Supp Materials: An Analytical Formula of Population Gradient for two-layered ReLU network and its Applications in Convergence and Critical Point Analysis

Yuandong Tian
Facebook AI Research
yuandong@fb.com

March 18, 2017

1 Introduction

No theorem is provided.

2 Related Works

No theorem is provided.

3 Problem Definition

No theorem is provided.

4 The Analytical Formula

Here we list all detailed proof for all the theorems.

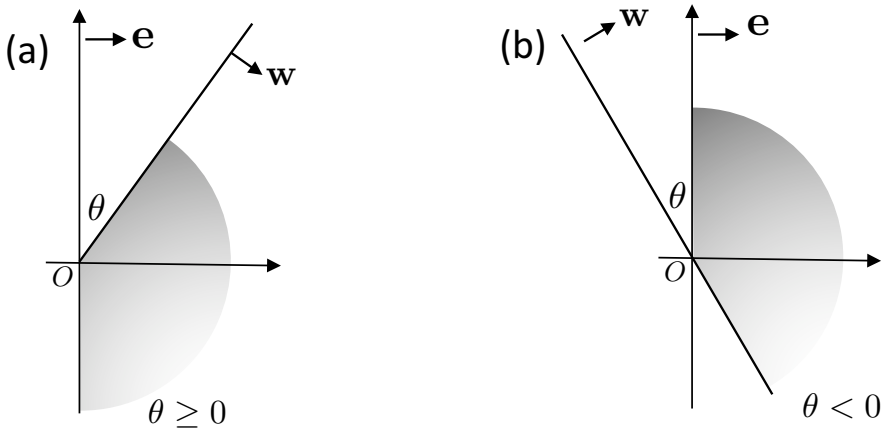

Figure 1: **(a)-(b)** Two cases in Thm. 1.

4.1 One ReLU Case

Theorem 1 *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* $\mathbf{x}_i \sim N(0, I)$*, then:*

$$\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right] = \frac{N}{2\pi}\left((\pi-\theta)\mathbf{w} + \|\mathbf{w}\|\sin\theta\mathbf{e}\right) \tag{1}$$

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:

$$F(\mathbf{e},\mathbf{w}) = \sum_{i:\mathbf{x}_i^\intercal\mathbf{e}\geq 0, \mathbf{x}_i^\intercal\mathbf{w}\geq 0} \mathbf{x}_i\mathbf{x}_i^\intercal\mathbf{w} \tag{2}$$

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. 1, 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: $\mathbf{e}$, $\mathbf{e}_\perp = -\frac{\mathbf{w}/\|\mathbf{w}\| - \mathbf{e}\cos\theta}{\sin\theta}$ (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:

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

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

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 spherical 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:

$$R(\phi_0) = \mathbb{E}\left[\frac{1}{N}\sum_{i:\phi_i\in[0,\phi_0]}\mathbf{x}_i\mathbf{x}_i^\intercal\right] = \frac{1}{4\pi}\begin{bmatrix} 2\phi_0 - \sin 2\phi_0 & 1-\cos 2\phi_0 & \mathbf{0} \\ 1-\cos 2\phi_0 & 2\phi_0 + \sin 2\phi_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 2\phi_0 I_{d-2}\end{bmatrix} \tag{3}$$

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

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. 1(a)). Using $\mathbf{w} = [\|\mathbf{w}\|\cos\theta, -\|\mathbf{w}\|\sin\theta, \mathbf{0}_{d-2}]$ and we have:

$$\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right] = N\left(R(\pi) - R(\theta)\right)\mathbf{w} \tag{5}$$

$$= \frac{N}{4\pi}\left(2(\pi-\theta)\mathbf{w} - \|\mathbf{w}\|\begin{bmatrix} -\sin 2\theta & 1-\cos 2\theta & \mathbf{0} \\ 1-\cos 2\theta & \sin 2\theta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 0\end{bmatrix}\begin{bmatrix}\cos\theta \\ -\sin\theta \\ \mathbf{0}\end{bmatrix}\right) \tag{6}$$

$$= \frac{N}{2\pi}\left((\pi-\theta)\mathbf{w} + \|\mathbf{w}\|\begin{bmatrix}\sin\theta \\ \mathbf{0}\end{bmatrix}\right) \tag{7}$$

$$= \frac{N}{2\pi}\left((\pi-\theta)\mathbf{w} + \|\mathbf{w}\|\sin\theta\mathbf{e}\right) \tag{8}$$

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. 1(b)), and similarly we get

$$\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right] = N\left(R(\pi+\theta) - R(0)\right)\mathbf{w} = \frac{N}{2\pi}\left((\pi+\theta)\mathbf{w} - \|\mathbf{w}\|\sin\theta\mathbf{e}\right) \tag{9}$$

Notice that by abuse of notation, the θ appears in Eqn. 1 is the absolute value and Eqn. 1 follows. ∎

5 Critical Point Analysis

Remember that we have: 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 $\mathbf{x}_i \sim N(0, I)$, then:

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

where $\theta \in [0, \pi]$ is the angle between $\mathbf{e}$ and $\mathbf{w}$. And the expected gradient for $g(\mathbf{x}) = \sum_{j=1}^{K} \sigma(\mathbf{w}_j^\intercal \mathbf{x})$ with respect to $\mathbf{w}_j$ is the following:

$$-\mathbb{E}\left[\nabla_{\mathbf{w}_j} J\right] = \sum_{j'=1}^{K} \mathbb{E}\left[F(\mathbf{e}_j, \mathbf{w}_{j'}^*)\right] - \mathbb{E}\left[F(\mathbf{e}_j, \mathbf{w}_{j'})\right] \tag{11}$$

where $\mathbf{e}_j = \mathbf{w}_j / \|\mathbf{w}_j\|$. By solving Eqn. 64 ($\mathbb{E}\left[\nabla_{\mathbf{w}_j} J\right] = 0 \quad j = 1, \ldots, K$), it is possible to identify all critical points of $g(\mathbf{x})$. In general Eqn. 64 is highly nonlinear and cannot be solved efficiently, however, we show that in our particular case, it is possible since Eqn. 64 has interesting properties.

