vol 107:1–32, 2020

Dying ReLU and Initialization: Theory and Numerical Examples

Lu Lu* LU_LU_1@BROWN.EDU
Yeonjong Shin* YEONJONG_SHIN@BROWN.EDU
Division of Applied Mathematics, Brown University, Providence, RI 02912, USA

Yanhui Su SUYH@FZU.EDU.CN
College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian 350116, China

George Em Karniadakis GEORGE_KARNIADAKIS@BROWN.EDU
Division of Applied Mathematics, Brown University, Providence, RI 02912, USA
Pacific North National Laboratory, Richland, WA 99354, USA

Abstract

The dying ReLU refers to the problem when ReLU neurons become inactive and only output 0 for any input. There are many empirical and heuristic explanations of why ReLU neurons die. However, little is known about its theoretical analysis. In this paper, we rigorously prove that a deep ReLU network will eventually die in probability as the depth goes to infinite. Several methods have been proposed to alleviate the dying ReLU. Perhaps, one of the simplest treatments is to modify the initialization procedure. One common way of initializing weights and biases uses symmetric probability distributions, which suffers from the dying ReLU. We thus propose a new initialization procedure, namely, a randomized asymmetric initialization. We show that the new initialization can effectively prevent the dying ReLU. All parameters required for the new initialization are theoretically designed. Numerical examples are provided to demonstrate the effectiveness of the new initialization procedure.

Keywords: Neural network, Dying ReLU, Vanishing/Exploding gradient, Length map, Randomized asymmetric initialization

1. Introduction

The rectified linear unit (ReLU), $\max\{x, 0\}$, is one of the most successful and widely-used activation functions in deep learning (LeCun et al., 2015; Ramachandran et al., 2017; Nair and Hinton, 2010). The success of ReLU is based on its superior training performance (Glorot et al., 2011; Sun et al., 2015) over other activation functions such as the logistic sigmoid and the hyperbolic tangent (Glorot and Bengio, 2010; LeCun et al., 1998). The ReLU has been used in various applications including image classification (Krizhevsky et al., 2012; Szegedy et al., 2015), natural language processes (Maas et al., 2013), speech recognition (Hinton et al., 2012), and game intelligence (Silver et al., 2016), to name a few.

The use of gradient-based optimization is inevitable in training deep neural networks. It has been widely known that the deeper a neural network is, the harder it is to train (Srivastava et al., 2015; Du et al., 2018a). A fundamental difficulty in training deep neural networks is the vanishing and exploding gradient problem (Poole et al., 2016; Hanin, 2018; Chen et al., 2018). The dying ReLU is a kind of vanishing gradient, which refers to a problem when ReLU neurons become in-

* Lu Lu and Yeonjong Shin contributed equally to this work.

active and only output 0 for any input. It has been known as one of the obstacles in training deep ReLU neural networks (Trottier et al., 2017; Agarap, 2018). To overcome this problem, a number of methods have been proposed. Broadly speaking, these can be categorized into three general approaches. One approach modifies the network architectures. This includes but not limited to the changes in the number of layers, the number of neurons, network connections, and activation functions. In particular, many activation functions have been proposed to replace the ReLU (Maas et al., 2013; He et al., 2015; Clevert et al., 2015; Klambauer et al., 2017). However, the performance of other activation functions varies on different tasks and data sets (Ramachandran et al., 2017) and it typically requires a parameter to be turned. Thus, the ReLU remains one of the popular activation functions due to its simplicity and reliability. Another approach introduces additional training steps. This includes several normalization techniques (Ioffe and Szegedy, 2015; Salimans and Kingma, 2016; Ulyanov et al., 2016; Ba et al., 2016; Wu and He, 2018) and dropout (Srivastava et al., 2014). One of the most successful normalization techniques is the batch normalization (Ioffe and Szegedy, 2015). It is a technique that inserts layers into the deep neural network that transform the output for the batch to be zero mean unit variance. However, batch normalization increases by 30% the computational overhead to each iteration (Mishkin and Matas, 2016). The third approach modifies only weights and biases initialization procedure without changing any network architectures or introducing additional training steps (LeCun et al., 1998; Glorot and Bengio, 2010; He et al., 2015; Saxe et al., 2014; Mishkin and Matas, 2016). The third approach is the topic of our work presented in this paper.

The intriguing ability of gradient-based optimization is perhaps one of the major contributors to the success of deep learning. Training deep neural networks using gradient-based optimization fall into the noncovex nonsmooth optimization. Since a gradient-based method is either a first- or a second-order method, and once converged, the optimizer is either a local minimum or a saddle point. The authors of (Fukumizu and Amari, 2000) proved that the existence of local minima poses a serious problem in training neural networks. Many researchers have been putting immense efforts to mathematically understand the gradient method and its ability to solve nonconvex nonsmooth problems. Under various assumptions, especially on the landscape, many results claim that the gradient method can find a global minimum, can escape saddle points, and can avoid spurious local minima (Lee et al., 2016; Amari et al., 2006; Ge et al., 2015, 2016; Zhou and Liang, 2017; Wu et al., 2018; Yun et al., 2018; Du et al., 2017, 2018b,a; Jin et al., 2017). However, these assumptions do not always hold and are provably false for deep neural networks (Safran and Shamir, 2018; Kawaguchi, 2016; Arora et al., 2018). This further limits our understanding on what contributes to the success of the deep neural networks. Often, theoretical conditions are impossible to be met in practice.

Where to start the optimization process plays a critical role in training and has a significant effect on the trained result (Nesterov, 2013). This paper focuses on a particular kind of bad local minima due to a bad initialization. Such a bad local minimum causes the dying ReLU. Specifically, we consider the worst case of dying ReLU, where the entire network dies, i.e., the network becomes a constant function. We refer this as *the dying ReLU neural networks* (NNs). This phenomenon could be well illustrated by a simple example. Suppose $f(x) = |x|$ is a target function we want to approximate using a ReLU network. Since $|x| = \text{ReLU}(x) + \text{ReLU}(-x)$, a 2-layer ReLU network of width 2 can exactly represent $|x|$. However, when we train a deep ReLU network, we frequently observe that the network is collapsed. This trained result is shown in Fig. 1. Our 1,000 independent simulations show that there is a high probability (more than 90%) for the deep ReLU network to

collapse to a constant function. In this example, we employ a 10-layer ReLU network of width 2 which should perfectly recover $f(x) = |x|$.

