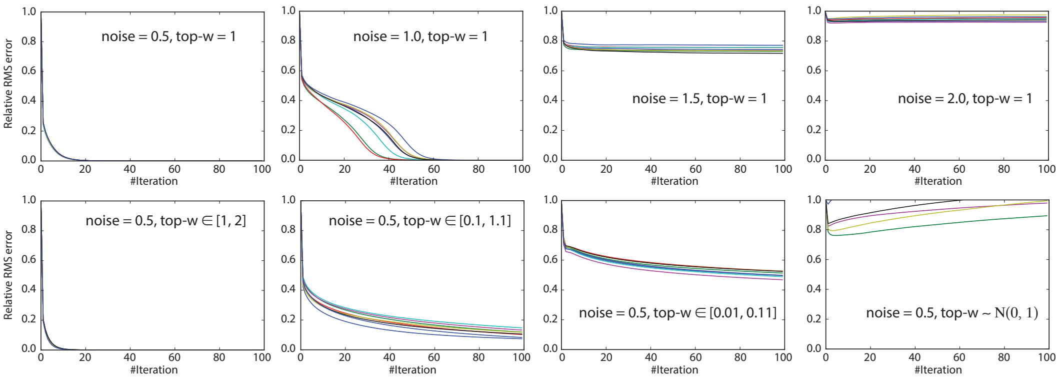

Figure 5: **Top row:** Convergence when the initial weights deviates from symmetric initialization: $\mathbf{w}^{(1)} = 10^{-3}\mathbf{w}^* + \epsilon$. Here $\epsilon \sim N(0, 10^{-3} * noise)$. The 2-layered network converges to $\mathbf{w}^*$ until very large noise is present. Both teacher and student networks use $g(\mathbf{x}) = \sum_{j=1}^{K} \sigma(\mathbf{w}_j^\intercal \mathbf{x})$. Each experiment has 8 runs. **Bottom row:** Convergence when we use $g_2(\mathbf{x}) = \sum_{j=1}^{K} a_j \sigma(\mathbf{w}_j^\intercal \mathbf{x})$. Here the top weights a_j is fixed at different numbers (rather than 1). Large positive a_j correponds to fast convergence. When a_j has positive/negative components, the network does not converge to $\mathbf{w}^*$.

$D(\mathbf{e})$ overlaps sufficiently, giving a reliable estimation for the gradient. When $\theta \to \pi$, $D(\mathbf{w})$ and $D(\mathbf{e})$ tend not to overlap, leaving very few data involved in the gradient computation. As a result, the variance grows. Note that all our analysis operate on $\theta \in [0, \pi/2]$ and is not affected by this behavior. In the following, angles are sampled from $[0, \pi/2]$.

Fig. 3(a) shows that the close form expression becomes more accurate with more samples. We also examine other zero-mean distributions of X, e.g., uniform distribution in $[-1/2, 1/2]$. As shown in Fig. 3(d), the close form expression still works for large d, showing that it could be quite general. Note that the error is computed up to a scaling constant, due to the difference in normalization constants among different distributions. We leave it to the future work to prove its usability for broader distributions.

5.2 Convergence for multiple ReLU nodes

Fig. 4(a) and (b) shows the 2D vector field given by the 2D dynamics (Eqn. 16) and Fig. 4(c) shows the 2D trajectory towards convergence to the teacher's parameters $\mathbf{w}^*$. Interestingly, even when we initialize the weights as $(10^{-3}, 0)$, aligning with $\mathbf{w}^*$, the gradient descent takes detours to reach the destination. One explanation is, at the beginning all nodes move similar direction trying to explain the data, once the data have been explained partly, specialization follows (y decreases).

Fig. 5 shows empirical convergence for $K \geq 2$, when the initialization deviates from symmetric initialization in Thm. 4.1. Unless the deviation is large, gradient descent converges to $\mathbf{w}^*$. We also check the convergence of a more general network $g_2(\mathbf{x}) = \sum_{j=1}^{K} a_j \sigma(\mathbf{w}_j^\intercal \mathbf{x})$. When $a_j > 0$ convergence follows; however, when some a_j is negative, the network does not converge to $\mathbf{w}^*$, even that the student network already knows the ground truth value of $\{a_j\}_{j=1}^K$.

6 Conclusion and Future Work

In this paper, we analyze the nonlinear dynamical behavior of certain two-layered bias-free ReLU networks in the form of $g(\mathbf{x}; \mathbf{w}) = \sum_{j=1}^{K} \sigma(\mathbf{w}_j^\intercal \mathbf{x})$, where $\sigma = \max(x, 0)$ is the ReLU node. We assume that the input $\mathbf{x}$ follows Gaussian distribution and the output is generated by a teacher network with parameters $\mathbf{w}^*$. In $K = 1$ we show a close-form nonlinear dynamics can be obtained and its convergence to $\mathbf{w}^*$ can be proven, if we sample the initialization properly. Such initialization is consistent with common practice [Glorot & Bengio (2010); He et al. (2015)] and is independent of the value of $\mathbf{w}^*$. For $K \geq 2$, when the teacher parameters $\{\mathbf{w}_j^*\}$ form a orthonormal bases, we prove that the trajectory from symmetric initialization is trapped into a saddle point, while certain symmetric breaking initialization converges to $\mathbf{w}^*$ without trapped into any local minima. Future work includes analysis of general cases (or symmetric case plus noise) for $K \geq 2$, and a generalization to multilayer ReLU (or other nonlinear) networks.