First of all, the system of equations

$$\mathbb{E}\left[\nabla_{\mathbf{w}_j} J\right] = 0 \quad , j = 1, \ldots, K \tag{12}$$

or

$$\sum_{j'=1}^{K} \mathbb{E}\left[F(\mathbf{e}_j, \mathbf{w}_{j'}\right] = \sum_{j'=1}^{K} \mathbb{E}\left[F(\mathbf{e}_j, \mathbf{w}_{j'}^*\right] \quad , j = 1, \ldots, K \tag{13}$$

is a linear combination of $\mathbf{w}_j$ and $\mathbf{w}_j^*$ but with varying coefficients. We could rewrite the system as follows:

$$\mathrm{diag}\mathbf{a} E^\intercal + B\mathrm{diag}\bar{\mathbf{w}} E^\intercal = \mathrm{diag}\mathbf{a}^* E^\intercal + B^* W^{*\intercal} \tag{14}$$

where E, W, W^*, $\mathbf{a}$, B, $\mathbf{a}^*$ and B^* are all K-by-K matrices:

$$\mathbf{a} = \sin\Theta^\intercal \bar{\mathbf{w}} \quad , \qquad \mathbf{a}^* = \sin\Theta^{*\intercal} \bar{\mathbf{w}}^* \tag{15}$$

$$B = \pi \mathbf{1}\mathbf{1}^\intercal - \Theta^\intercal \quad , \qquad B^* = \pi \mathbf{1}\mathbf{1}^\intercal - \Theta^{*\intercal} \tag{16}$$

$$E = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K] \tag{17}$$

$$W = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K] \quad , \qquad W^* = [\mathbf{w}_1^*, \mathbf{w}_2^*, \ldots, \mathbf{w}_K^*] \tag{18}$$

where $\theta_j^{*j'} \equiv \angle(\mathbf{w}_j, \mathbf{w}_{j'}^*)$ and $\theta_{j'}^{j} = \theta_j^{j'} \equiv \angle(\mathbf{w}_j, \mathbf{w}_{j'})$, $\Theta = [\theta_j^i]$ (the element at i-th row, j-th column of Θ is θ_j^i) and $\Theta^* = [\theta_j^{*i}]$, $\bar{\mathbf{w}} = [\|\mathbf{w}_1\|, \|\mathbf{w}_2\|, \ldots, \|\mathbf{w}_K\|]$ and $\bar{\mathbf{w}}^* = [\|\mathbf{w}_1^*\|, \|\mathbf{w}_2^*\|, \ldots, \|\mathbf{w}_K^*\|]$.

Eqn. 14 already has interesting properties. The first thing we consider is whether the critical point will fall outside the *Principle Hyperplane* Π_*, which is the plane spanned by the ground truth weight vectors W^*. The following theorem shows that the critical points outside Π_* must lie in a manifold:

Lemma 1 *If $\{\mathbf{w}_j\}$ is a critical point satisfying Eqn. 14, then for any orthogonal mapping R with $R|_{\Pi_*} = Id$, $\{R\mathbf{w}_j\}$ is also a critical point.*

Proof First of all, since R is an orthogonal transformation, it keeps all angles and magnitudes and $\mathbf{a}$, $\mathbf{a}^*$, B, $\bar{\mathbf{w}}$ and $\bar{\mathbf{w}}^*$ are invariant. For simplicity we write $Y = \mathrm{diag}\mathbf{a} + B\mathrm{diag}\bar{\mathbf{w}} - \mathrm{diag}\mathbf{a}^*$ and Y is also invariant under R. Since $R|_{\Pi_*} = Id$, we have $RW^* = W^*$ and

$$Y_R E_R^\intercal - B_R^* W^{*\intercal} = Y E^\intercal R^\intercal - B^* W^{*\intercal} R^\intercal = \left(Y E^\intercal - B^* W^{*\intercal}\right) R^\intercal = 0 \tag{19}$$

∎

Note that for $d \geq K + 2$, there always exists $R \neq Id$ and satisfy such a condition, which yield continuous critical points. Further, such a transformation forms a Lie group. Therefore we have:

Theorem 2 *If $d \geq K+2$, then any critical point satisfying Eqn. 14 and is outside Π_* must lie in a manifold.*

The intuition is simple. For any out-of-plane critical point, pick a matrix that satisfies the condition of the theorem, and tranforms it to a different yet infinitely close critical points. Such a matrix always exists, since for the $d - K$ subspace, if it is odd, then we can always pick a rotation whose fixed axis is not aligned with all K weights; if it is even, then there is a rotation matrix without a fixed point.

5.1 Characteristics within the Principle Plane

We could right-multiple E and turn the normal equation to a linear function with respect to the magnitude of weights $\|\mathbf{w}\|$. Note that we have:

$$E^\intercal E = \cos\Theta, \qquad W^\intercal E = \text{diag}\bar{\mathbf{w}}\cos\Theta, \qquad (W^*)^\intercal E = \text{diag}\bar{\mathbf{w}}^*\Theta^* \tag{20}$$

Therefore, Eqn. 14 becomes:

$$\text{diag}\mathbf{a}\cos\Theta + B\text{diag}\bar{\mathbf{w}}\cos\Theta = \text{diag}\mathbf{a}^*\cos\Theta + B^*\text{diag}\bar{\mathbf{w}}^*\cos\Theta^* \tag{21}$$

which is a homogenous linear equation with respect to the magnitude of the weights (note that $\mathbf{a}$ and $\mathbf{a}^*$ is linear to the magnitudes). In particular, the (i,j) entry of the LHS and RHS of this equality are:

$$LHS_{ij} = \cos\theta_j^i \left(\sum_{k=1}^{K}\sin\theta_i^k\|\mathbf{w}_k\|\right) + \sum_{k=1}^{K}(\pi-\theta_i^k)\|\mathbf{w}_k\|\cos\theta_j^k \tag{22}$$

$$RHS_{ij} = \cos\theta_j^i \left(\sum_{k=1}^{K}\sin\theta_i^{*k}\|\mathbf{w}_k^*\|\right) + \sum_{k=1}^{K}(\pi-\theta_i^{*k})\|\mathbf{w}_k^*\|\cos\theta_j^{*k} \tag{23}$$

Therefore, the following equation holds:

$$M\bar{\mathbf{w}} = M^*\bar{\mathbf{w}}^* \tag{24}$$

where M and M^* are K^2-by-K matrices. Each entry $m_{ij,k}$ that correponds to the coefficient of k-th weight vector at (i,j) entry of Eqn. 14 is defined as:

$$m_{ij,k} = (\pi-\theta_i^k)\cos\theta_j^k + \sin\theta_i^k\cos\theta_j^i \tag{25}$$

$$m_{ij,k}^* = (\pi-\theta_i^{*k})\cos\theta_j^{*k} + \sin\theta_i^{*k}\cos\theta_j^i \tag{26}$$