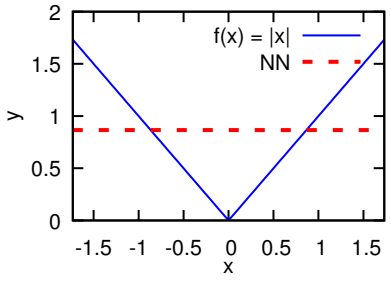


Figure 1: An approximation result for $f(x) = |x|$ using a 10-layer ReLU neural network of width 2. Among 1,000 independent simulations, this trained result is obtained with more than 90% probability. One of the most popular initialization procedures (He et al., 2015) is employed.

Almost all common initialization schemes in training deep neural networks use symmetric probability distributions around 0. For example, zero mean uniform distributions and zero mean normal distributions were proposed and used in (LeCun et al., 1998; Glorot and Bengio, 2010; He et al., 2015). We show that when weights and biases are initialized from symmetric probability distributions around 0, the dying ReLU NNs occurs in probability as the number of depth goes to infinite. To the best of our knowledge, this is the first theoretical work on the dying ReLU. This result explains why training extremely deep networks is challenging. Furthermore, it says that the dying ReLU is inevitable as long as the network is deep enough. Also, our result implies that it is the network architecture that decides whether an initialization procedure is good or bad. Our analysis reveals that a specific network architecture can avoid the dying ReLU NNs with high probability. That is, for any $\delta > 0$, when a symmetric initialization is used and $L = \Omega(\log_2(N/\delta))$ is satisfied where L is the number of depth and N is the number of width at each layer, with probability $1 - \delta$, the dying ReLU NNs will not happen.

Although there are other approaches to avoid the dying ReLU, we aim to overcome it without changing any network architectures or introducing additional training steps like normalizations. Perhaps, changing the initialization procedure might be one of the simplest treatments among others. We thus propose a new initialization procedure, namely, a randomized asymmetric initialization (RAI). The new initialization is designed to directly overcome the dying ReLU, while having similar generalization performance to the He initialization He et al. (2015). We show that our initialization has a smaller upper bound of the probability of the dying ReLU NNs. All parameters used in our initialization are theoretically chosen to avoid the exploding gradient problem. This is done by the second moment analysis where we derive the expected length map relations between layers (He et al., 2015; Hanin, 2018; Poole et al., 2016).

The rest of the paper is organized as follows. After setting up notation and terminology in Section 2, we present the main theoretical results in Section 3. In Section 4, upon introducing a randomized asymmetric initialization, we discuss its theoretical properties. Numerical examples are provided in Section 5, before the conclusion in Section 6.

2. Mathematical Setup

Let $\mathcal{N}^L : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^{d_{\text{out}}}$ be a feed-forward neural network with L layers and N_ℓ neurons in the ℓ-th layer ($N_0 = d_{\text{in}}$, $N_L = d_{\text{out}}$). Let us denote the weight matrix and bias vector in the ℓ-th layer by $\boldsymbol{W}^\ell \in \mathbf{R}^{N_\ell \times N_{\ell-1}}$ and $\mathbf{b}^\ell \in \mathbb{R}^{N_\ell}$, respectively. Given an activation function ϕ which is

applied element-wise, the feed-forward neural network is recursively defined as follows: $\mathcal{N}^1(\mathbf{x}) = \boldsymbol{W}^1\mathbf{x} + \boldsymbol{b}^1$ and

$$\mathcal{N}^\ell(\mathbf{x}) = \boldsymbol{W}^\ell\phi(\mathcal{N}^{\ell-1}(\mathbf{x})) + \boldsymbol{b}^\ell \in \mathbb{R}^{N_\ell}, \qquad \text{for} \quad 2 \le \ell \le L. \tag{1}$$

Here $\mathcal{N}^L$ is called a L-layer neural network or a $(L-1)$-hidden layer neural network. In this paper, the rectified linear unit (ReLU) activation function is employed, i.e.,

$$\phi(\mathbf{x}) = \text{ReLU}(\mathbf{x}) := (\max\{x_1, 0\}, \cdots, \max\{x_{\text{fan-in}}, 0\}), \quad \text{where} \quad \mathbf{x} = (x_1, \cdots, x_{\text{fan-in}}).$$

Let $\boldsymbol{\theta}_L = \{\boldsymbol{W}^\ell, \boldsymbol{b}^\ell\}_{1\le\ell\le L}$ be the set of all weight matrices and bias vectors. Let $\mathcal{T}_m = \{\mathbf{x}_i, y_i\}_{1\le i\le m}$ be the set of m training data and let $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^m$ be the training input data. We assume that $\mathcal{D} \subset B_r(0)$ for some $r > 0$. Given $\mathcal{T}_m$, in order to train $\boldsymbol{\theta}_L$, we consider the standard loss function $\mathcal{L}(\boldsymbol{\theta}_L, \mathcal{T}_m)$:

$$\mathcal{L}(\boldsymbol{\theta}_L, \mathcal{T}_m) = \frac{1}{m}\sum_{(\mathbf{x},y)\in\mathcal{T}_m} \ell(\mathcal{N}^L(\mathbf{x}; \boldsymbol{\theta}_L), y), \tag{2}$$

where $\ell : \mathbb{R}^{d_{\text{out}}} \times \mathbb{R}^{d_{\text{out}}} \mapsto \mathbb{R}$ is a loss criterion. In training neural networks, the gradient-based optimization is typically employed to minimize the loss $\mathcal{L}$. The first step for training would be to initialize weight matrices and bias vectors. Typically, they are initialized according to certain probability distributions. For example, uniform distributions around 0 or zero-mean normal distributions are common choices.

The dying ReLU refers to a problem when some ReLU neurons become inactive. In this paper, we focus on the worst case of dying ReLU, where the entire network dies, i.e., the network becomes a constant function. We refer this as the dying ReLU neural networks. We then define two phases: (1) a network is dead before training, and (2) a network is dead after training. The phase 1 implies the phase 2, but not vice versa. When the phase 1 happens, we say *the network is born dead* (BD).

3. Theoretical analysis

In this section, we present a theoretical analysis of the dying ReLU neural networks. We show that a deep ReLU network will eventually be BD in probability as the number of depth L goes to infinity.

