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

Yuandong Tian [1]

Abstract

In this paper, we explore theoretical properties of training a two-layered ReLU network $g(\mathbf{x}; \mathbf{w}) = \sum_{j=1}^{K} \sigma(\mathbf{w}_j^\intercal \mathbf{x})$ with centered d-dimensional spherical Gaussian input $\mathbf{x}$ (σ=ReLU). We train our network with gradient descent on $\mathbf{w}$ to mimic the output of a teacher network with the same architecture and fixed parameters $\mathbf{w}^*$. We show that its population gradient has an analytical formula, leading to interesting theoretical analysis of critical points and convergence behaviors. First, we prove that critical points outside the hyperplane spanned by the teacher parameters ("out-of-plane") are not isolated and form manifolds, and characterize in-plane critical-point-free regions for two ReLU case. On the other hand, convergence to $\mathbf{w}^*$ for one ReLU node is guaranteed with at least $(1-\epsilon)/2$ probability, if weights are initialized randomly with standard deviation upper-bounded by $O(\epsilon/\sqrt{d})$, consistent with empirical practice. For network with many ReLU nodes, we prove that an infinitesimal perturbation of weight initialization results in convergence towards $\mathbf{w}^*$ (or its permutation), a phenomenon known as spontaneous symmetric-breaking (SSB) in physics. We assume no independence of ReLU activations. Simulation verifies our findings.

1. Introduction

Despite empirical success of deep learning (e.g., 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)), it remains elusive how and why simple methods like gradient descent can solve the complicated non-convex optimization during training. In this paper, we focus on a two-layered ReLU network:

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

Here $\sigma(x) = \max(x, 0)$ is the ReLU nonlinearity. We consider the setting that a student network is optimized to minimize the l_2 distance between its prediction and the supervision provided by a teacher network of the same architecture with fixed parameters $\mathbf{w}^*$. Note that although the network prediction (Eqn. 1) is convex, when coupled with loss (e.g., l_2 loss Eqn. 2), the optimization becomes highly non-convex and has exponential number of critical points.

To analyze it, we introduce a simple analytic formula for population gradient in the case of l_2 loss, when inputs $\mathbf{x}$ are sampled from zero-mean spherical Gaussian. Using this formula, critical point and convergence analysis follow.

For critical points, we show that critical points outside the principal hyperplane (the subspace spanned by $\mathbf{w}^*$) form manifolds. We also characterize the region in the principal hyperplane that has no critical points, in two ReLU case.

We also analyze the convergence behavior under the population gradient. Using Lyapunov method (LaSalle & Lefschetz, 1961), for single ReLU case we prove that gradient descent converges to $\mathbf{w}^*$ with at least $(1-\epsilon)/2$ probability, if initialized randomly with standard deviation upper-bounded by $O(\epsilon/\sqrt{d})$, verifying common initialization techniques (Bottou, 1988; Glorot & Bengio, 2010; He et al., 2015; LeCun et al., 2012). For multiple ReLU case, when the teacher parameters $\{\mathbf{w}_j\}_{j=1}^K$ form an orthonormal basis, we prove that **(1)** a symmetric weight initialization gets stuck at a saddle point and **(2)** a particular infinitesimal perturbation of (1) leads to convergence towards $\mathbf{w}^*$ or its permutation. The behavior that the population gradient field is invariant under certain symmetry but the solution breaks it, is known as spontaneous symmetry breaking in physics. Although such behaviors are known practically, to our knowledge, we first formally characterize them in 2-layered ReLU network. Codes are available [1].

[1]Facebook AI Research. Correspondence to: Yuandong Tian <yuandong@fb.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

[1]`github.com/yuandong-tian/ICML17_ReLU`

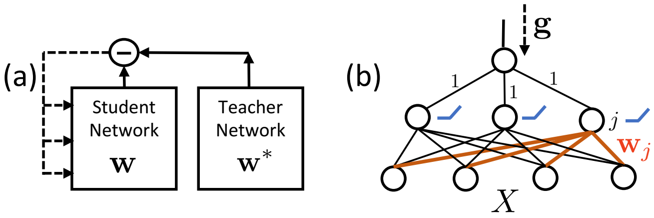

Figure 1. **(a)** We consider the student and teacher network as nonlinear neural networks with ReLU nonlinearity. The student network updates its weight $\mathbf{w}$ from the output of the teacher with fixed weights $\mathbf{w}^*$. **(b)** The 2-layered ReLU network structure (Eqn. 1) discussed in this paper. The first layer contains fixed weights of value 1, while the second layers has K ReLU nodes. Each node j has a d-dimensional weight $\mathbf{w}_j$ to be optimized. Teacher network has the same architecture as the student.

2. Related Works

For multilayer linear network, many works analyze its critical points and convergence behaviors. (Saxe et al., 2013) analyzes its dynamics of gradient descent and (Kawaguchi, 2016) shows every local minimum is global. On the other hand, very few theoretical works have been done for *nonlinear* networks. (Mei et al., 2016) shows the global convergence for a single nonlinear node whose derivatives of activation σ', σ'', σ''' are bounded and $\sigma' > 0$. Similar to our approach, (Saad & Solla, 1996) also uses the student-teacher setting and analyzes the student dynamics when the teacher's parameters $\mathbf{w}^*$ are orthonormal. However, their activation is Gaussian error function $\mathrm{erf}(x)$, and only the local behaviors of the two critical points (the initial saddle point near the origin and $\mathbf{w}^*$) are analyzed. Recent paper (Zhang et al., 2017) analyzes a similar teacher-student setting on 2-layered network when the involved function is harmonic, but it is unclear how the conclusion is generalized to ReLU case. To our knowledge, our close-form formula for 2-layered ReLU network is novel, as well as the critical point and convergence analysis. Concurrent work (Brutzkus & Globerson, 2017) proposes the same formula with a different approach, and provides similar convergence analysis for one node. For multiple nodes, they assume non-overlapping shared weights, a special case of our assumption (Sec. 6.2) that weights are cyclically symmetric and orthonormal.

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) relates 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, when independent multiplicative Bernoulli noise is applied to the activations. In practice, activations that share the input are highly dependent. Ignoring such dependency misses important behaviors, and may lead to misleading conclusions. In this paper, *no assumption of independent activations is made*. Instead, we assume input to follow spherical Gaussian distribution, which gives more realistic and interdependent activations during training.

For sigmoid activation, (Fukumizu & Amari, 2000) gives complicated conditions for a local minimum to be global when adding a new node to a 2-layered network. (Janzamin et al., 2015) gives guarantees for parameter recovery of a 2-layered network learnt with tensor decomposition. In comparison, we analyze ReLU networks trained with gradient descent, which is more popular in practice.