Special case on the diagonal. For diagonal element (i,i), $\cos\theta_i^i = 1$ and $m_{ii,k} = h(\theta_i^k)$, $m_{ii,k}^* = h(\theta_i^{*k})$, where

$$h(\theta) = (\pi-\theta)\cos\theta + \sin\theta. \tag{27}$$

Therefore, with only diagonal element, we arrive at the following subset of the constraints to be satisfied for any critical points:

$$M_r\bar{\mathbf{w}} = M_r^*\bar{\mathbf{w}}^* \tag{28}$$

where $M_r = h(\Theta^\intercal)$ and $M_r^* = h(\Theta^{*\intercal})$ are both K-by-K matrices. Note that if M_r is full-rank, then we could solve $\bar{\mathbf{w}}$ from Eqn. 28 and plug it back in Eqn. 24 to check whether it is indeed a critical point.

Lemma 2 *If $\bar{\mathbf{w}}^* \neq 0$ (no trivial ground truth solutions), and for a given (Θ,Θ^*), there exists a row (e.g. l-th row) of M and M^*, namely $\mathbf{m}_l^\intercal$ and $\mathbf{m}_l^{*\intercal}$, satisfying*

$$\mathbf{m}_l^* - M_r^{*\intercal}M_r^{-1}\mathbf{m}_l > 0 \quad \text{or} \quad \mathbf{m}_l^* - M_r^{*\intercal}M_r^{-1}\mathbf{m}_l < 0 \tag{29}$$

Then (Θ,Θ^*) cannot be a critical point.

Proof Suppose given (Θ,Θ^*), we get M_r and M_r^* and compute $\bar{\mathbf{w}}$ using Eqn. 28, then we have

$$(\bar{\mathbf{w}}^*)^\intercal M_r^{*\intercal}M_r^{-1}\mathbf{m}_l = (M_r^*\bar{\mathbf{w}}^*)^\intercal M_r^{-1}\mathbf{m}_l = \bar{\mathbf{w}}^\intercal M_r^\intercal M_r^{-1}\mathbf{m}_l = \bar{\mathbf{w}}^\intercal\mathbf{m}_l \tag{30}$$

Therefore, from the condition $\mathbf{m}_l^* - M_r^{*\intercal}M_r^{-1}\mathbf{m}_l > 0$ and $\bar{\mathbf{w}}^* \geq 0$ but $\bar{\mathbf{w}}^* \neq 0$, we have

$$(\bar{\mathbf{w}}^*)^\intercal(\mathbf{m}_l^* - M_r^{*\intercal}M_r^{-1}\mathbf{m}_l) = (\bar{\mathbf{w}}^*)^\intercal\mathbf{m}_l^* - \bar{\mathbf{w}}^\intercal\mathbf{m}_l > 0 \tag{31}$$

but this contradicts with the necessary condition for (Θ,Θ^*) to become a critical point (Eqn. 24). Similarly we can prove the other side. ∎

Separation property of Eqn. 29. Note that both the k-th element of $\mathbf{m}_l^*$ and $M_r^{*\intercal}M_r^{-1}\mathbf{m}_l$ in Eqn. 29 are only dependent on the k-th true weight vector $\mathbf{w}_k^*$ (and all $\{\mathbf{w}_j\}$).

- For $\mathbf{m}_l^*$, this can be seen by Eqn. 26, in which the k-th element is only related to the angles $\theta_\cdot^{*k}$ between $\mathbf{w}_k^*$ and $\{\mathbf{w}_j\}$.
- For $M_r^{*\intercal} M_r^{-1} \mathbf{m}_l$, notice that the k-th column of M_r^* (the k-th row of $M_r^{*\intercal}$) is only related to $\mathbf{w}_k^*$ but not other ground truth weight vectors. This separation property makes analysis much easier, as shown in the case of $K = 2$.

Therefore, we could consider the following function regarding to one (rather than K) ground truth unit weight vector $\mathbf{e}^*$ and all estimated unit vectors $\{\mathbf{e}_l\}$:

$$L_{ij}(\mathbf{e}^*, \{\mathbf{e}_l\}) = m_{ij}^* - \mathbf{v}^{*\intercal} M_r^{-1} \mathbf{m}_{ij} \tag{32}$$

where $\mathbf{v}^{*\intercal} = [h(\theta_1^*), h(\theta_2^*), \ldots, h(\theta_K^*)]$, $\theta_j^* = \angle(\mathbf{e}^*, \mathbf{w}_j)$ and $m_{ij}^* = (\pi - \theta_i^*)\cos\theta_j^* + \sin\theta_i^* \cos\theta_j^i$ (like Eqn. 26).

Proposition 1 *$L_{ij}(\mathbf{e}^*, \{\mathbf{e}_l\}) = 0$ for any $\mathbf{e}^* = \mathbf{e}_l$, $1 \le l \le K$. In addition, $L_{ii}(\mathbf{e}^*, \{\mathbf{e}_l\}) \equiv 0$.*

Proof When $\mathbf{e}^* = \mathbf{e}_l$, then $\theta_k^* = \theta_k^l$ and $\mathbf{v}^{*\intercal}$ becomes the l-th row of M_r. Since $M_r M_r^{-1} = I_{K\times K}$, $\mathbf{v}^{*\intercal} M_r^{-1}$ becomes a unit vector with only l-th element being 1. Therefore, again with $\theta_k^* = \theta_k^l$, we have:

$$L_{ij}(\mathbf{e}^*, \{\mathbf{e}_l\}) = m_{ij}^* - m_{ij,l} = 0 \tag{33}$$

For L_{ii}, by definition $\mathbf{m}_{ii}$ is i-th column of M_r, so $M_r^{-1}\mathbf{m}_{ii}$ is a unit vector with only i-th element being 1. Therefore

$$(\mathbf{v}^*)^\intercal M_r^{-1} \mathbf{m}_{ii} = h(\theta_i^*) = m_{ii}^* \tag{34}$$

∎

Then the previous lemma can be written as the following:

Theorem 3 *If $\bar{\mathbf{w}}^* \neq 0$, and for a given parameter $\mathbf{w}$, $L_{jj'}(\{\theta_l^{*k}\}, \Theta) > 0$ or < 0 for all $1 \le k \le K$, then $\mathbf{w}$ cannot be a critical point.*