References

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015a.

Anna Choromanska, Yann LeCun, and Gérard Ben Arous. Open problem: The landscape of the loss surfaces of multilayer networks. In *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3*, volume 6, pp. 1756–1760, 2015b.

Kenji Fukumizu and Shun-ichi Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13(3):317–327, 2000.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pp. 249–256, 2010.

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*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Computer Vision anad Pattern Recognition (CVPR)*, 2016.

Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

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

Kenji Kawaguchi. Deep learning without poor local minima. *Advances in Neural Information Processing Systems*, 2016.

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

Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pp. 9–48. Springer, 2012.

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

David Saad and Sara A Solla. Dynamics of on-line gradient descent learning for multilayer neural networks. *Advances in Neural Information Processing Systems*, pp. 302–308, 1996.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015.

Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

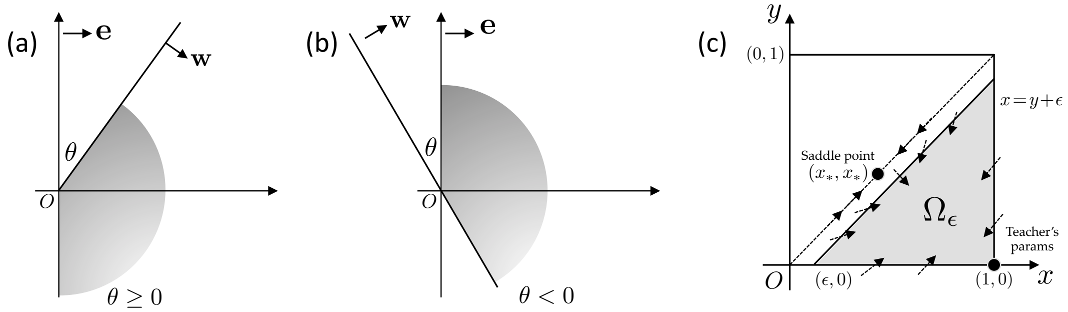

Figure 6: **(a)-(b)** Two cases in Lemma 7.2. **(c)** Convergence analysis in the symmetric two-layered case.

7 Appendix

Here we list all detailed proof for all the theorems.

7.1 Properties of ReLU networks

Lemma 7.1 *For neural network with ReLU nonlinearity and using l_2 loss to match with a teacher network of the same size, the negative gradient inflow $\mathbf{g}_j$ for node j at layer c has the following form:*

$$\mathbf{g}_j = L_j \sum_{j'} (L^*_{j'} \mathbf{u}_{j'} - L_{j'} \mathbf{v}_{j'}) \tag{18}$$

where L_j and L^*_j are N-by-N diagonal matrices. For any $k \in [c+1]$, $L_k = \sum_{j\in[c]} w_{jk} D_j L_j$ and similarly for L^*_k.

Proof We prove by induction on layer. For the first layer, there is only one node with $\mathbf{g} = \mathbf{u} - \mathbf{v}$, therefore $L_j = L_{j'} = I$. Suppose the condition holds for all node $j \in [c]$. Then for node $k \in [c+1]$, we have:

$$\begin{aligned}
\mathbf{g}_k &= \sum_j w_{jk} D_j \mathbf{g}_j = \sum_j w_{jk} D_j L_j \left(\sum_{j'} L^*_{j'} \mathbf{u}_{j'} - \sum_{j'} L_{j'} \mathbf{v}_{j'} \right) \\
&= \sum_j w_{jk} D_j L_j \left(\sum_{j'} L^*_{j'} \sum_{k'} D^*_{j'} w^*_{jk'} \mathbf{u}_{k'} - \sum_{j'} L_{j'} \sum_{k'} D_{j'} w_{jk'} \mathbf{v}_{k'} \right) \\
&= \sum_j w_{jk} D_j L_j \sum_{j'} L^*_{j'} D^*_{j'} \sum_{k'} w^*_{jk'} \mathbf{u}_{k'} - \sum_j w_{jk} D_j L_j \sum_{j'} L_{j'} D_{j'} \sum_{k'} w_{jk'} \mathbf{v}_{k'} \\
&= \sum_{k'} \left(\sum_j w_{jk} D_j L_j \right) \left(\sum_{j'} L^*_{j'} D^*_{j'} w^*_{jk'} \right) \mathbf{u}_{k'} - \sum_{k'} \left(\sum_j w_{jk} D_j L_j \right) \left(\sum_{j'} L_{j'} D_{j'} w_{jk'} \right) \mathbf{v}_{k'}
\end{aligned}$$

Setting $L_k = \sum_j w_{jk} D_j L_j$ and $L^*_k = \sum_j w^*_{jk} D^*_j L^*_j$ (both are diagonal matrices), we thus have:

$$\mathbf{g}_k = \sum_{k'} L_k L^*_{k'} \mathbf{u}_{k'} - L_k L_{k'} \mathbf{v}_{k'} = L_k \sum_{k'} L^*_{k'} \mathbf{u}_{k'} - L_{k'} \mathbf{v}_{k'} \tag{19}$$