3. Problem Definition

Denote N as the number of samples and d as the input dimension. The N-by-d matrix X is the input data and $\mathbf{w}^*$ is the fixed parameter of the teacher network. Given the current estimation $\mathbf{w}$, we have the following l_2 loss:

$$J(\mathbf{w}) = \frac{1}{2}\|g(X;\mathbf{w}^*) - g(X;\mathbf{w})\|^2, \quad (2)$$

Here we focus on *population loss* $\mathbb{E}_X[J]$, where the input X is assumed to follow spherical Gaussian distribution $\mathcal{N}(0, I)$. Its gradient is the *population gradient* $\mathbb{E}_X[\nabla J_{\mathbf{w}}(\mathbf{w})]$ (abbrev. $\mathbb{E}[\nabla J]$). In this paper, we study critical points $\mathbb{E}[\nabla J] = 0$ and vanilla gradient dynamics $\mathbf{w}^{t+1} = \mathbf{w}^t - \eta\mathbb{E}[\nabla J(\mathbf{w}^t)]$, where η is the learning rate.

4. The Analytical Formula

Properties of ReLU. ReLU nonlinearity has useful properties. We define the *gating function* $D(\mathbf{w}) \equiv \mathrm{diag}(X\mathbf{w} > 0)$ as an N-by-N binary diagonal matrix. Its l-th diagonal element is a binary variable showing whether the neuron is activated for sample l. Using this notation, $\sigma(X\mathbf{w}) = D(\mathbf{w})X\mathbf{w}$ which means $D(\mathbf{w})$ selects the output of a linear neuron, based on their activations. Note that $D(\mathbf{w})$ only depends on the direction of $\mathbf{w}$ but not its magnitude.

$D(\mathbf{w})$ is also "transparent" with respect to derivatives. For example, at differentiable regions, $\mathrm{Jacobian}_{\mathbf{w}}[\sigma(X\mathbf{w})] = \sigma'(X\mathbf{w})X = D(\mathbf{w})X$. This gives a very concise rule for gradient descent update in ReLU networks.

One ReLU node. Given the properties of ReLU, the population gradient $\mathbb{E}[\nabla J]$ can be written as:

$$\mathbb{E}[\nabla J] = \mathbb{E}_X\left[X^\intercal D(\mathbf{w})\left(D(\mathbf{w})X\mathbf{w} - D(\mathbf{w}^*)X\mathbf{w}^*\right)\right] \quad (3)$$

Intuitively, this term vanishes when $\mathbf{w} \to \mathbf{w}^*$, and should be around $\frac{N}{2}(\mathbf{w}-\mathbf{w}^*)$ if the data are evenly distributed, since roughly half of the samples are blocked. However, such an estimation fails to capture the nonlinear behavior.

If we define *Population Gating* (PG) function $F(\mathbf{e},\mathbf{w}) \equiv X^\intercal D(\mathbf{e})D(\mathbf{w})X\mathbf{w}$, then $\mathbb{E}[\nabla J]$ can be written as:

$$\mathbb{E}[\nabla J] = \mathbb{E}[F(\mathbf{w}/\|\mathbf{w}\|,\mathbf{w})] - \mathbb{E}[F(\mathbf{w}/\|\mathbf{w}\|,\mathbf{w}^*)]. \quad (4)$$

Interestingly, $F(\mathbf{e},\mathbf{w})$ has an analytic formula if the data X follow spherical Gaussian distribution:

Theorem 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 the* N*-by-*d *data matrix and* $D(\mathbf{w}) = \mathrm{diag}(X\mathbf{w}>0)$ *is a binary diagonal matrix. If* $\mathbf{x}_i \sim \mathcal{N}(0,I)$ *(and thus bias-free), then:*

$$\mathbb{E}[F(\mathbf{e},\mathbf{w})] = \frac{N}{2\pi}\left[(\pi-\theta)\mathbf{w} + \|\mathbf{w}\|\sin\theta\mathbf{e}\right] \quad (5)$$

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

See the link[2] for the proof of all theorems. Note that we do not require X to be independent between samples. Intuitively, the first *mass* term $\frac{N}{2\pi}(\pi-\theta)\mathbf{w}$ aligns with $\mathbf{w}$ and is proportional to the amount of *activated* data whose ReLU are on. When $\theta=0$, the gating function is fully on and half of the data contribute to the term; when $\theta=\pi$, the gating function is completely switched off. The gate is controlled by the angle between $\mathbf{w}$ and the *control* signal $\mathbf{e}$. The second *asymmetric* term is aligned with $\mathbf{e}$, and is proportional to the asymmetry of the activated data samples (Fig. 2).

Note that the expectation analysis smooths out ReLU and leaves only one singularity at the origin, where $\mathbb{E}[\nabla J]$ is not continuous. That is, if approaching from different directions towards $\mathbf{w}=0$, $\mathbb{E}[\nabla J]$ is different.

With the close form of F, $\mathbb{E}[\nabla J]$ also has a close form:

$$\mathbb{E}[\nabla J] = \frac{N}{2}(\mathbf{w}-\mathbf{w}^*) + \frac{N}{2\pi}\left(\theta\mathbf{w}^* - \frac{\|\mathbf{w}^*\|}{\|\mathbf{w}\|}\sin\theta\mathbf{w}\right) \quad (6)$$

where $\theta = \angle(\mathbf{w},\mathbf{w}^*) \in [0,\pi]$. The first term is from linear approximation, while the second term shows the nonlinear behavior.

For linear case, $D \equiv I$ (no gating) and thus $\nabla J \propto X^\intercal X(\mathbf{w}-\mathbf{w}^*)$. For spherical Gaussian input X, $\mathbb{E}_X[X^\intercal X] = I$ and $\mathbb{E}[\nabla J] \propto \mathbf{w}-\mathbf{w}^*$. Therefore, the dynamics has only one critical point and global convergence follows, which is consistent with its convex nature.

Extension to other distributions. From its definition, $\mathbb{E}[F(\mathbf{e},\mathbf{w})] = \mathbb{E}[X^\intercal D(\mathbf{e})D(\mathbf{w})X\mathbf{w}]$ is linear to $\|\mathbf{w}\|$, regardless of the distribution of X. On the other hand, isotropy in spherical Gaussian distribution leads to the fact that $\mathbb{E}[F(\mathbf{e},\mathbf{w})]$ only depends on angles between vectors. For other isotropic distributions, we could similarly derive:

$$\mathbb{E}[F(\mathbf{e},\mathbf{w})] = A(\theta)\mathbf{w} + \|\mathbf{w}\|B(\theta)\mathbf{e} \quad (7)$$

where $A(0)=N/2$ (gating fully on), $A(\pi)=0$ (gating fully off), and $B(0)=B(\pi)=0$ (no asymmetry when $\mathbf{w}$ and $\mathbf{e}$ are aligned). Although we focus on spherical Gaussian case, many following analysis, in particular critical point analysis, can also be applied to Eqn. 7.