5.2 ReLU network with two hidden nodes ($K = 2$)

For $K = 2$, we have 4-by-2 matrix (the row order is $(1,1),(1,2),(2,1),(2,2)$):

$$\begin{aligned} M &= \begin{bmatrix} (\pi - \theta_1^1)\cos\theta_1^1 + \sin\theta_1^1\cos\theta_1^1 & (\pi - \theta_1^2)\cos\theta_1^2 + \sin\theta_1^2\cos\theta_1^1 \\ (\pi - \theta_1^1)\cos\theta_2^1 + \sin\theta_1^1\cos\theta_2^1 & (\pi - \theta_1^2)\cos\theta_2^2 + \sin\theta_1^2\cos\theta_2^1 \\ (\pi - \theta_2^1)\cos\theta_1^1 + \sin\theta_2^1\cos\theta_1^2 & (\pi - \theta_2^2)\cos\theta_1^2 + \sin\theta_2^2\cos\theta_1^2 \\ (\pi - \theta_2^1)\cos\theta_2^1 + \sin\theta_2^1\cos\theta_2^2 & (\pi - \theta_2^2)\cos\theta_2^2 + \sin\theta_2^2\cos\theta_2^2 \end{bmatrix} \qquad (35) \\ &= \begin{bmatrix} \pi & (\pi - \theta)\cos\theta + \sin\theta \\ \pi\cos\theta & (\pi - \theta) + \sin\theta\cos\theta \\ (\pi - \theta) + \sin\theta\cos\theta & \pi\cos\theta \\ (\pi - \theta)\cos\theta + \sin\theta & \pi \end{bmatrix} \qquad (36) \end{aligned}$$

since $\theta_1^2 = \theta_2^1 = \theta$, $\theta_1^1 = \theta_2^2 = 0$. Similarly we could write M^*:

$$M^* = \begin{bmatrix} (\pi - \theta_1^{*1})\cos\theta_1^{*1} + \sin\theta_1^{*1}\cos\theta_1^1 & (\pi - \theta_1^{*2})\cos\theta_1^{*2} + \sin\theta_1^{*2}\cos\theta_1^1 \\ (\pi - \theta_1^{*1})\cos\theta_2^{*1} + \sin\theta_1^{*1}\cos\theta_2^1 & (\pi - \theta_1^{*2})\cos\theta_2^{*2} + \sin\theta_1^{*2}\cos\theta_2^1 \\ (\pi - \theta_2^{*1})\cos\theta_1^{*1} + \sin\theta_2^{*1}\cos\theta_1^2 & (\pi - \theta_2^{*2})\cos\theta_1^{*2} + \sin\theta_2^{*2}\cos\theta_1^2 \\ (\pi - \theta_2^{*1})\cos\theta_2^{*1} + \sin\theta_2^{*1}\cos\theta_2^2 & (\pi - \theta_2^{*2})\cos\theta_2^{*2} + \sin\theta_2^{*2}\cos\theta_2^2 \end{bmatrix} \tag{37}$$

In this case,

$$M_r = \begin{bmatrix} h(\theta_1^1) & h(\theta_1^2) \\ h(\theta_2^1) & h(\theta_2^2) \end{bmatrix}, \qquad M_r^* = \begin{bmatrix} h(\theta_1^{*1}) & h(\theta_1^{*2}) \\ h(\theta_2^{*1}) & h(\theta_2^{*2}) \end{bmatrix} \tag{38}$$

Therefore, if we know $\theta_1^2 = \theta_2^1$, θ_1^{*1}, θ_1^{*2}, θ_2^{*1} and θ_2^{*2}, then we could compute M and M^* and solve a linear equation to get the magnitude of $\mathbf{w}_1$ and $\mathbf{w}_2$, which collectly identify the critical points. Note that M is a 4-by-2 matrix, so critical point only happens if the matrix has singular structure.

Global Optimum. One special case is when $\theta_1^2 = \theta_2^1 = \theta_1^{*2} = \theta_2^{*1} = \pi/2$ and $\theta_1^{*1} = \theta_2^{*2} = 0$, in this case, we have:

$$M = M^* = \begin{bmatrix} \pi & 1 \\ 0 & \pi/2 \\ \pi/2 & 0 \\ 1 & \pi \end{bmatrix} \tag{39}$$

and thus $\|\mathbf{w}_j\| = \|\mathbf{w}_j^*\|$ is the unique solution.

When $K = 2$, the following conjecture is empirically correct.

Conjecture 1 *If $\mathbf{e}^*$ is in the interior of* $\text{Cone}(\mathbf{e}_1, \mathbf{e}_2)$*, then $L_{12}(\theta_1^*, \theta_2^*, \theta_2^1) > 0$. If $\mathbf{e}^*$ is in the exterior, then $L_{12} < 0$. If $\mathbf{e}^*$ is on the boundary then $L_{12} = 0$. Same for L_{21}.*

Remark Note that L_{1j} can be written as the following:

$$\begin{aligned} L_{1j}(\mathbf{e}^*, \{\mathbf{e}_1, \mathbf{e}_2\}) &= m_{1j}^* - [h(\theta_1^*), h(\theta_2^*)]\, M_r^{-1} \mathbf{m}_{1j} && (40) \\ &= [(\pi - \theta_1^*)\mathbf{e}^* + \sin\theta_1^* \mathbf{e}_1]^\intercal\, \mathbf{e}_j && (41) \\ &- [\alpha(\theta_1^*, \theta_2^*, \theta), \beta(\theta_1^*, \theta_2^*, \theta)] \begin{bmatrix} (\pi - \theta_1^1)\mathbf{e}_1^\intercal + \sin\theta_1^1 \mathbf{e}_1^\intercal \\ (\pi - \theta_1^2)\mathbf{e}_2^\intercal + \sin\theta_1^2 \mathbf{e}_1^\intercal \end{bmatrix} \mathbf{e}_j && (42) \end{aligned}$$

Here we have

$$[\alpha, \beta] = [\alpha(\theta_1^*, \theta_2^*, \theta), \beta(\theta_1^*, \theta_2^*, \theta)] \equiv [h(\theta_1^*), h(\theta_2^*)] M_r^{-1} \tag{43}$$