Theorem 1 *Let $\mathcal{N}^L(\boldsymbol{x})$ be a ReLU neural network with L layers, each having N neurons. Suppose that all weights and biases are randomly initialized from probability distributions, which satisfy*

$$P\left(\boldsymbol{W}_j^\ell \in \mathbb{R}_-^N, \boldsymbol{b}_j^\ell < 0\right) \ge p > 0, \qquad \forall 1 \le j \le N, \tag{3}$$

for some constant $p > 0$, where $\boldsymbol{W}_j^\ell$ is the j-th row of the ℓ-th layer weight matrix and $\boldsymbol{b}_j^\ell$ is the j-th component of the ℓ-th layer bias vector. Then

$$\lim_{L\to\infty} P\left(\mathcal{N}^L(\boldsymbol{x}; \boldsymbol{\theta}_L) \text{ is born dead in } \mathcal{D}\right) = 1,$$

where $\mathcal{D}$ is the training input data.

Proof The proof can be found in Appendix A. ■

We remark that Equation 3 is a very mild condition and it can be satisfied in many cases. For example, when symmetric probability distributions around 0 are employed, the condition is met with $p = 2^{-N-1}$. Theorem 1 implies that the fully connected ReLU network will be dead at the initialization as long as the network is deep enough. This explains theoretically why training a very deep network is hard.

Theorem 1 shows that the ReLU network asymptotically will be dead. Thus, we are now concerned with the convergence behavior of the probability of NNs being BD. Since almost all common initialization procedures use symmetric probability distributions around 0, we derive an upper bound of the born dead probability (BDP) for symmetric initialization.

Theorem 2 *Let $\mathcal{N}^L(\boldsymbol{x})$ be a ReLU neural network with L layers, each having $N_1, \cdots, N_L$ neurons. Suppose that all weights are independently initialized from symmetric probability distributions around 0 and all biases are either drawn from a symmetric distribution or set to zero. Then*

$$P\left(\mathcal{N}^L(\boldsymbol{x};\boldsymbol{\theta}_L) \text{ is born dead in } \mathcal{D}\right) \leq 1 - \prod_{\ell=1}^{L-1}(1-(1/2)^{N_\ell}), \tag{4}$$

where $\mathcal{D}$ is the training input data. Furthermore, assuming $N_\ell = N$ for all ℓ,

$$\lim_{L\to\infty} P\left(\mathcal{N}^L(\boldsymbol{x};\boldsymbol{\theta}_L) \text{ is born dead in } \mathcal{D}\right) = 1, \qquad \lim_{N\to\infty} P\left(\mathcal{N}^L(\boldsymbol{x};\boldsymbol{\theta}_L) \text{ is born dead in } \mathcal{D}\right) = 0.$$

Proof The proof can be found in Appendix B. ■

Theorem 2 provides an upper bound of the BDP. It shows that at a fixed depth L, the network will not be BD in probability as the number of width N goes to infinite. In order to understand how this probability behaves with respect to the number of width and depth, a lower bound is needed. We thus provide a lower bound of the BDP of ReLU NNs at $d_{\text{in}} = 1$.

Theorem 3 *Let $\mathcal{N}^L(\boldsymbol{x})$ be a bias-free ReLU neural network with $L \geq 2$ layers, each having N neurons at $d_{in} = 1$. Suppose that all weights are independently initialized from continuous symmetric probability distributions around 0, which satisfies*

$$P(\langle \boldsymbol{W}_j^\ell, \boldsymbol{v}_1\rangle > 0, \langle \boldsymbol{W}_j^\ell, \boldsymbol{v}_2\rangle < 0 | \boldsymbol{v}_1, \boldsymbol{v}_2) \leq \frac{1}{4}, \quad \boldsymbol{0} \neq \boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}_+^N, \quad \forall 1 \leq j \leq N,$$

where $\boldsymbol{W}_j^\ell$ is the j-th row of the ℓ-th layer weight matrix. Then

$$p_{low}(L,N) \leq P\left(\mathcal{N}^L(\boldsymbol{x};\boldsymbol{\theta}_L) \text{ is born dead in } \mathcal{D}\right) \leq 1 - \prod_{\ell=1}^{L-1}(1-(1/2)^N),$$

where $a_1 = 1-(1/2)^N$, $a_2 = 1-(1/2)^{N-1} - (N-1)(1/4)^N$, and

$$p_{low}(L,N) = 1 - a_1^{L-2} + \frac{(1-2^{-N+1})(1-2^{-N})}{1+(N-1)2^{-N}}(-a_1^{L-2} + a_2^{L-2}).$$

Proof The proof can be found in Appendix C. ■

Theorem 3 reveals that the BDP behavior depends on the network architecture. In Fig. 2, we plot the BDP with respect to increasing the number of layers at varying width from $N = 2$ to $N = 5$. A bias-free ReLU feed-forward NN with $d_{\text{in}} = 1$ is employed with weights randomly initialized from symmetric distributions. The results of one million independent simulations are used to calculate each probability estimation. Numerical estimations are shown as symbols. The upper and lower bounds from Theorem 3 are also plotted with dash and dash-dot lines, respectively. We see that when the NN gets narrower, the probability of NN being BD grows faster as the depth increases. Also, at a fixed width N, the BDP grows as the number of layer increases. This is expected by Theorems 1 and 3.

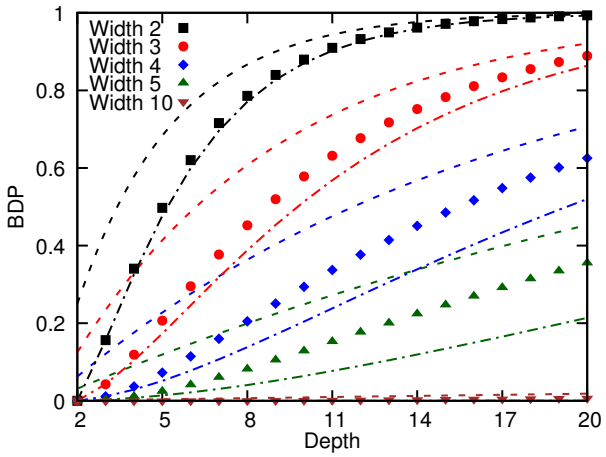


Figure 2: Probability of a ReLU NN to be born dead as a function of the number of layers for different widths. The dash lines represent the upper and lower bounds from Theorem 3. The symbols represent our numerical estimations. Similar colors correspond to the same width.

Once the network is BD, we have no hope to train the network successfully. Here we provide a formal statement of the consequence of the network being BD.