[2] http://yuandong-tian.com/ssb-supp.pdf

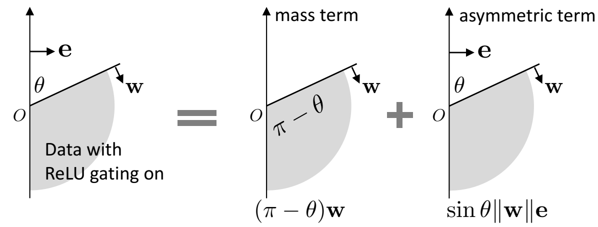

Figure 2. Decomposition of *Population Gating* (PG) function $F(\mathbf{e},\mathbf{w})$ (Eqn. 5) into mass term and asymmetric term. $F(\mathbf{e},\mathbf{w})$ is computed from the portion of data with ReLU gate on. The mass term is proportional to the amount of data, while the asymmetric term is related to the data asymmetry with respect to $\mathbf{e}$.

Multiple ReLU node. For Eqn. 1 that contains K ReLU node, we could similarly write down the population gradient with respect to $\mathbf{w}_j$ (note that $\mathbf{e}_j = \mathbf{w}_j/\|\mathbf{w}_j\|$):

$$\mathbb{E}\left[\nabla_{\mathbf{w}_j} J\right] = \sum_{j'=1}^{K}\mathbb{E}[F(\mathbf{e}_j,\mathbf{w}_{j'})] - \sum_{j'=1}^{K}\mathbb{E}\left[F(\mathbf{e}_j,\mathbf{w}^*_{j'})\right] \quad (8)$$

5. Critical Point Analysis

By solving Eqn. 8 (the *normal equation*, $\mathbb{E}\left[\nabla_{\mathbf{w}_j} J\right]=0$), we could identify all critical points of $g(\mathbf{x})$. However, it is highly nonlinear and cannot be solved easily. In this paper, we provide conditions for critical points using the structure of Eqn. 8. The case study for $K=2$ gives examples for saddle points and regions without critical points.

For convenience, we define Π_* as the *Principal Hyperplane* spanned by K ground truth weight vectors. Note that Π_* is at most K dimensional. $\{\mathbf{w}_j\}_{j=1}^K$ is said to be *in-plane*, if all $\mathbf{w}_j \in \Pi_*$. Otherwise it is *out-of-plane*.

5.1. Normal Equation

The normal equation $\{\mathbb{E}\left[\nabla_{\mathbf{w}_j} J\right]=0\}_{j=1}^K$ contain Kd scalar equations and can be written as the following:

$$YE^\intercal = B^*W^{*\intercal} \quad (9)$$

where $Y = \mathrm{diag}(\sin\Theta^\intercal\bar{\mathbf{w}} - \sin\Theta^{*\intercal}\bar{\mathbf{w}}^*) + (\pi\mathbf{1}\mathbf{1}^\intercal - \Theta^\intercal)\mathrm{diag}\bar{\mathbf{w}}$ and $B^* = \pi\mathbf{1}\mathbf{1}^\intercal - (\Theta^*)^\intercal$. Here $\theta_j^{*j'} \equiv \angle(\mathbf{w}_j,\mathbf{w}^*_{j'})$, $\theta_j^{j'} \equiv \angle(\mathbf{w}_j,\mathbf{w}_{j'})$, $\Theta = [\theta_j^i]$ (i-th row, j-th column of Θ is θ_j^i) and $\Theta^* = [\theta_j^{*i}]$.

Note that Y and B^* are both K-by-K matrices that only depend on angles and magnitudes, and hence rotational invariant. This leads to the following theorem characterizing the structure of out-of-plane critical points:

Theorem 2 *If $d \geq K+2$, then out-of-plane critical points (solutions of Eqn. 9) are non-isolated and lie in a manifold.*

The intuition is to construct a rotational matrix that is not identity matrix but keeps Π_* invariant. Such matrices form a Lie group $\mathcal{L}$ that transforms critical points to critical points. Then for any out-of-plane critical point, there is one matrix in $\mathcal{L}$ that changes at least one of its weights, yielding a non-isolated different critical point.

Note that Thm. 2 also works for any general isotropic distribution, in which $\mathbb{E}\left[F(\mathbf{e}, \mathbf{w})\right]$ has the form of Eqn. 7. This is due to the symmetry of the input X, which in turn affects the geometry of critical points. The theorem also explains why we have flat minima (Hochreiter et al., 1995; Dauphin et al., 2014) often occuring in practice.

5.2. In-Plane Normal Equation

To analyze in-plane critical points, it suffices to study gradient projections on Π_*. When $\{\mathbf{w}_j\}$ is full-rank, the projections could be achieved by right-multiplying both sides by $\{\mathbf{e}_{j'}\}$, which gives K^2 equations:

$$M(\Theta)\bar{\mathbf{w}} = M^*(\Theta, \Theta^*)\bar{\mathbf{w}}^* \tag{10}$$

This again shows decomposition of angles and magnitudes, and linearity with respect to the norms of weight vectors. Here $\bar{\mathbf{w}} = [\|\mathbf{w}_1\|, \|\mathbf{w}_2\|, \ldots, \|\mathbf{w}_K\|]^\intercal$ and similarly for $\bar{\mathbf{w}}^*$. M and M^* are K^2-by-K matrices that only depend on angles. Entries of M and M^* are:

$$\begin{aligned} m_{jj',k} &= (\pi - \theta_j^k)\cos\theta_{j'}^k + \sin\theta_j^k \cos\theta_{j'}^j & (11)\\ m^*_{jj',k} &= (\pi - \theta_j^{*k})\cos\theta_{j'}^{*k} + \sin\theta_j^{*k}\cos\theta_{j'}^j & (12) \end{aligned}$$

Here index j is the j-th column of Eqn. 9, j' is from projection vector $\mathbf{e}_{j'}$ and k is the k-th weight magnitude.

Diagnoal constraints. For "diagonal" constraints (j, j) of Eqn. 10, we have $\cos\theta_j^j = 1$ and $m_{jj,k} = h(\theta_j^k)$, $m^*_{jj,k} = h(\theta_j^{*k})$, where $h(\theta) = (\pi - \theta)\cos\theta + \sin\theta$. Therefore, we arrive at the following subset of the constraints:

$$M_r\bar{\mathbf{w}} = M_r^*\bar{\mathbf{w}}^* \tag{13}$$

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. 13 and plug it back in Eqn. 10 to check whether it is indeed a critical point. This gives necessary conditions for critical points that only depend on angles.