We know that $L_{11} = 0$ by Proposition 1. Therefore

$$\begin{aligned} \mathbf{u}_{1j} &\equiv (\pi - \theta_1^*)\mathbf{e}^* + \sin\theta_1^* \mathbf{e}_1 - \left[(\pi - \theta_1^1)\mathbf{e}_1 + \sin\theta_1^1 \mathbf{e}_1, (\pi - \theta_1^2)\mathbf{e}_2 + \sin\theta_1^2 \mathbf{e}_1\right] \begin{bmatrix} \alpha(\theta_1^*, \theta_2^*, \theta_2^1) \\ \beta(\theta_1^*, \theta_2^*, \theta_2^1) \end{bmatrix} \\ &= (\pi - \theta_1^*)\mathbf{e}^* + \sin\theta_1^* \mathbf{e}_1 - [\pi \mathbf{e}_1, (\pi - \theta)\mathbf{e}_2 + \sin\theta \mathbf{e}_1] \begin{bmatrix} \alpha(\theta_1^*, \theta_2^*, \theta) \\ \beta(\theta_1^*, \theta_2^*, \theta) \end{bmatrix} && (44) \end{aligned}$$

is perpendicular to $\mathbf{e}_1$. So if we compute the inner product between $\mathbf{u}_{12}$ and $\mathbf{e}_1^\perp$ (the unit vector that is in Π_* and is orthogonal to $\mathbf{e}_1$), we get

$$\mathbf{u}_{12}^\intercal \mathbf{e}_1^\perp = (\pi - \theta_1^*)\sin\theta_1^* - [(\pi - \theta)\sin\theta]\, \beta \tag{45}$$

Since $\mathbf{e}_2 = \cos\theta \mathbf{e}_1 + \sin\theta \mathbf{e}_1^\perp$ so we have:

$$L_{12}(\mathbf{e}^*, \{\mathbf{e}_1, \mathbf{e}_2\}) = \mathbf{u}_{12}^\intercal \mathbf{e}_2 = \sin\theta(\mathbf{u}_{12}^\intercal \mathbf{e}_1^\perp) \tag{46}$$

Note that Eqn. 45 is a function with 2-variables θ and θ_1^* (θ_2^* is determined by θ and θ_1^*, depending on whether $\mathbf{e}^*$ is inside or outside $Cone(\mathbf{e}_1, \mathbf{e}_2)$). And we could verify it numerically.

Theorem 4 *If Conjecture 1 is correct, then for 2 ReLU network, $(\mathbf{w}_1, \mathbf{w}_2)$ ($\mathbf{w}_1 \neq \mathbf{w}_2$) is not a critical point, if they both are in* $\text{Cone}(\mathbf{w}_1^*, \mathbf{w}_2^*)$*, or both out of it.*

Proof If both $\mathbf{w}_1^*$ and $\mathbf{w}_2^*$ are inside $\text{Cone}(\mathbf{w}_1, \mathbf{w}_2)$, then from Conjecture 1, we have

$$L_{12}(\theta_1^{k*}, \theta_2^{k*}, \theta_2^1) > 0 \tag{47}$$

for $k = 1, 2$. Since $K = 2$ we could simply apply Thm. **??** to say $(\mathbf{w}_1, \mathbf{w}_2)$ is not a critical point. Similary we prove the case for both $\mathbf{w}_1^*$ and $\mathbf{w}_2^*$ outside $\text{Cone}(\mathbf{w}_1, \mathbf{w}_2)$. ∎

6 Convergence Analysis

6.1 Single ReLU case

In this subsection, we mainly deal with the following dynamics:

$$\mathbb{E}\left[\nabla_{\mathbf{w}} J\right] = \frac{N}{2}(\mathbf{w} - \mathbf{w}^*) + \frac{N}{2\pi}\left(\theta \mathbf{w}^* - \frac{\|\mathbf{w}^*\|}{\|\mathbf{w}\|}\sin\theta \mathbf{w}\right) \tag{48}$$

Theorem 5 *In the region* $\|\mathbf{w}^0 - \mathbf{w}^*\| < \|\mathbf{w}^*\|$*, following the dynamics (Eqn. 48), 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 $\Omega = \{\mathbf{w} : \|\mathbf{w}^0 - \mathbf{w}^*\| < \|\mathbf{w}^*\|\}$. Note that

$$\dot{V} = -(\mathbf{w} - \mathbf{w}^*)^\intercal \nabla_{\mathbf{w}} J = -\mathbf{y}^\intercal M \mathbf{y} \tag{49}$$

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

$$M = \frac{1}{2}\begin{bmatrix} \sin 2\theta + 2\pi - 2\theta & -(2\pi - \theta)\cos\theta - \sin\theta \\ -(2\pi - \theta)\cos\theta - \sin\theta & 2\pi \end{bmatrix} \tag{50}$$

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:

$$\begin{aligned} 4\det(M) &= 2\pi(\sin 2\theta + 2\pi - 2\theta) - [(2\pi - \theta)\cos\theta + \sin\theta]^2 && (51)\\ &= 2\pi(\sin 2\theta + 2\pi - 2\theta) - \left[(2\pi - \theta)^2\cos^2\theta + (2\pi - \theta)\sin 2\theta + \sin^2\theta\right] && (52)\\ &= (4\pi^2 - 1)\sin^2\theta - 4\pi\theta + 4\pi\theta\cos^2\theta - \theta^2\cos^2\theta + \theta\sin 2\theta && (53)\\ &= (4\pi^2 - 4\pi\theta - 1)\sin^2\theta + \theta\cos\theta(2\sin\theta - \theta\cos\theta) && (54) \end{aligned}$$

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. ∎

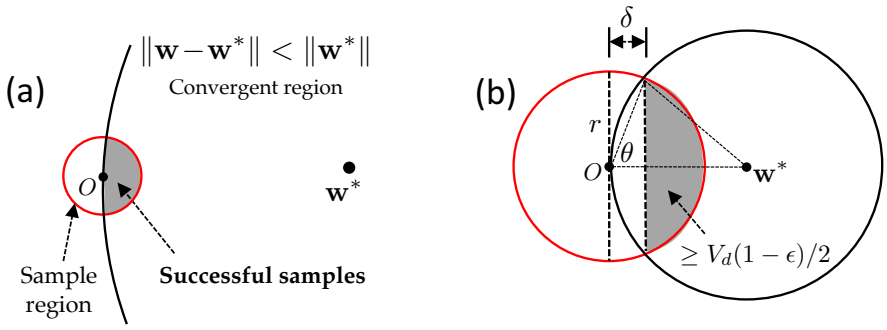

Figure 2: **(a)** Sampling strategy to maximize the probability of convergence. **(b)** Relationship between sampling range r and desired probability of success $(1 - \epsilon)/2$.