Theorem 4 *Suppose that the feed-forward ReLU neural network is BD. Then, for any loss function $\mathcal{L}$, and for any gradient based method, the ReLU network is optimized to be a constant function, which minimizes the loss.*

Proof The proof can be found in Appendix D. ■

Theorem 4 implies that no matter what gradient-based optimizers are employed including stochastic gradient desecent (SGD), SGD-Nesterov (Sutskever et al., 2013), AdaGrad (Duchi et al., 2011), AdaDelta (Zeiler, 2012), RMSProp (Hinton, 2014), Adam (Kingma and Ba, 2015), BFGS (Nocedal and Wright, 2006), L-BFGS (Byrd et al., 1995), the network is trained to be a constant function which minimizes the loss.

If the online-learning or the stochastic gradient method is employed, where the training data are independently drawn from a probability distribution $P_{\mathcal{D}}$, the optimized network is

$$\mathcal{N}^L(\mathbf{x}; \boldsymbol{\theta}^*) = \boldsymbol{c}^* = \underset{\boldsymbol{c} \in \mathbb{R}^{N_L}}{\operatorname{argmin}} \, \mathbb{E}\left[\ell(\boldsymbol{c}, f(\mathbf{x}))\right],$$

where the expectation $\mathbb{E}$ is taken with respect to $\mathbf{x} \sim P_{\mathcal{D}}$. For example, if L^2-loss is employed, i.e., $\ell(\mathcal{N}^L(\mathbf{x}), f(\mathbf{x})) = (\mathcal{N}^L(\mathbf{x}) - f(\mathbf{x}))^2$, the resulting network is $\mathbb{E}[f(\mathbf{x})]$. If L^1 loss is employed,

i.e., $\ell(\mathcal{N}^L(\mathbf{x}), f(\mathbf{x})) = |\mathcal{N}^L(\mathbf{x}) - f(\mathbf{x}))|$, the resulting network is the median of $f(\mathbf{x})$ with respect to $\mathbf{x} \sim P_{\mathcal{D}}$. Note that the mean absolute error (MAE) and the mean squared error (MSE) used in practice are discrete versions of L^1 and L^2 loss, respectively, if the size of minibatch is large.

When we design a neural network, we want the BDP to be small, say, less than 1% or 10%. Then, the upper bound (Equation 4) of Theorem 2 can be used for designing a specific network architecture, which has a small probability of NNs being born dead.

Corollary 5 *Suppose $N_\ell = N$ for all ℓ. For fixed depth L and $\delta > 0$, if the width N is $N = \log_2 L/\delta$, with probability exceeding $1 - \delta$, the ReLU neural network will not be initialized to be dead in $\mathcal{D}$.*

Proof This readily follows from

$$P\left(\mathcal{N}^L(\mathbf{x}; \boldsymbol{\theta}_L) \text{ is born dead in } \mathcal{D}\right) \leq 1 - (1 - 2^{-N})^{L-1} \leq 1 - (1 - (L-1)2^{-N}) \leq L2^{-N} = \delta.$$

■

As a practical guide, we constructed a diagram shown in Fig. 3 that includes both theoretical predictions and our numerical tests. We see that as the number of layers increases, the numerical tests match closer the theoretical results. It is clear from the diagram that a 10-layer NN of width 10 has a probability of dying less than 1% whereas a 10-layer NN of width 5 has a probability of dying greater than 10%; for width of three the probability is about 60%. Note that the growth rate of the maximum number of layers is exponential which is expected by Corollary 5.

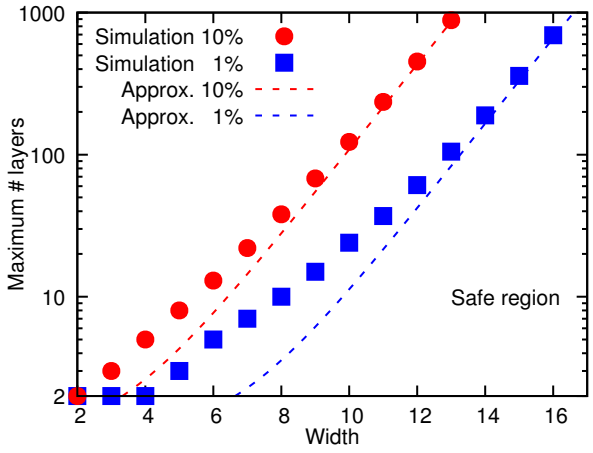


Figure 3: Diagram indicating safe operating regions for a ReLU NN. The dash lines represent Corollary 5 while the symbols represent our numerical tests. The maximum number of layers of a neural network can be used at different width to keep the probability of collapse less than 1% or 10%. The region below the blue line is the safe region when we design a neural network. As the width increases the theoretical predictions match closer with our numerical simulations.

4. Randomized Asymmetric Initialization

The so-called 'He initialization' (He et al., 2015) is perhaps one of the most popular initialization schemes in deep learning community, especially when the ReLU activation function is concerned. The effectiveness of the He initialization has been shown in many machine learning applications.

The He initialization uses mean zero normal distributions. Thus, as we discussed earlier, it suffers from the dying ReLU. We thus propose a new initialization procedure, namely, a randomized asymmetric initialization. The motivation is in twofolds. One is to mimics the He initialization so that the new scheme can produce similar generalization performance. The other is to alleviate the problem of dying ReLU neural networks.

For ease of discussion, we introduce some notation. For any vector $\boldsymbol{v} \in \mathbb{R}^{n+1}$ and $k \in \{1, \cdots, n+1\}$, we define

$$\boldsymbol{v}_{-k} = (v_1, \cdots, v_{k-1}, v_{k+1}, \cdots, v_{n+1})^T \in \mathbb{R}^n. \tag{5}$$

In order to train a L-layer neural network, we need to initialize $\boldsymbol{\theta}_L = \{\boldsymbol{W}^\ell, \boldsymbol{b}^\ell\}_{1 \le \ell \le L}$. At each layer, let $\mathbf{V}^\ell = [\boldsymbol{W}^\ell, \boldsymbol{b}^\ell] \in \mathbb{R}^{N_\ell \times (N_{\ell-1}+1)}$. We denote the j-th row of $\mathbf{V}^\ell$ by $\mathbf{V}_j^\ell = [\boldsymbol{W}_j^\ell, \boldsymbol{b}_j^\ell] \in \mathbb{R}^{N_{\ell-1}+1}$, $j = 1, \cdots, N_\ell$ where $N_0 = d_{\text{in}}$ and $N_L = d_{\text{out}}$.