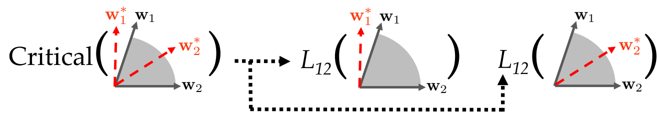

Figure 3. Separable property of critical points using $L_{jj'}$ function (Eqn. 14). Checking the criticability of $\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_1^*, \mathbf{w}_2^*\}$ can be decomposed into two subproblems, one related to $\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_1^*\}$ and the other is related to $\{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_2^*\}$.

Separable Property. Interestingly, the plugging back operation leads to conditions that are separable with respect to ground truth weight (Fig. 3). To see this, we first define the following quantity $L_{jj'}$ which is a function between *a single* (rather than K) ground truth unit weight vector $\mathbf{e}^*$ and all current unit weights $\{\mathbf{e}_l\}_{l=1}^K$:

$$L_{jj'}(\{\theta_l^*\}, \Theta) = m^*_{jj'} - \mathbf{v}^\intercal M_r^{-1}\mathbf{m}_{jj'} \tag{14}$$

where $\theta_l^* = \angle(\mathbf{e}^*, \mathbf{e}_l)$ is the angle between $\mathbf{e}^*$ and $\mathbf{e}_l$, $\mathbf{v} = \mathbf{v}(\{\theta_l^*\}) = [h(\theta_1^*), \ldots, h(\theta_K^*)]^\intercal$, and $m^*_{jj'} = (\pi - \theta_j^*)\cos\theta_{j'}^* + \sin\theta_j^*\cos\theta_{j'}^j$ (like Eqn. 12). Note that $\mathbf{v}(\{\theta_l^{*j}\})$ is the j-th column of M_r^*. Fig. 3 illustrates the case when $K = 2$. $L_{jj'}$ has the following properties:

Proposition 1 $L_{jj'}(\{\theta_l^*\}, \Theta) = 0$ *when there exists l so that $\mathbf{e}^* = \mathbf{e}_l$. In addition, $L_{jj}(\{\theta_l^*\}, \Theta) = 0$ always.*

Intuitively, $L_{jj'}$ characterizes the relative geometric relationship among $\mathbf{e}^*$ and $\{\mathbf{e}_l\}$. It is like determinant of a matrix whose columns are $\{\mathbf{e}_l\}$ and $\mathbf{e}^*$. With $L_{jj'}$, we have the following necessary conditions for critical points:

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 $1 \leq k \leq K$, then $\mathbf{w}$ cannot be a critical point.*

5.3. Case study: $K = 2$ network

In this case, M_r and M_r^* are 2-by-2 matrices. Here we discuss the case that both $\mathbf{w}_1$ and $\mathbf{w}_2$ are in Π_*.

Saddle points. When $\theta_2^1 = 0$ ($\mathbf{w}_1$ and $\mathbf{w}_2$ are collinear), $M_r = \pi\mathbf{1}\mathbf{1}^\intercal$ is singular since $\mathbf{e}_1$ and $\mathbf{e}_2$ are identical. From Eqn. 9, if $\theta_1^{*1} = \theta_1^{*2}$, i.e., they are both aligned with the bisector angle of $\mathbf{w}_1^*$ and $\mathbf{w}_2^*$, and $\pi\bar{\mathbf{w}}^\intercal\mathbf{1} = h\left(\theta_{*2}^{*1}/2\right)(\bar{\mathbf{w}}^*)^\intercal\mathbf{1}$, then the current solution is a saddle point. Note that this gives one constraint for two weight magnitudes, and thus there exist infinite solutions.

Region without critical points. We rely on the following conjecture that is verified empirically in an exhaustive manner (Sec. 7.2). It characterizes zero-crossings of a 2D function on a closed region $[0, 2\pi] \times [0, \pi]$. In comparison, in-plane 2 ReLU network has 6 parameters and is more difficult to handle: 8 for $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_1^*$ and $\mathbf{w}_2^*$, minus the rotational and scaling symmetries.

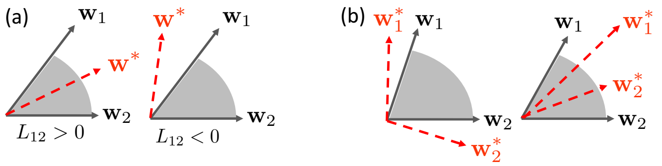

Figure 4. Critical point analysis for $K = 2$. **(a)** L_{12} changes sign when $\mathbf{w}^*$ is in/out of the cone spanned by weights $\mathbf{w}_1$ and $\mathbf{w}_2$. **(b)** Two cases that $(\mathbf{w}_1, \mathbf{w}_2)$ cannot be critical points.

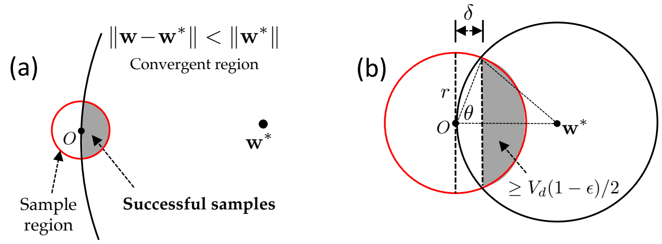

Figure 5. **(a)** Sampling strategy to maximize the probability of convergence. **(b)** Relationship between sampling range r and desired probability of success $(1 - \epsilon)/2$.

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 $L_{12} < 0$.*

This is also empirically true for L_{21}. Combined with Thm. 3, we know that (Fig. 4):

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.*

When exact one $\mathbf{w}^*$ is inside $\text{Cone}(\mathbf{w}_1, \mathbf{w}_2)$, whether $(\mathbf{w}_1, \mathbf{w}_2)$ is a critical point remains open.

6. Convergence Analysis

Application of Eqn. 5 also yields interesting convergence analysis. We focus on infinitesimal analysis, i.e., when learning rate $\eta \to 0$ and the gradient update becomes a first-order differential equation:

$$\mathrm{d}\mathbf{w}/\mathrm{d}t = -\mathbb{E}_X \left[\nabla_{\mathbf{w}} J(\mathbf{w})\right] \tag{15}$$

Then the populated objective $\mathbb{E}_X[J]$ does not increase:

$$\mathrm{d}\mathbb{E}[J]/\mathrm{d}t = -\mathbb{E}[\nabla J]^\intercal \mathrm{d}\mathbf{w}/\mathrm{d}t = -\mathbb{E}[\nabla J]^\intercal \mathbb{E}[\nabla J] \leq 0 \tag{16}$$

The goal of convergence analysis is to determine specific weight initializations $\mathbf{w}^0$ that leads to convergence to $\mathbf{w}^*$ following the gradient descent dynamics (Eqn. 15).