Theorem 6 *When* $K = 1$*, the dynamics in Eqn. 64 converges to* $\mathbf{w}^*$ *with probability at least* $(1 - \epsilon)/2$*, if the initial value* $\mathbf{w}^0$ *is sampled uniformly from* $B_r = \{\mathbf{w} : \|\mathbf{w}\| \leq r\}$ *with:*

$$r \leq \epsilon\sqrt{\frac{2\pi}{d+1}}\|\mathbf{w}^*\| \tag{55}$$

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:

$$\frac{1}{2}V_d(r) - \delta V_{d-1} \geq \frac{1-\epsilon}{2}V_d(r) \tag{56}$$

which gives

$$\delta \leq \frac{\epsilon}{2}\frac{V_d}{V_{d-1}} \tag{57}$$

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

$$r \leq \epsilon \frac{V_d(1)}{V_{d-1}(1)}\|\mathbf{w}^*\| \tag{58}$$

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

$$V_d(1) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \tag{59}$$

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

$$\frac{V_d(1)}{V_{d-1}(1)} = \sqrt{\pi}\frac{\Gamma(d/2+1/2)}{\Gamma(d/2+1)} \tag{60}$$

From Gautschi's Inequality

$$x^{1-s} < \frac{\Gamma(x+1)}{\Gamma(x+s)} < (x+s)^{1-s} \qquad x > 0, 0 < s < 1 \tag{61}$$

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

$$\left(\frac{d+1}{2}\right)^{-1/2} < \frac{\Gamma(d/2+1/2)}{\Gamma(d/2+1)} < \left(\frac{d}{2}\right)^{-1/2} \tag{62}$$

Therefore, it suffices to have

$$r \leq \epsilon\sqrt{\frac{2\pi}{d+1}}\|\mathbf{w}^*\| \tag{63}$$

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

6.2 Multiple ReLU case

Explanation of Eqn. 18. We first write down the dynamics to be studied:

$$-\mathbb{E}\left[\nabla_{\mathbf{w}_j} J\right] = \sum_{j'=1}^{K} \mathbb{E}\left[F(\mathbf{e}_j, \mathbf{w}^*_{j'})\right] - \mathbb{E}\left[F(\mathbf{e}_j, \mathbf{w}_{j'})\right] \tag{64}$$

We first define $f(\mathbf{w}_j, \mathbf{w}_{j'}, \mathbf{w}^*_{j'}) = F(\mathbf{w}_j/\|\mathbf{w}_j\|, \mathbf{w}^*_{j'}) - F(\mathbf{w}_j/\|\mathbf{w}_j\|, \mathbf{w}_{j'})$. Therefore, the dynamics can be written as:

$$-\mathbb{E}\left[\nabla_{\mathbf{w}_j} J\right] = \sum_{j'} \mathbb{E}\left[f(\mathbf{w}_j, \mathbf{w}_{j'}, \mathbf{w}^*_{j'})\right] \tag{65}$$

Suppose we have a finite group $\mathcal{P} = \{P_j\}$ which is a subgroup of orthognoal group $O(d)$. P_1 is the identity element. If $\mathbf{w}$ and $\mathbf{w}^*$ have the following symmetry: $\mathbf{w}_j = P_j\mathbf{w}$ and $\mathbf{w}^*_j = P_j\mathbf{w}^*$, then RHS of Eqn. 64 can be simplified as follows:

$$\begin{aligned}
-\mathbb{E}\left[\nabla_{\mathbf{w}_j} J\right] &= \sum_{j'} \mathbb{E}\left[f(\mathbf{w}_j, \mathbf{w}_{j'}, \mathbf{w}^*_{j'})\right] = \sum_{j'} \mathbb{E}\left[f(P_j\mathbf{w}, P_{j'}\mathbf{w}, P_{j'}\mathbf{w}^*)\right] \\
&= \sum_{j''} \mathbb{E}\left[f(P_j\mathbf{w}, P_jP_{j''}\mathbf{w}, P_jP_{j''}\mathbf{w}^*)\right] \quad (\{P_j\}_{j=1}^K \text{ is a group}) \\
&= P_j \sum_{j''} \mathbb{E}\left[f(\mathbf{w}, P_{j''}\mathbf{w}, P_{j''}\mathbf{w}^*)\right] \quad (\|P\mathbf{w}_1\| = \|\mathbf{w}_1\|,\ \angle(P\mathbf{w}_1, P\mathbf{w}_2) = \angle(\mathbf{w}_1, \mathbf{w}_2)) \\
&= -P_j\mathbb{E}\left[\nabla_{\mathbf{w}_1} J\right]
\end{aligned} \tag{66}$$

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:

$$-\mathbb{E}\left[\nabla_{\mathbf{w}} J\right] = \sum_{j=1}^{K} \mathbb{E}\left[f(\mathbf{w}, P_j\mathbf{w}, P_j\mathbf{w}^*)\right] \tag{67}$$

Theorem 7 *For a bias-free two-layered ReLU network*

$$g(\mathbf{x};\mathbf{w}) = \sum_j \sigma(\mathbf{w}_j^{\intercal}\mathbf{x}) \tag{68}$$

that takes spherical Gaussian inputs, if the teacher's parameters $\{\mathbf{w}_j^*\}$ form a set of orthnomal bases, then:

(1) When the student parameters is initialized to be $[x^0, y^0, \ldots, y^0]$ under the basis of $\mathbf{w}^*$, where $(x^0, y^0) \in \Omega = \{x \in (0,1], y \in [0,1], x > y\}$, then the dynamics (Eqn. 64) converges to teacher's parameters $\{\mathbf{w}_j^*\}$ (or $(x,y) = (1,0)$);

(2) when $x^0 = y^0 \in (0,1]$, then it converges to a saddle point $x = y = \frac{1}{\pi K}(\sqrt{K-1} - \arccos(1/\sqrt{K}) + \pi)$.

This theorem is quite complicated. We will start with a few lemmas and gradually come to the conclusion.