4.1. Proposed initialization

We propose to initialize $\mathbf{V}^\ell$ as follows. Let P_ℓ be a probability distribution defined on $[0, M_\ell]$ for some $M_\ell > 0$ or $[0, \infty)$. Note that P_ℓ is asymmetric around 0. At the first layer of $\ell = 1$, we employ the 'He initialization' (He et al., 2015), i.e., $\mathbf{W}_{ij}^1 \sim N(0, 2/d_{\text{in}})$ and $\mathbf{b}^1 = \mathbf{0}$. For $\ell \ge 2$, and each $1 \le j \le N_\ell$, we initialize $\boldsymbol{V}_j^\ell$ as follows:

1. Randomly choose k_j^ℓ in $\{1, 2, \cdots, N_{\ell-1}, N_{\ell-1}+1\}$.
2. Initialize $(\mathbf{V}_j^\ell)_{-k_j^\ell} \sim \mathcal{N}(0, \sigma_\ell^2 \boldsymbol{I})$ and $(\mathbf{V}_j^\ell)_{k_j^\ell} \sim \mathrm{P}_\ell$.

Since an index is randomly chosen at each ℓ and j and a positive number is randomly drawn from an asymmetric probability distribution around 0, we name this new initialization a randomized asymmetric initialization. Only for the first layer, the He initialization is employed. This is because since an input could have a negative value, if the weight which corresponds to the negative input were to be initialized from P_ℓ, this could cause the dying ReLU. We note that the new initialization requires us to choose σ_ℓ^2 and P_ℓ. In Subsection 4.2, these will be theoretically determined. One could choose multiple indices in the step 1 of the new initialization. However, for simplicity, we constraint ourselves to a single index case.

We first show that this new initialization procedure results in a smaller upper bound of the BDP.

Theorem 6 *If a ReLU feed-forward neural network $\mathcal{N}^L$ with L layers, each having width $N_1, \cdots, N_L$, is initialized by the randomized asymmetric initialization, then*

$$P\left(\mathcal{N}^L(\boldsymbol{x}; \boldsymbol{\theta}_L) \text{ is born dead in } \mathcal{D}\right) \le 1 - \prod_{\ell=1}^{L-1} \left(1 - (1/2 - \gamma_\ell)^{N_\ell}\right),$$

where $\gamma_1 = 0$ and γ_j's are some constants in $(0, 0.5]$, which depend on $\{N_\ell\}_{\ell=1}^{L-1}$ and the training input data $\mathcal{D}$.

Proof The proof can be found in Appendix E. ■

When a symmetric initialization is employed, $\gamma_j = 0$ for all $1 \le j < N_L$, which results in Equation 4 of Theorem 2. Although the new initialization has a smaller upper bound compared to those by symmetric initialization, as Theorem 1 suggests, it also asymptotically suffers from the dying ReLU.

Corollary 7 *Assuming the same conditions in Theorem 6, and $N_\ell = N$ for all ℓ. Then, there exists $2 < \gamma$, which depends on N, L and the training input data $\mathcal{D}$, such that*

$$P\left(\mathcal{N}^L(\boldsymbol{x};\boldsymbol{\theta}_L) \text{ is born dead in } \mathcal{D}\right) \leq 1 - \prod_{\ell=1}^{L-1}\left(1-(1/\gamma)^N\right).$$

For fixed depth L and $\delta > 0$, if the width N is $N = \log_\gamma L/\delta$, with probability exceeding $1-\delta$, the ReLU neural network will not be initialized to be dead. Furthermore,

$$\lim_{L\to\infty} P\left(\mathcal{N}^L(\boldsymbol{x};\boldsymbol{\theta}_L) \text{ is born dead in } \mathcal{D}\right) = 1, \qquad \lim_{N\to\infty} P\left(\mathcal{N}^L(\boldsymbol{x};\boldsymbol{\theta}_L) \text{ is born dead in } \mathcal{D}\right) = 0.$$

Proof The proof is readily followed from Theorem 1, 6 and Corollary 5. ■

4.2. Second moment analysis

The proposed randomized asymmetric initialization described in Subsection 4.1 requires us to determine σ_ℓ^2 and P_ℓ. Similar to the He initialization (He et al. (2016)), we aim to properly choose initialization parameters from the length map analysis. Following the work of Poole et al. (2016), we present the analysis of a single input propagation through the deep ReLU network. To be more precise, we track the expectation of the normalized squared length of the input vector at each layer, $\mathbb{E}[q^\ell(\mathbf{x})]$, where $q^\ell(\mathbf{x}) = \frac{\|\mathcal{N}^\ell(\mathbf{x})\|^2}{N_\ell}$. The expectation $\mathbb{E}$ is taken with respect to all weights and biases.

Theorem 8 *Let P_ℓ be a probability distribution whose support is $[0, M_\ell] \subset \mathbb{R}^+$. Let $X_\ell \sim P_\ell$ have finite first and second moments, i.e., $\mu'_{\ell,i} = E[X_\ell^i] < \infty$, for $i = 1, 2$, and $\mu'_{\ell,1} \geq M_\ell/2$. Suppose the ℓ-th layer weights and biases are initialized by the randomized asymmetric initialization described in Subsection 4.1. Then for any input $\boldsymbol{x} \in \mathbb{R}^{d_{in}}$, we have*

$$\frac{\mathcal{A}_{low,\ell}}{2}\mathbb{E}[q^\ell(\boldsymbol{x})] + \sigma_{b,\ell}^2 \leq \mathbb{E}[q^{\ell+1}(\boldsymbol{x})] \leq \frac{\mathcal{A}_{upp,\ell}}{2}\mathbb{E}[q^\ell(\boldsymbol{x})] + \sigma_{b,\ell}^2,$$

where $\sigma_{b,\ell}^2 = \frac{\mu'_{\ell+1,2}+\sigma_{\ell+1}^2 N_{\ell+1}}{N_{\ell+1}+1}$, $\mathcal{A}_{low,\ell} = \frac{\sigma_{b,\ell+1}^2}{\sigma_{b,\ell}^2}\left(\frac{N_\ell \mu'_{\ell,2}+N_{\ell-1}\sigma_w^2}{N_{\ell-1}+1}\right)$, and

$$\mathcal{A}_{upp,\ell} = \frac{\sigma_{b,\ell+1}^2}{\sigma_{b,\ell}^2}\left(\frac{N_{\ell-1}\sigma_w^2 + 2\sqrt{2/\pi}N_\ell \mu'_{\ell,1}\sigma_w + 2N_\ell\mu'_{\ell,2}}{N_{\ell-1}+1}\right).$$