6.1. Single ReLU case

Using Lyapunov method (LaSalle & Lefschetz, 1961), we show that the gradient dynamics (Eqn. 15) converges to $\mathbf{w}^*$ when $\mathbf{w}^0 \in \Omega = \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}^*\| < \|\mathbf{w}^*\|\}$:

Theorem 5 *When $\mathbf{w}^0 \in \Omega = \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}^*\| < \|\mathbf{w}^*\|\}$, following the dynamics of Eqn. 15, the Lyapunov function $V(\mathbf{w}) = \frac{1}{2}\|\mathbf{w} - \mathbf{w}^*\|^2$ has $\mathrm{d}V/\mathrm{d}t < 0$ and the system is asymptotically stable and thus $\mathbf{w}^t \to \mathbf{w}^*$ when $t \to +\infty$.*

The intuition is to represent $\mathrm{d}V/\mathrm{d}t$ as a 2-by-2 bilinear form of vector $[\|\mathbf{w}\|, \|\mathbf{w}^*\|]$, and the bilinear coefficient matrix, as a function of angles, is negative definite (except for $\mathbf{w} = \mathbf{w}^*$). Note that similar approaches do not apply to regions including the origin because at the origin, the population gradient is discontinuous. Ω does not include the origin and for any initialization $\mathbf{w}^0 \in \Omega$, we could always find a slightly smaller subset $\Omega'_\delta = \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}^*\| \leq \|\mathbf{w}^*\| - \delta\}$ with $\delta > 0$ that covers $\mathbf{w}^0$, and apply Lyapunov method within. Note that the global convergence claim in (Mei et al., 2016) for l_2 loss does not apply to ReLU, since it requires $\sigma'(x) > 0$.

Random Initialization. How to sample $\mathbf{w}^0 \in \Omega$ without knowing $\mathbf{w}^*$? Uniform sampling around origin with radius $r \geq \gamma\|\mathbf{w}^*\|$ for any $\gamma > 1$ results in exponentially small success rate $(r/\|\mathbf{w}^*\|)^d \leq \gamma^{-d}$ in high-dimensional space. A better idea is to sample around the origin with very small radius (but not at $\mathbf{w} = 0$), so that Ω looks like a hyperplane near the origin, and thus almost half samples are useful (Fig. 5(a)), as shown in the following theorem:

Theorem 6 *The dynamics in Eqn. 6 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 $r \leq \epsilon\sqrt{\frac{2\pi}{d+1}}\|\mathbf{w}^*\|$.*

The idea is to lower-bound the probability of the shaded area (Fig. 5(b)). Thm. 6 gives an explanation for common initialization techniques (Glorot & Bengio, 2010; He et al., 2015; LeCun et al., 2012; Bottou, 1988) that uses random variables with $O(1/\sqrt{d})$ standard deviation.

6.2. Multiple ReLU case

For multiple ReLUs, Lyapunov method on Eqn. 8 yields no decisive conclusion. Here we focus on the symmetric property of Eqn. 8 and discuss a special case, that the teacher parameters $\{\mathbf{w}_j^*\}_{j=1}^K$ and the initial weights $\{\mathbf{w}_j^0\}_{j=1}^K$ respect the following symmetry: $\mathbf{w}_j = P_j\mathbf{w}$ and $\mathbf{w}_j^* = P_j\mathbf{w}^*$, where P_j is an orthogonal matrix whose collection $\mathcal{P} \equiv \{P_j\}_{j=1}^K$ forms a group. Without loss of generality, we set P_1 as the identity. Then from Eqn. 8 the population gradient becomes:

$$\mathbb{E}\left[\nabla_{\mathbf{w}_j} J\right] = P_j \mathbb{E}\left[\nabla_{\mathbf{w}_1} J\right] \tag{17}$$

This means that if all $\mathbf{w}_j$ and $\mathbf{w}_j^*$ are symmetric under group actions, so does their population gradients. There-

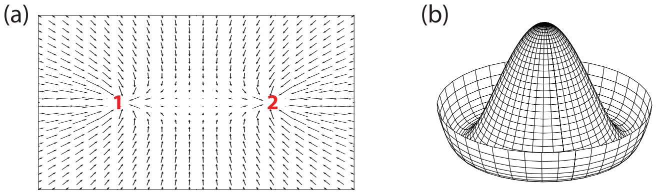

Figure 6. Spontaneous Symmetric-Breaking (SSB): Objective / gradient field is symmetric but the solution is not. **(a)** Reflection keeps the gradient field invariant, but transforms 1 to 2 and vice versa. **(b)** The Mexican hat example. Rotation keeps the objective invariant, but transforms any local minimum to a different one.

fore, the trajectory $\{\mathbf{w}^t\}$ also respects the symmetry (i.e., $P_j\mathbf{w}_1^t = \mathbf{w}_j^t$) and we only need to solve one equation for $\mathbb{E}\left[\nabla_{\mathbf{w}} J\right]$ instead of K (here $\mathbf{e} = \mathbf{w}/\|\mathbf{w}\|$):

$$\mathbb{E}\left[\nabla_{\mathbf{w}} J\right] = \sum_{j'=1}^{K} \mathbb{E}\left[F(\mathbf{e}, P_{j'}\mathbf{w})\right] - \mathbb{E}\left[F(\mathbf{e}, P_{j'}\mathbf{w}^*)\right] \quad (18)$$

Eqn. 18 has interesting properties, known as Spontaneous Symmetric-Breaking (SSB) in physics (Brading & Castellani, 2003), in which the equations of motion respect a certain symmetry but its solution breaks it (Fig. 6). In our language, despite that the population gradient field $\mathbb{E}\left[\nabla_{\mathbf{w}} J\right]$ and the objective $\mathbb{E}\left[J\right]$ are invariant to the group transformation $\mathcal{P}$, i.e., for $\mathbf{w}^* \rightarrow P_j\mathbf{w}^*$, $\mathbb{E}\left[J\right]$ and $\mathbb{E}\left[\nabla_{\mathbf{w}} J\right]$ remain the same, its solution is not ($P_j\mathbf{w} \neq \mathbf{w}$). Furthermore, since $\mathcal{P}$ is finite, as we will see, the final solution converges to different permutations of $\mathbf{w}^*$ due to infinitesimal perturbations of initialization.