∎

7.2 One ReLU Case

Lemma 7.2 *Suppose* $F(\mathbf{e},\mathbf{w}) = X^\intercal D(\mathbf{e})D(\mathbf{w})X\mathbf{w}$ *where* $\mathbf{e}$ *is a unit vector and* $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N]^\intercal$ *is* N*-by-*d *sample matrix. If* $\mathbf{x}_i \sim N(0, I)$ *and are i.i.d, then:*

$$\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right] = \frac{N}{2\pi}\left((\pi-\theta)\mathbf{w} + \|\mathbf{w}\|\sin\theta\mathbf{e}\right) \tag{20}$$

where $\theta \in [0, \pi]$ *is the angle between* $\mathbf{e}$ *and* $\mathbf{w}$.

Proof Note that F can be written in the following form:

$$F(\mathbf{e},\mathbf{w}) = \sum_{i:\mathbf{x}_i^\intercal\mathbf{e}\geq 0, \mathbf{x}_i^\intercal\mathbf{w}\geq 0} \mathbf{x}_i\mathbf{x}_i^\intercal\mathbf{w} \tag{21}$$

where $\mathbf{x}_i$ are samples so that $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^\intercal$. We set up the axes related to $\mathbf{e}$ and $\mathbf{w}$ as in Fig. 6, while the rest of the axis are prependicular to the plane. In this coordinate system, any vector $\mathbf{x} = [r\sin\phi, r\cos\phi, x_3, \ldots, x_d]$. We have an orthonomal set of bases: $\mathbf{e}$, $\mathbf{e}_\perp = -\frac{\mathbf{e}-\mathbf{w}/\|\mathbf{w}\|\cos\theta}{\sin\theta}$ (and any set of bases that span the rest of the space). Under the basis, the representation for $\mathbf{e}$ and $\mathbf{w}$ is $[1, \mathbf{0}_{d-1}]$ and $[\|\mathbf{w}\|\cos\theta, -\|\mathbf{w}\|\sin\theta, \mathbf{0}_{d-2}]$. Note that here $\theta \in (-\pi, \pi]$. The angle θ is positive when $\mathbf{e}$ "chases after" $\mathbf{w}$, and is otherwise negative.

Now we consider the quality $R(\phi_0) = \mathbb{E}\left[\frac{1}{N}\sum_{i:\phi_i\in[0,\phi_0]}\mathbf{x}_i\mathbf{x}_i^\intercal\right]$. If we take the expectation and use polar coordinate only in the first two dimensions, we have:

$$R(\phi_0) = \mathbb{E}\left[\frac{1}{N}\sum_{i:\phi_i\in[0,\phi_0]}\mathbf{x}_i\mathbf{x}_i^\intercal\right] = \mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\intercal|\phi_i\in[0,\phi_0]\right]\mathbb{P}\left[\phi_i\in[0,\phi_0]\right]$$

$$= \int_0^{+\infty}\iint_{-\infty}^{+\infty}\int_0^{\phi_0}\begin{bmatrix} r\sin\phi \\ r\cos\phi \\ \ldots \\ x_d\end{bmatrix}\begin{bmatrix} r\sin\phi & r\cos\phi & \ldots & x_d\end{bmatrix} p(r)p(\theta)\prod_{k=3}^d p(x_k) r\mathrm{d}r\mathrm{d}\phi\mathrm{d}x_3\ldots\mathrm{d}x_d$$

where $p(r) = e^{-r^2/2}$ and $p(\theta) = 1/2\pi$. Note that $R(\phi_0)$ is a d-by-d matrix. The first 2-by-2 block can be computed in close form (note that $\int_0^{+\infty} r^2 p(r)r\mathrm{d}r = 2$). Any off-diagonal element except for the first 2-by-2 block is zero due to symmetric property of i.i.d Gaussian variables. Any diagonal element outside the first 2-by-2 block will be $\mathbb{P}\left[\phi_i\in[0,\phi_0]\right] = \phi_0/2\pi$. Finally, we have:

$$R(\phi_0) = \mathbb{E}\left[\frac{1}{N}\sum_{i:\phi_i\in[0,\phi_0]}\mathbf{x}_i\mathbf{x}_i^\intercal\right] = \frac{1}{4\pi}\begin{bmatrix} 2\phi_0 - \sin 2\phi_0 & 1-\cos 2\phi_0 & \mathbf{0} \\ 1-\cos 2\phi_0 & 2\phi_0+\sin 2\phi_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 2\phi_0 I_{d-2}\end{bmatrix} \tag{22}$$

$$= \frac{\phi_0}{2\pi}I_d + \frac{1}{4\pi}\begin{bmatrix} -\sin 2\phi_0 & 1-\cos 2\phi_0 & \mathbf{0} \\ 1-\cos 2\phi_0 & \sin 2\phi_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 0\end{bmatrix} \tag{23}$$