First, if $\mathbf{w}^0 = [x, y, y, \ldots, y]$ under the bases $\{\mathbf{w}_j^*\}_{j=1}^K$, then from simple computation we know that $\mathbf{w}^t$ also follows this pattern. Therefore, we only need to study the following 2D dynamics related to x and y:

$$\begin{aligned} -\frac{2\pi}{N}\mathbb{E}\begin{bmatrix}\nabla_x J\\ \nabla_y J\end{bmatrix} &= -\left\{[(\pi-\phi)(x-1+(K-1)y)]\begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}\theta\\ \phi^*-\phi\end{bmatrix} + \phi\begin{bmatrix}x-1\\ y\end{bmatrix}\right\} \\ &+ [(K-1)(\alpha\sin\phi^* - \sin\phi) + \alpha\sin\theta]\begin{bmatrix}x\\ y\end{bmatrix} \end{aligned} \tag{69}$$

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:

$$\alpha = (x^2 + (K-1)y^2)^{-1/2}, \quad \cos\theta = \alpha x, \quad \cos\phi^* = \alpha y, \quad \cos\phi = \alpha^2(2xy + (K-2)y^2) \tag{70}$$

Now we start a sequence of lemmas.

Lemma 3 *For ϕ^*, θ and ϕ defined in Eqn. 70:*

$$\alpha \equiv (x^2 + (K-1)y^2)^{-1/2} \tag{71}$$
$$\cos\theta \equiv \alpha x \tag{72}$$
$$\cos\phi^* \equiv \alpha y \tag{73}$$
$$\cos\phi \equiv \alpha^2(2xy + (K-2)y^2) \tag{74}$$

we have the following relations in the triangular region $\Omega_{\epsilon_0} = \{(x,y) : x \geq 0, y \geq 0, x \geq y + \epsilon_0\}$ (Fig. 1(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) $\phi^* \geq \phi$ (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 $x > y \geq 0$, 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

$$\frac{\cos\phi}{\cos\phi^*} = \frac{\alpha^2(2xy + (K-2)y^2)}{\alpha y} = \alpha(2x + (K-2)y) > \alpha(x + (K-1)y) > 1 \tag{75}$$

The final inequality is because $K \geq 2$, $x, y > 0$ 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 $\phi^* = \phi$. ∎

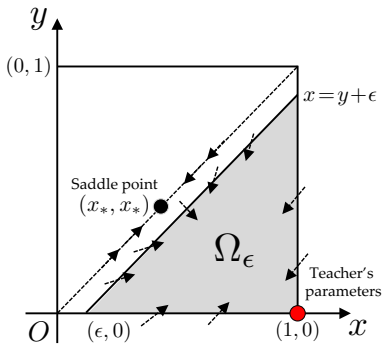

Figure 3: The region Ω_ϵ considered in the analysis of Eqn. 69.

Lemma 4 *For the dynamics defined in Eqn. 69, there exists $\epsilon_0 > 0$ so that the trianglar region $\Omega_{\epsilon_0} = \{(x, y) : x \geq 0, y \geq 0, x \geq y + \epsilon_0\}$ (Fig. 3) 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: $y = 0, 0 \leq x \leq 1$, horizontal line. In this case, $\theta = 0$, $\phi = \pi/2$ and $\phi^* = \pi/2$. The y component of the dynamics in this line is:

$$f_1 \equiv -\frac{2\pi}{N}\nabla_y J = -\frac{\pi}{2}(x-1) \geq 0 \tag{76}$$

So $-\nabla_y J$ points to the interior of Ω.

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

$$\begin{aligned} f_2 \equiv -\frac{2\pi}{N}\nabla_x J &= -(\pi-\phi)(K-1)y - \theta + (K-1)(\alpha\sin\phi^* - \sin\phi) + \alpha\sin\theta & (77)\\ &= -(K-1)\left[(\pi-\phi)y - (\alpha\sin\phi^* - \sin\phi)\right] + (\alpha\sin\theta - \theta) & (78)\end{aligned}$$

Note that since $\alpha \leq 1$, $\alpha\sin\theta \leq \sin\theta \leq \theta$, 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

$$\begin{aligned} & (\pi-\phi)y - (\alpha\sin\phi^* - \sin\phi) & (79)\\ = \;& (\pi-\phi)y + \alpha(x-y)\sqrt{2-\alpha^2(x-y)^2} - \alpha\sqrt{1-\alpha^2 y^2} & (80)\\ = \;& y\left[\pi - \phi - \alpha\sqrt{2-\alpha^2(x-y)^2}\right] + \alpha\left[x\sqrt{2-\alpha^2(x-y)^2} - \sqrt{1-\alpha^2y^2}\right] & (81)\end{aligned}$$

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

$$\pi - \phi - \alpha\sqrt{2-\alpha^2(x-y)^2} \geq \pi - \frac{\pi}{2} - \sqrt{2} > 0 \tag{82}$$

So $(\pi-\phi)y - (\alpha\sin\phi^* - \sin\phi) \geq 0$ and $-\nabla_x J \leq 0$, pointing inwards.

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

$$\begin{aligned} f_3(y,\epsilon) \equiv -\frac{2\pi}{N}\begin{bmatrix}\nabla_x J\\ \nabla_y J\end{bmatrix}^\intercal\begin{bmatrix}1\\-1\end{bmatrix} &= \phi^* - \theta - \epsilon\phi + \left[(K-1)(\alpha\sin\phi^* - \sin\phi) + \alpha\sin\theta\right]\epsilon & (83)\\ &\geq \epsilon(K-1)\left[\alpha\sin\phi^* - \left(1 + \frac{\pi}{2(K-1)}\right)\sin\phi\right]\\ &= \epsilon\alpha(K-1)\left[\sqrt{1-\alpha^2y^2} - \epsilon\left(1+\frac{\pi}{2(K-1)}\right)\sqrt{2-\alpha^2\epsilon^2}\right]\end{aligned}$$

Note that for $y > 0$:

$$\alpha y = \frac{1}{\sqrt{(x/y)^2 + (K-1)}} = \frac{1}{\sqrt{(1+\epsilon/y)^2 + (K-1)}} \leq \frac{1}{\sqrt{K}} \tag{84}$$

For $y = 0$, $\alpha y = 0 < \sqrt{1/K}$. So we have $\sqrt{1-\alpha^2 y^2} \geq \sqrt{1-1/K}$. And $\sqrt{2-\alpha^2\epsilon^2} \leq \sqrt{2}$. Therefore $f_3 \geq \epsilon\alpha(K-1)(C_1 - \epsilon C_2)$ 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 5 (Reparametrization) *Denote $\epsilon = x - y > 0$. The terms αx, αy and $\alpha\epsilon$ involved in the trigometric functions in Eqn. 69 has the following parameterization:*

$$\alpha \begin{bmatrix} y \\ x \\ \epsilon \end{bmatrix} = \frac{1}{K} \begin{bmatrix} \beta - \beta_2 \\ \beta + (K-1)\beta_2 \\ K\beta_2 \end{bmatrix} \tag{85}$$