To illustrate such behaviors, consider the following example in which $\{\mathbf{w}_j^*\}_{j=1}^K$ forms an orthonormal basis and under this basis, $\mathcal{P}$ is a cyclic group in which P_j circularly shifts dimension by $j-1$ (e.g., $P_2[1,2,3]^\intercal = [3,1,2]^\intercal$). In this case, if we start with $\mathbf{w}^0 = x^0\mathbf{w}^* + \sum_{j\neq 1} P_j\mathbf{w}_j^* = [x^0, y^0, \ldots, y^0]$ under the basis of $\mathbf{w}^*$, then Eqn. 18 is further reduced to a convergent 2D nonlinear dynamics and Thm. 7 holds (Please check Supplementary Materials for the associated close-form of the 2D dynamics):

Theorem 7 *For a bias-free two-layered ReLU network $g(\mathbf{x};\mathbf{w}) = \sum_j \sigma(\mathbf{w}_j^\intercal\mathbf{x})$ that takes spherical Gaussian inputs, if the teacher's parameters $\{\mathbf{w}_j^*\}$ form 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 Eqn. 8 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 $x = y = \frac{1}{\pi K}(\sqrt{K-1} - \arccos(1/\sqrt{K}) + \pi)$.*

Thm. 7 suggests that when $\mathbf{w}^0 = [y^0, x^0, \ldots, y^0]$, the system converges to $P_2\mathbf{w}^*$, etc. Since $|x^0 - y^0|$ can be arbitrarily small, a slightest perturbation around $x^0 = y^0$ leads to a different fixed point $P_j\mathbf{w}^*$ for some j. Unlike single ReLU case, the initialization in Thm. 7 is $\mathbf{w}^*$-dependent, and serves as an example for the branching behavior.

Thm. 7 also suggests that for convergence, x^0 and y^0 can be arbitrarily small, regardless of the magnitude of $\mathbf{w}^*$, showing a global convergence behavior. In comparison, (Saad & Solla, 1996) uses Gaussian error function ($\sigma = \mathrm{erf}$) as the activation, and only analyzes local behaviors near the two fixed points (origin and $\mathbf{w}^*$).

In practice, even with noisy initialization, Eqn. 18 and the original dynamics (Eqn. 8) still converge to $\mathbf{w}^*$ (and its transformations). We leave it as a conjecture, whose proof may lead to an initialization technique for 2-layered ReLU that is $\mathbf{w}^*$-independent.

Conjecture 2 *If the initialization $\mathbf{w}^0 = x^0\mathbf{w}^* + y^0\sum_{j\neq 1} P_j\mathbf{w}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is noise and $(x^0, y^0) \in \Omega$, then Eqn. 8 also converges to $\mathbf{w}^*$ with high probability.*

7. Simulations

7.1. The analytical solution to $F(\mathbf{e}, \mathbf{w})$

We verify $\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right] = \mathbb{E}\left[X^\intercal D(\mathbf{e})D(\mathbf{w})X\mathbf{w}\right]$ (Eqn. 5) with simulation. We randomly pick $\mathbf{e}$ and $\mathbf{w}$ so that their angle $\angle(\mathbf{e},\mathbf{w})$ is uniformly distributed in $[0,\pi]$. The analytical formula $\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right]$ is compared with $F(\mathbf{e},\mathbf{w})$, which is computed via sampling on the input X that follows spherical Gaussian distribution. We use relative RMS error: $err = \|\mathbb{E}\left[F(\mathbf{e},\mathbf{w})\right] - F(\mathbf{e},\mathbf{w})\| / \|F(\mathbf{e},\mathbf{w})\|$. Fig. 7(a) shows the error distribution with respect to angles. For small θ, the gating function $D(\mathbf{w})$ and $D(\mathbf{e})$ mostly overlap and give a reliable estimation. When $\theta \rightarrow \pi$, $D(\mathbf{w})$ and $D(\mathbf{e})$overlap less and the variance grows. Note that our convergence analysis operate on $\theta \in [0, \pi/2]$ and is not affected. In the following, we sample angles from $[0,\pi/2]$.

Fig. 7(a) shows that the formula is more accurate with more samples. We also examine other zero-mean distributions of X, e.g., $U[-1/2, 1/2]$. As shown in Fig. 7(d), the formula still works for large d. Note that the error is computed up to a global scale, due to different normalization constants in probability distributions. Whether Eqn. 5 applies for more general distributions remains open.

7.2. Empirical Results in critical point analysis $K = 2$

Conjecture 1 can be reduced to enumerate a complicated but 2D function via exhaustive sampling. In comparison, a full optimization of 2-ReLU network constrained on principal hyperplane Π_* involves 6 parameters (8 parameters minus 2 degrees of symmetry) and is more difficult to handle. Fig. 10 shows that empirically L_{12} has no extra zero-crossing other than $\mathbf{e}^* = \mathbf{e}_1$ or $\mathbf{e}_2$. As shown in Fig. 10(c), we have densely enumerated $\theta_2^1 \in [0,\pi]$ and $\mathbf{e}^*$ on a

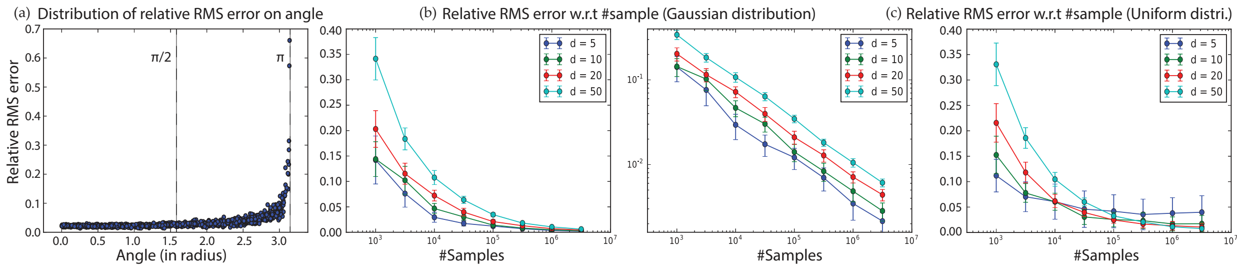

Figure 7. **(a)** Distribution of relative RMS error with respect to $\theta = \angle(\mathbf{w}, \mathbf{e})$. **(b)** Relative RMS error decreases with sample size, showing the asympototic behavior of the analytical formula (Eqn. 5). Note that the y-axis of the right plot is in log scale. **(c)** Eqn. 5 also works well when the input data X are generated by other zero-mean distribution X, e.g., uniform distribution in $[-1/2, 1/2]$.