Proof The proof can be found in Appendix F. ■

Corollary 9 *Under the same conditions of Theroem 8, if $N_\ell = N$, $M_\ell = M$, $E[X_\ell^i] = \mu'_i$, $i = 1, 2$ for all ℓ, and $\mu'_1 \geq M/2$, we have*

$$\frac{\mathcal{A}_{low,\ell}}{2}E[q^\ell(\boldsymbol{x})] + \sigma_{b,\ell}^2 \leq E[q^{\ell+1}(\boldsymbol{x})] \leq \frac{\mathcal{A}_{upp,\ell}}{2}E[q^\ell(\boldsymbol{x})] + \sigma_{b,\ell}^2,$$

where $\sigma_{b,\ell}^2 = \frac{\mu'_2+\sigma_w^2}{N+1}$, $\mathcal{A}_{low,\ell} = \frac{N(\mu'_2+\sigma_w^2)}{N+1}$, and $\mathcal{A}_{upp,\ell} = \frac{N(\sigma_w^2+2\sqrt{2/\pi}\mu'_1\sigma_w+2\mu'_2)}{N+1}$.

Since $\sigma_{b,\ell}^2 > 0$, $\lim_{\ell\to\infty} E[q^\ell(\mathbf{x})]$ cannot be zero. In order for $\lim_{\ell\to\infty} E[q^\ell(\mathbf{x})] < \infty$, the initialization parameters $(\sigma_w^2, \mu'_{\ell,1}, \mu'_{\ell,2})$ have to be chosen to satisfy $\mathcal{A}_{upp,\ell} < 2$. Assuming $\mu'_2 < 1$, if σ_w is chosen to be

$$\sigma_w = \sqrt{2}\left(-\frac{\mu'_1}{\sqrt{\pi}} + \sqrt{\frac{\mu_1'^2}{\pi} + 1 - \mu'_2}\right), \tag{6}$$

we have $\frac{\mathcal{A}_{upp,\ell}}{2} = \frac{N}{N+1}$ which satisfies the condition.

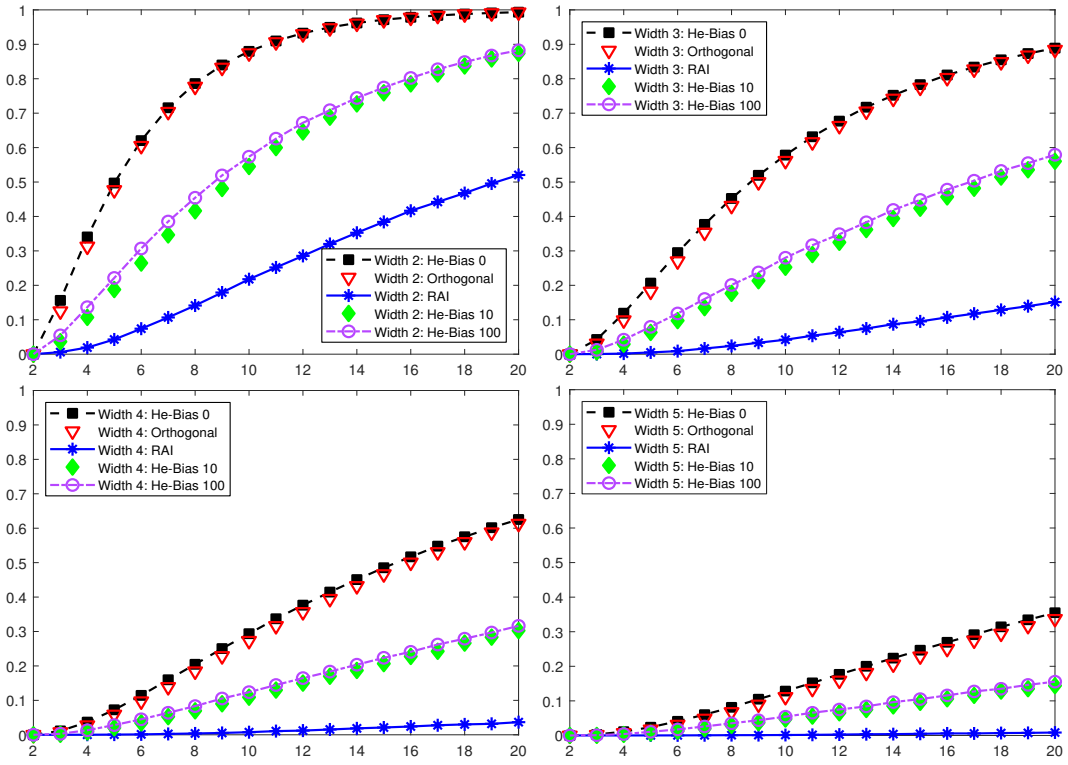


Figure 4: The BDPs are plotted with respect to increasing the number of depth L at varying width $N = 2$ (top left), $N = 3$ (top right), $N = 4$ (bottom left) and $N = 5$ (bottom right). The ReLU neural networks in $d_{\text{in}} = 1$ are employed. The square, diamond and circle symbols correspond to the He initialization (He et al., 2015) with constant bias 0, 10 and 100, respectively. The inverted triangle symbols correspond to the orthogonal initialization (Saxe et al., 2014). The asterisk symbols correspond to the proposed randomized asymmetric initialization (RAI).

4.3. Comparison against other initialization procedures

In Fig. 4, we demonstrate the probability that the network is BD by the proposed randomized asymmetric initialization (RAI) method. Here we employ $\mathrm{P} = \mathrm{Beta}(2, 1)$ and $\sigma_w = -\frac{2\sqrt{2}}{3\sqrt{\pi}} + \sqrt{1 + \frac{8}{9\pi}} \approx 0.6007$ from Equation 6. To compare against other procedures, we present the results by the He

initialization (He et al., 2015). We also present the results of existing asymmetric initialization procedures; the orthogonal (Saxe et al., 2014) and the layer-sequential unit-variance (LSUV) (Mishkin and Matas, 2016) initializations. The LSUV is the orthogonal initialization combined with rescaling of weights such that the output of each layer has unit variance. Because weight rescaling cannot make the output escape from the negative part of ReLU, it is sufficient to consider the orthogonal initialization. We see that the BDPs by the orthogonal initialization are very close to and a little lower than those by the He initialization. This implies that the orthogonal initialization cannot prevent the dying ReLU network. Furthermore, we show the results by the He initialization with positive constant bias of 10 and 100. Naively speaking, having a big positive bias will help in preventing dying ReLU neurons, as the input of each layer is pushed to be positive, although this might cause the exploding gradient problem in training. We see that the BDPs by the He with bias 10 and 100 are lower than those by the He with bias 0 and the orthogonal initialization. However, it is clearly observed that our proposed initialization (RAI) drastically drops the BDPs compared to all others. This is implied by Theorem 6.