With this equation, we could then compute $\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right]$. When $\theta \geq 0$, the condition $\{i : \mathbf{x}_i^\intercal\mathbf{e} \geq 0, \mathbf{x}_i^\intercal\mathbf{w} \geq 0\}$ is equivalent to $\{i : \phi_i \in [\theta, \pi]\}$ (Fig. 6(a)). Using $\mathbf{w} = [\|\mathbf{w}\|\cos\theta, -\|\mathbf{w}\|\sin\theta, \mathbf{0}_{d-2}]$ and we have:

$$\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right] = N\left(R(\pi) - R(\theta)\right)\mathbf{w} \tag{24}$$

$$= \frac{N}{4\pi}\left(2(\pi-\theta)\mathbf{w} - \|\mathbf{w}\|\begin{bmatrix} -\sin 2\theta & 1-\cos 2\theta & \mathbf{0} \\ 1-\cos 2\theta & \sin 2\theta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 0\end{bmatrix}\begin{bmatrix}\cos\theta \\ -\sin\theta \\ \mathbf{0}\end{bmatrix}\right) \tag{25}$$

$$= \frac{N}{2\pi}\left((\pi-\theta)\mathbf{w} + \|\mathbf{w}\|\begin{bmatrix}\sin\theta \\ \mathbf{0}\end{bmatrix}\right) \tag{26}$$

$$= \frac{N}{2\pi}\left((\pi-\theta)\mathbf{w} + \|\mathbf{w}\|\sin\theta\mathbf{e}\right) \tag{27}$$

For $\theta < 0$, the condition $\{i : \mathbf{x}_i^\intercal\mathbf{e} \geq 0, \mathbf{x}_i^\intercal\mathbf{w} \geq 0\}$ is equivalent to $\{i : \phi_i \in [0, \pi+\theta]\}$ (Fig. 6(b)), and similarly we get

$$\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right] = N\left(R(\pi+\theta) - R(0)\right)\mathbf{w} = \frac{N}{2\pi}\left((\pi+\theta)\mathbf{w} - \|\mathbf{w}\|\sin\theta\mathbf{e}\right) \tag{28}$$

Notice that by abuse of notation, the θ appears in Eqn. 20 is the absolute value and Eqn. 20 follows. ■

Lemma 7.3 *In the region $\|\mathbf{w}^{(1)} - \mathbf{w}^*\| < \|\mathbf{w}^*\|$, following the dynamics (Eqn. 11), the Lyapunov function $V(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{w}^*\|^2$ has $\dot{V} < 0$ and the system is asymptotically stable and thus $\mathbf{w}^{(t)} \to \mathbf{w}^*$ when $t \to +\infty$.*

Proof Denote that $\Omega = \{\mathbf{w} : \|\mathbf{w}^{(1)} - \mathbf{w}^*\| < \|\mathbf{w}^*\|\}$. Note that

$$\dot{V} = (\mathbf{w} - \mathbf{w}^*)^\intercal \Delta\mathbf{w} = -\mathbf{y}^\intercal M \mathbf{y} \tag{29}$$

where $\mathbf{y} = [\|\mathbf{w}^*\|, \|\mathbf{w}\|]^\intercal$ and M is the following 2-by-2 matrix:

$$M = \frac{1}{2}\begin{bmatrix} \sin 2\theta + 2\pi - 2\theta & -(2\pi - \theta)\cos\theta - \sin\theta \\ -(2\pi - \theta)\cos\theta - \sin\theta & 2\pi \end{bmatrix} \tag{30}$$

In the following we will show that M is positive definite when $\theta \in (0, \pi/2]$. It suffices to show that $M_{11} > 0$, $M_{22} > 0$ and $\det(M) > 0$. The first two are trivial, while the last one is:

$$\begin{aligned}
4\det(M) &= 2\pi(\sin 2\theta + 2\pi - 2\theta) - [(2\pi - \theta)\cos\theta + \sin\theta]^2 && (31)\\
&= 2\pi(\sin 2\theta + 2\pi - 2\theta) - \left[(2\pi - \theta)^2\cos^2\theta + (2\pi - \theta)\sin 2\theta + \sin^2\theta\right] && (32)\\
&= (4\pi^2 - 1)\sin^2\theta - 4\pi\theta + 4\pi\theta\cos^2\theta - \theta^2\cos^2\theta + \theta\sin 2\theta && (33)\\
&= (4\pi^2 - 4\pi\theta - 1)\sin^2\theta + \theta\cos\theta(2\sin\theta - \theta\cos\theta) && (34)
\end{aligned}$$

Note that $4\pi^2 - 4\pi\theta - 1 = 4\pi(\pi - \theta) - 1 > 0$ for $\theta \in [0, \pi/2]$, and $g(\theta) = \sin\theta - \theta\cos\theta \geq 0$ for $\theta \in [0, \pi/2]$ since $g(0) = 0$ and $g'(\theta) \geq 0$ in this region. Therefore, when $\theta \in (0, \pi/2]$, M is positive definite.