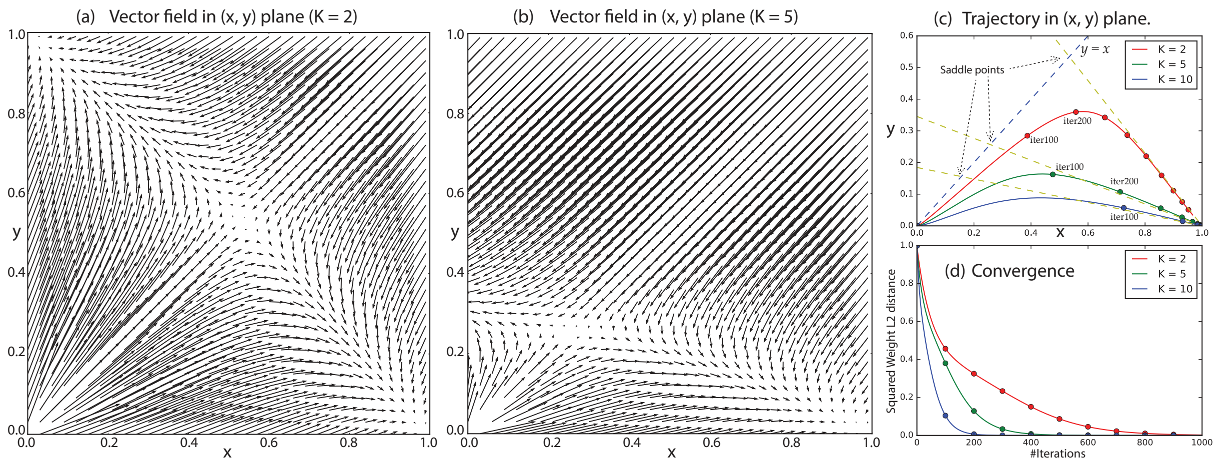

Figure 8. **(a)-(b)** Vector field in (x, y) plane following 2D dynamics (Thm. 7, See Supplementary Materials for the close-form formula) for $K = 2$ and $K = 5$. Saddle points are visible. The parameters of teacher's network are at $\mathbf{w}^* = (1, 0)$. **(c)** Trajectory in (x, y) plane for $K = 2$, $K = 5$, and $K = 10$. All trajectories start from $\mathbf{w}^0 = (10^{-3}, 0)$. Even $\mathbf{w}^0$ is aligned with $\mathbf{w}^*$, gradient descent takes detours. **(d)** Training curve. Interestingly, when K is larger the convergence is faster.

$10^4 \times 10^4$ grid without finding any counterexamples.

7.3. Convergence analysis for multiple ReLU nodes

Fig. 8(a) and (b) shows the 2D vector field in Thm 7. Fig. 8(c) shows the 2D trajectory towards convergence to the teacher's parameters $\mathbf{w}^*$. Interestingly, even when we initialize the weights as $[10^{-3}, 0]^\intercal$, whose direction is aligned with $\mathbf{w}^*$ at $[1, 0]^\intercal$, the gradient descent still takes detours to reach the destination. This is because at the beginning of optimization, all ReLU nodes explain the training error in the same way (both x and y increases); when the "obvious" component is explained, the error pushes some nodes to explain other components. Hence, specialization follows (x increases but y decreases).

Fig. 9 shows empirical convergence for $K \geq 2$, when the initialization deviates from initialization $[x, y, \ldots, y]$ in Thm. 7. Unless the deviation is large, $\mathbf{w}$ converges to $\mathbf{w}^*$. For 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. When some a_j is negative, the network fails to converge to $\mathbf{w}^*$, even when the student is initialized with the true values $\{a_j^*\}_{j=1}^K$.

8. Extension to multilayer ReLU network

A natural question is whether the proposed method can be extended to multilayer ReLU network. In this case, there is similar subtraction structure for gradient as Eqn. 3:

Proposition 2 *Denote $[c]$ as all nodes in layer c. Denote $\mathbf{u}_j^*$ and $\mathbf{u}_j$ as the output of node j at layer c of the teacher and student network, then the gradient of the parameters $\mathbf{w}_j$ immediate under node $j \in [c]$ is:*

$$\nabla_{\mathbf{w}_j} J = X_c^\intercal D_j Q_j \sum_{j' \in [c]} (Q_{j'} \mathbf{u}_{j'} - Q_{j'}^* \mathbf{u}_{j'}^*) \tag{19}$$

where X_c is the data fed into node j, Q_j and Q_j^* are N-by-N diagonal matrices. For any node $k \in [c+1]$, $Q_k = \sum_{j \in [c]} w_{jk} D_j Q_j$ and similarly for Q_k^*.

The 2-layered network in this paper is a special case with $Q_j = Q_j^* = I$. Despite the difficulty that Q_j is now depends on the weights of upper layers, and the input X_c is not necessarily Gaussian distributed, Proposition 2 gives a mathematical framework to explore the structure of gradient. For example, a similar definition of Population Gradi-

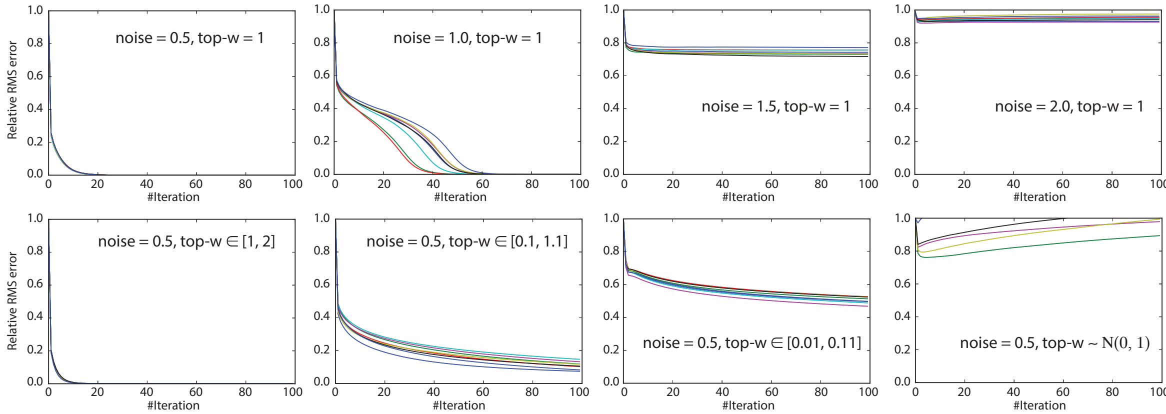

Figure 9. **Top row:** Convergence when weights are initialized with noise: $\mathbf{w}^0 = 10^{-3}\mathbf{w}^* + \epsilon$, where $\epsilon \sim N(0, 10^{-3} * noise)$. The 2-layered network converges to $\mathbf{w}^*$ until huge noise. Both teacher and student networks use $g(\mathbf{x}) = \sum_{j=1}^{K} \sigma(\mathbf{w}_j^\intercal \mathbf{x})$. Each experiment has 8 runs. **Bottom row:** Convergence for $g_2(\mathbf{x}) = \sum_{j=1}^{K} a_j \sigma(\mathbf{w}_j^\intercal \mathbf{x})$. Here we fix top weights a_j at different numbers (rather than 1). Large positive a_j corresponds to fast convergence. When $\{a_j\}$ contains mixture signs, convergence to $\mathbf{w}^*$ is not achieved.