5. Numerical examples

We demonstrate the effectiveness of the proposed randomized asymmetric initialization (RAI) in training deep ReLU networks.

Test functions include one- and two-dimensional functions of different regularities. The following test functions are employed as unknown target functions. For one dimensional cases,

$$f_1(x) = |x|, \qquad f_2(x) = x\sin(5x), \qquad f_3(x) = 1_{\{x>0\}}(x) + 0.2\sin(5x). \tag{7}$$

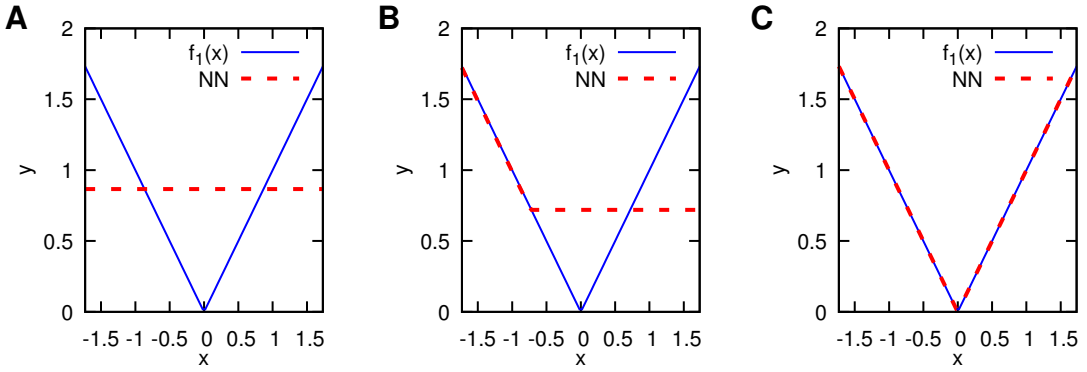
For two dimensional case,

$$f_4(x_1, x_2) = \begin{bmatrix} |x_1 + x_2| \\ |x_1 - x_2| \end{bmatrix}. \tag{8}$$

We employ the network architecture having the width of $d_{\text{in}} + d_{\text{out}}$ at all layers. Here d_{in} and d_{out} are the dimensions of the input and output, respectively. It was shown in (Hanin and Sellke, 2017) that the minimum number of width required for the universal approximation is less than or equal to $d_{\text{in}} + d_{\text{out}}$. We thus choose this specific network architecture. as it theoretically guarantees to approximate any continuous function. In all numerical examples, we employ one of the most popular first-order gradient-based optimization, `Adam` (Kingma and Ba, 2015) with the default parameters. The minibatch size is chosen to be either 64 or 128. The standard L_2-loss function $\ell(\mathcal{N}^L(\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}, y)) = (\mathcal{N}^L(\mathbf{x}; \boldsymbol{\theta}) - y)^2$ is used on 3,000 training data. The training data are randomly uniformly drawn from $[-\sqrt{3}, \sqrt{3}]^{d_{\text{in}}}$. Without changing any setups described above, we present the approximation results based on different initialization procedures. The results by our proposed randomized asymmetric initialization are referred to 'Rand. Asymmetric' or 'RAI'. Specifically, we use $\mathrm{P} = \mathrm{Beta}(2, 1)$ with σ_w defined in Equation 6. To compare against other methods, we also show the results by the He initialization (He et al., 2015). We present the ensemble of 1,000 independent training simulations.

In one dimensional examples, we employ a 10-layer ReLU network of width 2. It follows from Fig. 4 that we expect to observe at least 88% training results by the symmetric initialization and 22% training results by the RAI are collapsed. In the two dimensional example, we employ a 20-layer ReLU network of width 4. According to Fig. 4, we expect to see at least 63% training results by the symmetric initialization and 3.7% training results by the RAI are collapsed.

Fig. 5 shows all training outcomes of our 1,000 simulations for approximating $f_1(x)$ and its corresponding empirical probabilities by different initialization schemes. For this specific test function, we observe only 3 trained results shown in A, B, C. In fact, $f_1(x)$ can be represented exactly by a 2-layer ReLU network with width 2, $f_1(x) = |x| = \text{ReLU}(x) + \text{ReLU}(-x)$. It can clearly be seen that the He initialization results in the collapse with probability more than 90%. However, this probability is drastically reduced to 40% by the RAI. These probabilities are different from the probability that the network is BD. This implies that even though the network wasn't BD, there are cases that after training, the network dies. In this example, 5.6% and 18.3% of results by the symmetric and our method, respectively, are not dead at the initialization, however, they are ended up with collapsing after training. The 37.3% of training results by the RAI perfectly recover the target function $f_1(x)$, however, only 2.2% of results by the He initialization achieve this success. Also, 22.4% of the RAI and 4.2% of the He initialization produce the half-trained results which correspond to Fig. 5 (B). We remark that the only difference in training is the initialization. This implies that our new initialization does not only prevent the dying ReLU network but also improves the quality of the training in this case.

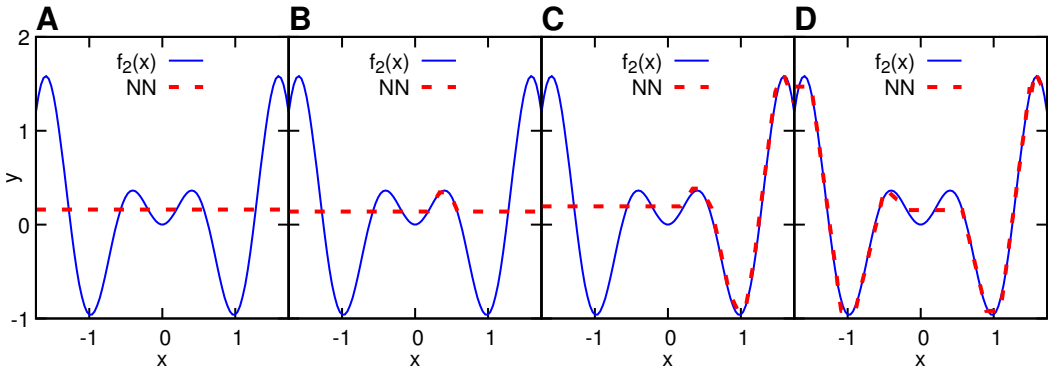


$f_1(x)$	**Collapse (A)**	**Half-Trained (B)**	**Success (C)**
Symmetric (He init.)	93.6%	4.2%	2.2%
Rand. Asymmetric	**40.3%**	**22.4%**	**37.3%**

Figure 5: The approximation results for $f_1(x)$ using a 10-layer ReLU network of width 2. For this specific test function, we observe only 3 trained results shown in A, B, C. The table shows the corresponding empirical probabilities from 1,000 independent simulations. The only difference is the initialization.