When $\theta = 0$, $M(\theta) = \pi[1, -1; -1, 1]$ and is semi-positive definite, with the null eigenvector being $\frac{\sqrt{2}}{2}[1, 1]$, i.e., $\|\mathbf{w}\| = \|\mathbf{w}^*\|$. However, along $\theta = 0$, the only $\mathbf{w}$ that satisfies $\|\mathbf{w}\| = \|\mathbf{w}^*\|$ is $\mathbf{w} = \mathbf{w}^*$. Therefore, $\dot{V} = -\mathbf{y}^\intercal M\mathbf{y} < 0$ in Ω. Note that although this region could be expanded to the entire open half-space $\mathcal{H} = \{\mathbf{w} : \mathbf{w}^\intercal\mathbf{w}^* > 0\}$, it is not straightforward to prove the convergence in $\mathcal{H}$, since the trajectory might go outside $\mathcal{H}$. On the other hand, Ω is the level set $V < \frac{1}{2}\|\mathbf{w}^*\|^2$ so the trajectory starting within Ω remains inside. ■

Theorem 7.4 *The dynamics in Eqn. 11 converges to $\mathbf{w}^*$ with probability at least $(1 - \epsilon)/2$, if the initial value $\mathbf{w}^{(1)}$ is sampled uniformly from $B_r = \{\mathbf{w} : \|\mathbf{w}\| \leq r\}$ with:*

$$r \leq \epsilon\sqrt{\frac{2\pi}{d+1}}\|\mathbf{w}^*\| \tag{35}$$

Proof Given a ball of radius r, we first compute the "gap" δ of sphere cap (Fig. 2(b)). First $\cos\theta = \frac{r}{2\|\mathbf{w}^*\|}$, so $\delta = r\cos\theta = \frac{r^2}{2\|\mathbf{w}^*\|}$. Then a sufficient condition for the probability argument to hold, is to ensure that the volume V_{shaded} of the shaded area is greater than $\frac{1-\epsilon}{2}V_d(r)$, where $V_d(r)$ is the volume of d-dimensional ball of radius r. Since $V_{\text{shaded}} \geq \frac{1}{2}V_d(r) - \delta V_{d-1}$, it suffices to have:

$$\frac{1}{2}V_d(r) - \delta V_{d-1} \geq \frac{1-\epsilon}{2}V_d(r) \tag{36}$$

which gives

$$\delta \leq \frac{\epsilon}{2}\frac{V_d}{V_{d-1}} \tag{37}$$

Using $\delta = \frac{r^2}{2\|\mathbf{w}^*\|}$ and $V_d(r) = V_d(1)r^d$, we thus have:

$$r \leq \epsilon\frac{V_d(1)}{V_{d-1}(1)}\|\mathbf{w}^*\| \tag{38}$$

where $V_d(1)$ is the volume of the unit ball. Since the volume of d-dimensional unit ball is

$$V_d(1) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \tag{39}$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\mathrm{d}t$. So we have

$$\frac{V_d(1)}{V_{d-1}(1)} = \sqrt{\pi}\frac{\Gamma(d/2+1/2)}{\Gamma(d/2+1)} \tag{40}$$

From Gautschi's Inequality

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+s)^{1-s} \qquad x > 0, 0 < s < 1 \tag{41}$$

with $s = 1/2$ and $x = d/2$ we have:

$$\left(\frac{d+1}{2}\right)^{-1/2} < \frac{\Gamma(d/2+1/2)}{\Gamma(d/2+1)} < \left(\frac{d}{2}\right)^{-1/2} \tag{42}$$

Therefore, it suffices to have

$$r \le \epsilon\sqrt{\frac{2\pi}{d+1}}\|\mathbf{w}^*\| \tag{43}$$

Note that this upper bound is tight when $\delta \to 0$ and $d \to +\infty$, since all inequality involved asymptotically becomes equal. ■

7.3 Two Layer Case

Lemma 7.5 *For ϕ^*, θ and ϕ defined in Eqn. 17:*

$$\begin{aligned} \alpha &\equiv (x^2 + (K-1)y^2)^{-1/2} && (44) \\ \cos\theta &\equiv \alpha x && (45) \\ \cos\phi^* &\equiv \alpha y && (46) \\ \cos\phi &\equiv \alpha^2(2xy + (K-2)y^2) && (47) \end{aligned}$$

we have the following relations in the triangular region $\Omega_{\epsilon_0} = \{(x, y) : x \ge 0, y \ge 0, x \ge y + \epsilon_0\}$ (Fig. 6(c)):

(1) $\phi, \phi^* \in [0, \pi/2]$ and $\theta \in [0, \theta_0)$ where $\theta_0 = \arccos\frac{1}{\sqrt{K}}$.

(2) $\cos\phi = 1 - \alpha^2(x-y)^2$ and $\sin\phi = \alpha(x-y)\sqrt{2 - \alpha^2(x-y)^2}$.

(3) $\phi^* \ge \phi$ (equality holds only when $y = 0$) and $\phi^* > \theta$.