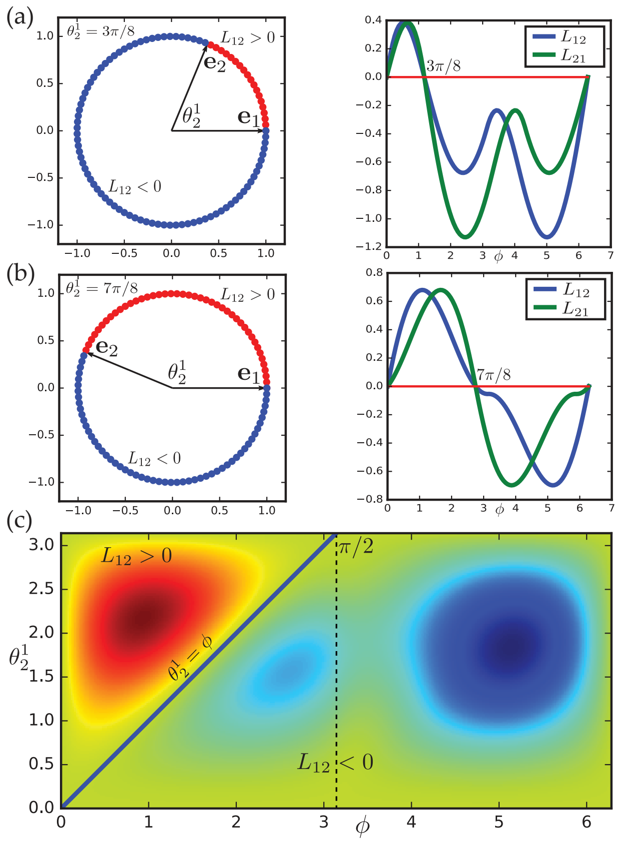

Figure 10. Quantity $L_{12}(\theta_1^*, \theta_2^*, \theta_2^1)$ and $L_{21}(\theta_1^*, \theta_2^*, \theta_2^1)$ in 2 ReLU network. We fix $\theta_2^1 = \angle(\mathbf{e}_1, \mathbf{e}_2)$ and vary $\mathbf{e}^* = [\cos\phi, \sin\phi]^\intercal$. In this case, θ_1^* and θ_2^* are both dependent variables with respect to ϕ. When $\mathbf{e}^* \in \mathrm{Cone}(\mathbf{e}_1, \mathbf{e}_2)$, L_{12} and $L_{21} > 0$, otherwise negative. There are no extra zero-crossings. **(a)-(b)** Examples: $\theta_2^1 = 3\pi/8$ and $\theta_2^1 = 7\pi/8$. **(c)** Empirical evaluation on $(\theta_2^1, \phi) \in [0, \pi] \times [0, 2\pi]$ with grid size $10^4 \times 10^4$.

ent function is possible.

9. Conclusion and Future Work

In this paper, we study the gradient descent dynamics of a 2-layered bias-free ReLU network. The network is trained using gradient descent to reproduce the output of a teacher network with fixed parameters $\mathbf{w}^*$ in the sense of l_2 norm. We propose a novel analytic formula for population gradient when the input follows zero-mean spherical Gaussian distribution. This formula leads to interesting critical point and convergence analysis. Specifically, we show that critical points out of the hyperplane spanned by $\mathbf{w}^*$ are not isolated and form manifolds. For two ReLU case, we characterize regions that contain no critical points. For convergence analysis, we show guaranteed convergence for a single ReLU case with random initialization whose standard deviation is on the order of $O(1/\sqrt{d})$. For multiple ReLU case, we show that an infinitesimal change of weight initialization leads to convergence to different optima.

Our work opens many future directions. First, Thm. 2 characterizes the non-isolating nature of critical points in the case of isotropic input distribution, which explains why often practical solutions of NN are degenerated. What if the input distribution has different symmetries? Will such symmetries determine the geometry of critical points? Second, empirically we see convergence cases that are not covered by the theorems, suggesting the conditions imposed by the theorems can be weaker. Finally, how to apply similar analysis to broader distributions and how to generalize the analysis to multiple layers are also open problems.

Acknowledgement We thank Léon Bottou, Ruoyu Sun, Jason Lee, Yann Dauphin and Nicolas Usunier for discussions and insightful suggestions.

References

Bottou, Léon. Reconnaissance de la parole par reseaux connexionnistes. In *Proceedings of Neuro Nimes 88*, pp. 197–218, Nimes, France, 1988. URL `http://leon.bottou.org/papers/bottou-88b`.

Brading, Katherine and Castellani, Elena. *Symmetries in physics: philosophical reflections*. Cambridge University Press, 2003.

Brutzkus, Alon and Globerson, Amir. Globally optimal gradient descent for a convnet with gaussian inputs. *International Conference on Machine Learning (ICML)*, 2017.

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

Choromanska, Anna, LeCun, Yann, and Arous, Gérard Ben. 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.

Dauphin, Yann N, Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, Ganguli, Surya, and Bengio, Yoshua. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in neural information processing systems*, pp. 2933–2941, 2014.

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

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

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. 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.

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

Hinton, Geoffrey, Deng, Li, Yu, Dong, Dahl, George E, Mohamed, Abdel-rahman, Jaitly, Navdeep, Senior, Andrew, Vanhoucke, Vincent, Nguyen, Patrick, Sainath, Tara N, 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.

Hochreiter, Sepp, Schmidhuber, Jürgen, et al. Simplifying neural nets by discovering flat minima. *Advances in Neural Information Processing Systems*, pp. 529–536, 1995.

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

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

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

LaSalle, J. P. and Lefschetz, S. Stability by lyapunov's second method with applications. *New York: Academic Press.*, 1961.

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

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

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

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

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

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

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

Szegedy, Christian, Liu, Wei, Jia, Yangqing, Sermanet, Pierre, Reed, Scott, Anguelov, Dragomir, Erhan, Dumitru, Vanhoucke, Vincent, and Rabinovich, Andrew.

Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

Zhang, Qiuyi, Panigrahy, Rina, and Sachdeva, Sushant. Electron-proton dynamics in deep learning. *arXiv preprint arXiv:1702.00458*, 2017.