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. 69 only depend on the single variable β. In particular, the following relationship is useful:

$$\beta = \cos\theta + \sqrt{K-1}\sin\theta \tag{86}$$

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

$$\frac{1}{\alpha K}(\beta - \beta_2) = \frac{1}{K}\left(\sqrt{\frac{K}{\alpha^2} - (K-1)\epsilon^2} - \epsilon\right) = \frac{1}{K}\left(\sqrt{(Ky+\epsilon)^2} - \epsilon\right) = y \tag{87}$$

To prove Eqn. 86, 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:

$$\beta = \cos\theta + \sqrt{\cos^2\theta + K\sin^2\theta - 1} = \cos\theta + \sqrt{K-1}\sin\theta \tag{88}$$

∎

Lemma 6 *After reparametrization (Eqn. 85), $f_3(\beta, \epsilon) \geq 0$ 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. 85) to Eqn. 83 and notice that $\alpha\epsilon = \beta_2 = \beta_2(\beta)$, we could write

$$f_3 = h_1(\beta) - (\phi + (K-1)\sin\phi)\epsilon \tag{89}$$

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

$$f_{31}(\beta, \epsilon') = \phi^* - \theta - \epsilon'\phi + \epsilon'\alpha\sin\theta \tag{90}$$

$$f_{32}(\beta, \epsilon') = (K-1)(\alpha\sin\phi^* - \sin\phi)\epsilon' \tag{91}$$

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:

$$f_{31} = \epsilon'(\phi^* - \phi) + (1-\epsilon')(\phi^* - \theta) - \epsilon'\theta + \beta_2\sin\theta \geq -\epsilon'\theta + \beta_2\sin\theta \geq \beta_2\left(\sin\theta - \frac{\theta}{\beta}\right) \tag{92}$$

And it reduces to showing whether $\beta\sin\theta - \theta$ is nonnegative. Using Eqn. 86, we have:

$$f_{33}(\theta) = \beta\sin\theta - \theta = \frac{1}{2}\sin 2\theta + \sqrt{K-1}\sin^2\theta - \theta \tag{93}$$

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 3, $\theta \in [0, \theta_0)$. 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 $(y, \epsilon) = (0, 1)$. ∎

Lemma 7 *For the dynamics defined in Eqn. 69, the only critical point ($\nabla_x J = 0$ and $\nabla_y J = 0$) 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. 83). By Lemma 7, $\epsilon > \epsilon' = \beta_2/\beta > 0$ and

$$\epsilon - 1 + Ky = \frac{1}{\alpha}(\beta_2 - \alpha + \beta - \beta_2) = \frac{\beta - \alpha}{\alpha} = \frac{\beta - \beta_2/\epsilon}{\alpha} > \frac{\beta - \beta_2/\epsilon'}{\alpha} = 0 \tag{94}$$

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

$$((K-1)(\alpha \sin\phi^* - \sin\phi) + \alpha \sin\theta) = -\frac{1}{\epsilon}(\phi^* - \theta) + \phi \tag{95}$$

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

$$\begin{aligned} -\frac{2\pi}{N}\nabla_y J &= -(\pi - \phi)(\epsilon - 1 + Ky) - (\phi^* - \phi) - \phi y + ((K-1)(\alpha \sin\phi^* - \sin\phi) + \alpha \sin\theta)\, y \\ &= -(\pi - \phi)(\epsilon - 1 + Ky) - (\phi^* - \phi) - \frac{1}{\epsilon}(\phi^* - \theta)y < 0 \end{aligned} \tag{96}$$

So the current point (β, ϵ) cannot be a critical point. ∎

Lemma 8 *Any trajectory in Ω_{ϵ_0} converges to $(y, \epsilon) = (1, 0)$, following the dynamics defined in Eqn. 69.*

Proof We have Lyaponov function $V = \mathbb{E}[E]$ so that $\dot{V} = -\mathbb{E}[\nabla_{\mathbf{w}} J^\intercal \nabla_{\mathbf{w}} J] \leq -\mathbb{E}[\nabla_{\mathbf{w}} J]^\intercal \mathbb{E}[\nabla_{\mathbf{w}} J] \leq 0$. By Lemma 7, other than the optimal solution $\mathbf{w}^*$, there is no other symmetric critical point, $\nabla_{\mathbf{w}} J \neq 0$ and thus $\dot{V} < 0$. On the other hand, by Lemma 4, 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. 64). ∎

Lemma 9 *When $x = y \in (0, 1]$, the 2D dynamics (Eqn. 69) reduces to the following 1D case:*

$$-\frac{2\pi}{N}\nabla_x J = -\pi K(x - x_*) \tag{97}$$

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. ∎

Combining Lemma 8 and Lemma 9 yields Thm. 7.

7 Simulations

No theorems is provided.

8 Extension to multilayer ReLU network

Proposition 2 *For neural network with ReLU nonlinearity and using l_2 loss to match with a teacher network of the same size, the gradient inflow $\mathbf{g}_j$ for node j at layer c has the following form:*

$$\mathbf{g}_j = Q_j \sum_{j'} (Q_{j'} \mathbf{u}_{j'} - Q^*_{j'} \mathbf{u}^*_{j'}) \tag{98}$$

where Q_j and Q^*_j are N-by-N diagonal matrices. For any $k \in [c+1]$, $Q_k = \sum_{j \in [c]} w_{jk} D_j Q_j$ and similarly for Q^*_k. The gradient with respect to $\mathbf{w}_j$ (the parameters immediately under node j), is computed as:

$$\nabla_{\mathbf{w}_j} J = X_c^T D_j^\intercal \mathbf{g}_j \tag{99}$$

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

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

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

$$\mathbf{g}_k = \sum_{k'} Q_k Q_{k'} \mathbf{u}_{k'} - Q_k Q^*_{k'} \mathbf{u}^*_{k'} = Q_k \sum_{k'} Q_{k'} \mathbf{u}_{k'} - Q^*_{k'} \mathbf{u}^*_{k'} \tag{100}$$

∎