Proof Propositions (1) and (2) are computed by direct calculations. In particular, note that since $\cos\theta = \alpha x = 1/\sqrt{1 + (K-1)(y/x)^2}$ and $x > y \ge 0$, we have $\cos\theta \in (1/\sqrt{K}, 1]$ and $\theta \in [0, \theta_0)$. For Preposition (3), $\phi^* = \arccos\alpha y > \theta = \arccos\alpha x$ because $x > y$. Finally, for $x > y > 0$, we have

$$\frac{\cos\phi}{\cos\phi^*} = \frac{\alpha^2(2xy + (K-2)y^2)}{\alpha y} = \alpha(2x + (K-2)y) > \alpha(x + (K-1)y) > 1 \tag{48}$$

The final inequality is because $K \ge 2$, $x, y > 0$ and thus $(x + (K-1)y)^2 > x^2 + (K-1)^2y^2 > x^2 + (K-1)y^2 = \alpha^{-2}$. Therefore $\phi^* > \phi$. If $y = 0$ then $\phi^* = \phi$. ■

Theorem 7.6 *For the dynamics defined in Eqn. 16, there exists $\epsilon_0 > 0$ so that the trianglar region $\Omega_{\epsilon_0} = \{(x, y) : x \ge 0, y \ge 0, x \ge y + \epsilon_0\}$ (Fig. 6(c)) is a convergent region. That is, the flow goes inwards for all three edges and any trajectory starting in Ω_{ϵ_0} stays.*

Proof We discuss the three boundaries as follows:

Case 1: $y = 0, 0 \le x \le 1$, **horizontal line**. In this case, $\theta = 0$, $\phi = \pi/2$ and $\phi^* = \pi/2$. The y component of the dynamics in this line is:

$$f_1 \equiv \frac{2\pi}{N}\Delta y = -\frac{\pi}{2}(x-1) \ge 0 \tag{49}$$

So Δy points to the interior of Ω.

Case 2: $x = 1, 0 \le y \le 1$**, vertical line**. In this case, $\alpha \le 1$ and the x component of the dynamics is:

$$
\begin{aligned}
f_2 \equiv \frac{2\pi}{N}\Delta x &= -(\pi - \phi)(K-1)y - \theta + (K-1)(\alpha \sin\phi^* - \sin\phi) + \alpha\sin\theta && (50)\\
&= -(K-1)\left[(\pi-\phi)y - (\alpha\sin\phi^* - \sin\phi)\right] + (\alpha\sin\theta - \theta) && (51)
\end{aligned}
$$

Note that since $\alpha \le 1$, $\alpha\sin\theta \le \sin\theta \le \theta$, so the second term is non-positive. For the first term, we only need to check whether $(\pi-\phi)y - (\alpha\sin\phi^* - \sin\phi)$ is nonnegative. Note that

$$
\begin{aligned}
& (\pi-\phi)y - (\alpha\sin\phi^* - \sin\phi) && (52)\\
= \;& (\pi-\phi)y + \alpha(x-y)\sqrt{2-\alpha^2(x-y)^2} - \alpha\sqrt{1-\alpha^2y^2} && (53)\\
= \;& y\left[\pi - \phi - \alpha\sqrt{2-\alpha^2(x-y)^2}\right] + \alpha\left[x\sqrt{2-\alpha^2(x-y)^2} - \sqrt{1-\alpha^2y^2}\right] && (54)
\end{aligned}
$$

In Ω we have $(x-y)^2 \le 1$, combined with $\alpha \le 1$, we have $1 \le \sqrt{2-\alpha^2(x-y)^2} \le \sqrt{2}$ and $\sqrt{1-\alpha^2y^2} \le 1$. Since $x = 1$, the second term is nonnegative. For the first term, since $\alpha \le 1$,

$$
\pi - \phi - \alpha\sqrt{2-\alpha^2(x-y)^2} \ge \pi - \frac{\pi}{2} - \sqrt{2} > 0 \qquad (55)
$$

So $(\pi-\phi)y - (\alpha\sin\phi^* - \sin\phi) \ge 0$ and $\Delta x \le 0$, pointing inwards.

Case 3: $x = y + \epsilon$**,** $0 \le y \le 1$**, diagonal line**. We compute the inner product between $(\Delta x, \Delta y)$ and $(1, -1)$, the inward normal of Ω at the line. Using $\phi \le \frac{\pi}{2}\sin\phi$ for $\phi \in [0, \pi/2]$ and $\phi^* - \theta = \arccos\alpha y - \arccos\alpha x \ge 0$ when $x \ge y$, we have:

$$
\begin{aligned}
f_3(y,\epsilon) \equiv \frac{2\pi}{N}\begin{bmatrix}\Delta x\\ \Delta y\end{bmatrix}^\intercal \begin{bmatrix}1\\ -1\end{bmatrix} &= \phi^* - \theta - \epsilon\phi + \left[(K-1)(\alpha\sin\phi^* - \sin\phi) + \alpha\sin\theta\right]\epsilon && (56)\\
&\ge \epsilon(K-1)\left[\alpha\sin\phi^* - \left(1 + \frac{\pi}{2(K-1)}\right)\sin\phi\right]\\
&= \epsilon\alpha(K-1)\left[\sqrt{1-\alpha^2y^2} - \epsilon\left(1 + \frac{\pi}{2(K-1)}\right)\sqrt{2-\alpha^2\epsilon^2}\right]
\end{aligned}
$$

Note that for $y > 0$:

$$
\alpha y = \frac{1}{\sqrt{(x/y)^2 + (K-1)}} = \frac{1}{\sqrt{(1+\epsilon/y)^2 + (K-1)}} \le \frac{1}{\sqrt{K}} \qquad (57)
$$

For $y = 0$, $\alpha y = 0 < \sqrt{1/K}$. So we have $\sqrt{1-\alpha^2y^2} \ge \sqrt{1-1/K}$. And $\sqrt{2-\alpha^2\epsilon^2} \le \sqrt{2}$. Therefore $f_3 \ge \epsilon\alpha(K-1)(C_1 - \epsilon C_2)$ with $C_1 \equiv \sqrt{1-1/K} > 0$ and $C_2 \equiv \sqrt{2}(1 + \pi/2(K-1)) > 0$. With $\epsilon = \epsilon_0 > 0$ sufficiently small, $f_3 > 0$. ■

Lemma 7.7 (Reparametrization) *Denote* $\epsilon = x - y > 0$*. The terms* αx*,* αy *and* $\alpha\epsilon$ *involved in the trigometric functions in Eqn. 16 has the following parameterization:*

$$
\alpha\begin{bmatrix}y\\ x\\ \epsilon\end{bmatrix} = \frac{1}{K}\begin{bmatrix}\beta - \beta_2\\ \beta + (K-1)\beta_2\\ K\beta_2\end{bmatrix} \qquad (58)
$$

where $\beta_2 = \sqrt{(K-\beta^2)/(K-1)}$*. The reverse transformation is given by* $\beta = \sqrt{K-(K-1)\alpha^2\epsilon^2}$*. Here* $\beta \in [1, \sqrt{K})$ *and* $\beta_2 \in (0, 1]$*. In particular, the critical point* $(x, y) = (1, 0)$ *corresponds to* $(\beta, \epsilon) = (1, 1)$*. As a result, all trigometric functions in Eqn. 16 only depend on the single variable* β*. In particular, the following relationship is useful:*

$$
\beta = \cos\theta + \sqrt{K-1}\sin\theta \qquad (59)
$$

Proof This transformation can be checked by simple algebraic manipulation. For example:

$$
\frac{1}{\alpha K}(\beta - \beta_2) = \frac{1}{K}\left(\sqrt{\frac{K}{\alpha^2} - (K-1)\epsilon^2} - \epsilon\right) = \frac{1}{K}\left(\sqrt{(Ky+\epsilon)^2} - \epsilon\right) = y \qquad (60)
$$

To prove Eqn. 59, first we notice that $K\cos\theta = K\alpha x = \beta + (K-1)\beta_2$. Therefore, we have $(K\cos\theta - \beta)^2 - (K-1)^2\beta_2^2 = 0$, which gives $\beta^2 - 2\beta\cos\theta + 1 - K\sin^2\theta = 0$. Solving this quadratic equation and notice that $\beta \geq 1$, $\theta \in [0, \pi/2]$ and we get:

$$\beta = \cos\theta + \sqrt{\cos^2\theta + K\sin^2\theta - 1} = \cos\theta + \sqrt{K-1}\sin\theta \tag{61}$$

$\blacksquare$

Lemma 7.8 *After reparametrization (Eqn. 58),* $f_3(\beta, \epsilon) \geq 0$ *for* $\epsilon \in (0, \beta_2/\beta]$*. Furthermore, the equality is true only if* $(\beta, \epsilon) = (1, 1)$ *or* $(y, \epsilon) = (0, 1)$.

Proof Applying the parametrization (Eqn. 58) to Eqn. 56 and notice that $\alpha\epsilon = \beta_2 = \beta_2(\beta)$, we could write

$$f_3 = h_1(\beta) - (\phi + (K-1)\sin\phi)\epsilon \tag{62}$$

When β is fixed, f_3 now is a monotonously decreasing function with respect to $\epsilon > 0$. Therefore, $f_3(\beta, \epsilon) \geq f_3(\beta, \epsilon')$ for $0 < \epsilon \leq \epsilon' \equiv \beta_2/\beta$. If we could prove $f_3(\beta, \epsilon') \geq 0$ and only attain zero at known critical point $(\beta, \epsilon) = (1, 1)$, the proof is complete.

Denote $f_3(\beta, \epsilon') = f_{31} + f_{32}$ where

$$\begin{aligned} f_{31}(\beta, \epsilon') &= \phi^* - \theta - \epsilon'\phi + \epsilon'\alpha\sin\theta & (63) \\ f_{32}(\beta, \epsilon') &= (K-1)(\alpha\sin\phi^* - \sin\phi)\epsilon' & (64) \end{aligned}$$

For f_{32} it suffices to prove that $\epsilon'(\alpha\sin\phi^* - \sin\phi) = \beta_2\sin\phi^* - \frac{\beta_2}{\beta}\sin\phi \geq 0$, which is equivalent to $\sin\phi^* - \sin\phi/\beta \geq 0$. But this is trivially true since $\phi^* \geq \phi$ and $\beta \geq 1$. Therefore, $f_{32} \geq 0$. Note that the equality only holds when $\phi^* = \phi$ and $\beta = 1$, which corresponds to the horizontal line $x \in (0, 1], y = 0$.

For f_{31}, since $\phi^* \geq \phi$, $\phi^* > \theta$ and $\epsilon' \in (0, 1]$, we have the following:

$$f_{31} = \epsilon'(\phi^* - \phi) + (1-\epsilon')(\phi^* - \theta) - \epsilon'\theta + \beta_2\sin\theta \geq -\epsilon'\theta + \beta_2\sin\theta \geq \beta_2\left(\sin\theta - \frac{\theta}{\beta}\right) \tag{65}$$

And it reduces to showing whether $\beta\sin\theta - \theta$ is nonnegative. Using Eqn. 59, we have:

$$f_{33}(\theta) = \beta\sin\theta - \theta = \frac{1}{2}\sin 2\theta + \sqrt{K-1}\sin^2\theta - \theta \tag{66}$$

Note that $f_{33}' = \cos 2\theta + \sqrt{K-1}\sin 2\theta - 1 = \sqrt{K}\cos(2\theta - \theta_0) - 1$, where $\theta_0 = \arccos\frac{1}{\sqrt{K}}$. By Prepositions 1 in Lemma 7.5, $\theta \in [0, \theta_0)$. Therefore, $f_{33}' \geq 0$ and since $f_{33}(0) = 0$, $f_{33} \geq 0$. Again the equity holds when $\theta = 0$, $\phi^* = \phi$ and $\epsilon' = 1$, which is the critical point $(\beta, \epsilon) = (1, 1)$ or $(y, \epsilon) = (0, 1)$. $\blacksquare$

Theorem 7.9 *For the dynamics defined in Eqn. 16, the only critical point (*$\Delta x = 0$ *and* $\Delta y = 0$*) within* Ω_ϵ *is* $(y, \epsilon) = (0, 1)$.

Proof We prove by contradiction. Suppose (β, ϵ) is a critical point other than $\mathbf{w}^*$. A necessary condition for this to hold is $f_3 = 0$ (Eqn. 56). By Lemma 7.8, $\epsilon > \epsilon' = \beta_2/\beta > 0$ and

$$\epsilon - 1 + Ky = \frac{1}{\alpha}(\beta_2 - \alpha + \beta - \beta_2) = \frac{\beta - \alpha}{\alpha} = \frac{\beta - \beta_2/\epsilon}{\alpha} > \frac{\beta - \beta_2/\epsilon'}{\alpha} = 0 \tag{67}$$

So $\epsilon - 1 + Ky$ is strictly greater than zero. On the other hand, the condition $f_3 = 0$ implies that

$$((K-1)(\alpha\sin\phi^* - \sin\phi) + \alpha\sin\theta) = -\frac{1}{\epsilon}(\phi^* - \theta) + \phi \tag{68}$$

Using $\phi \in [0, \pi/2]$, $\phi^* \geq \phi$ and $\phi^* > \theta$, we have:

$$\begin{aligned} \frac{2\pi}{N}\Delta y &= -(\pi - \phi)(\epsilon - 1 + Ky) - (\phi^* - \phi) - \phi y + ((K-1)(\alpha\sin\phi^* - \sin\phi) + \alpha\sin\theta)\, y \\ &= -(\pi - \phi)(\epsilon - 1 + Ky) - (\phi^* - \phi) - \frac{1}{\epsilon}(\phi^* - \theta)y < 0 & (69) \end{aligned}$$

So the current point (β, ϵ) cannot be a critical point. $\blacksquare$

Theorem 7.10 *Any trajectory in Ω_{ϵ_0} converges to $(y, \epsilon) = (1, 0)$, following the dynamics defined in Eqn. 16.*

Proof We have Lyaponov function $V = \mathbb{E}[E]$ so that $\dot{V} = -\mathbb{E}[\Delta\mathbf{w}^\intercal\Delta\mathbf{w}] \leq -\mathbb{E}[\Delta\mathbf{w}]^\intercal\,\mathbb{E}[\Delta\mathbf{w}] \leq 0$. By Thm. 7.9, other than the optimal solution $\mathbf{w}^*$, there is no other symmetric critical point, $\Delta\mathbf{w} \neq 0$ and thus $\dot{V} < 0$. On the other hand, by Thm. 7.6, the triangular region Ω_{ϵ_0} is convergent, in which the 2D dynamics is C^∞ differentiable. Therefore, any 2D solution curve $\xi(t)$ will stay within. By PoincareBendixson theorem, when there is a unique critical point, the curve either converges to a limit circle or the critical point. However, limit cycle is not possible since V is strictly monotonous decreasing along the curve. Therefore, $\xi(t)$ will converge to the unique critical point, which is $(y, \epsilon) = (1, 0)$ and so does the symmetric system (Eqn. 12).

Theorem 7.11 *When $x = y \in (0, 1]$, the 2D dynamics (Eqn. 16) reduces to the following 1D case:*

$$\frac{2\pi}{N}\Delta x = -\pi K(x - x_*) \tag{70}$$

where $x_* = \frac{1}{\pi K}(\sqrt{K-1} - \arccos(1/\sqrt{K}) + \pi)$. Furthermore, x_* is a convergent critical point.

Proof The 1D system can be computed with simple algebraic manipulations (note that when $x = y$, $\phi = 0$ and $\theta = \phi^* = \arccos(1/\sqrt{K})$). Note that the 1D system is linear and its close form solution is $x^{(t)} = x_0 + Ce^{-K/2Nt}$ and thus convergent.