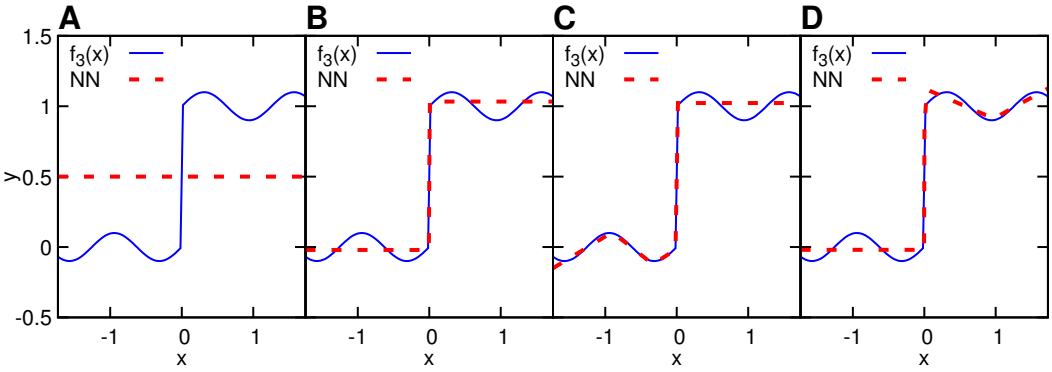
The approximation results for $f_2(x)$ are shown in Fig. 6. Note that f_2 is a C^∞ function. It can be seen that 91.9% of training results by the symmetric initialization and 29.2% of training results by the RAI are collapsed which correspond to Fig. 6 (A). This indicates that the RAI can effectively alleviate the dying ReLU. In this example, 3.9% and 7.2% of results by the symmetric and our method, respectively, are not dead at the initialization, however, they are ended up with collapsing after training. Except for the collapse, other training outcomes are not easy to be classified. Fig. 6 (B,C,D) show three training results among many others. We observe that the behavior and result of training are not easily predictable in general. However, we consistently observe partially collapsed

results after training. Such partial collapses are also observed in Fig. 6 (B,C,D). We believe that this requires more attention and postpone the study of this partial collapse to future work.



$f_2(x)$	Collapsed (A)	Not collapsed (B,C,D)
Symmetric (He init.)	91.9%	8.1%
Rand. Asymmetric	**29.2%**	**70.8%**

Figure 6: The approximation results for $f_2(x)$ using a 10-layer ReLU network of width 2. Among many trained results, four are shown. The table shows the corresponding empirical probabilities from 1,000 independent simulations. The only difference is the initialization.



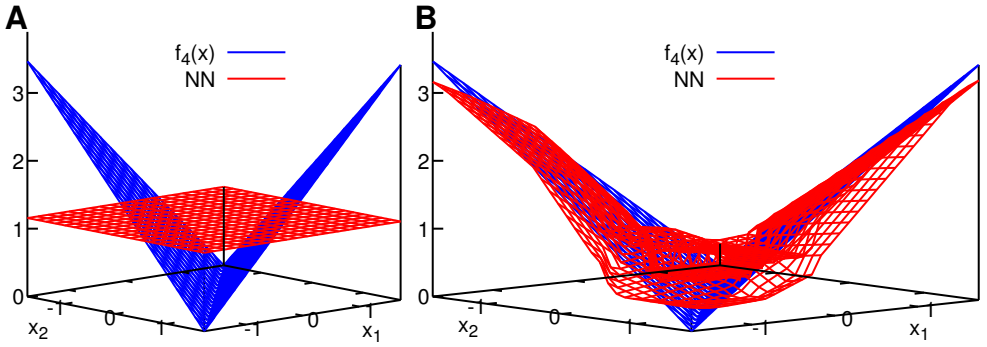
$f_3(x)$	Collapsed (A)	Not collapsed (B,C,D)
Symmetric (He init.)	93.8%	6.2%
Rand. Asymmetric	**32.6%**	**67.4%**

Figure 7: The approximation results for $f_3(x)$ using a 10-layer ReLU network of width 2. Among many trained results, four are shown. The table shows the corresponding empirical probabilities from 1,000 independent simulations. The only difference is the initialization.

Similar behavior is observed for approximating a discontinuous function $f_3(x)$. The approximation results for $f_3(x)$ and its corresponding empirical probabilities are shown in Fig. 7. We see that 93.8% of training results by the He initialization and 32.6% of training results by the RAI are

collapsed which correspond to Fig. 7 (A). In this example, the RAI drops the probability of collapsing by 60.3 percentage point. Again, this implies that the RAI can effectively avoid the dying ReLU, especially when deep and narrow ReLU networks are employed. Fig. 7 (B,C,D) show three trained results among many others. Again, we observe partially collapsed results.

Next we show the approximation result for a multi-dimensional inputs and outputs function $f_4(\mathbf{x})$ defined in Equation 8. We observe similar behavior. Fig. 8 shows some of approximation results for f_4 and its corresponding probabilities. Among 1,000 independent simulations, the collapsed results are obtained by the He initialization with 76.8% probability and by the RAI with 9.6% probability. From Fig. 4, we expect to observe at least 63% and 3.7% of results by the symmetric and the RAI to be collapsed. Thus, in this example, 13.8% and 5.9% of results by the symmetric and our method, respectively, are not dead at the initialization, however, they are ended up with Fig. 8 (A) after training. This indicates that the RAI can also effectively overcome the dying ReLU in multi-dimensional inputs and outputs tasks.



$f_4(x)$	Collapsed (A)	Not collapsed (B)
Symmetric (He init.)	76.8%	23.2%
Rand. Asymmetric	**9.6%**	**90.4%**

Figure 8: The approximation results for $f_4(\mathbf{x})$ using a 20-layer ReLU network of width 4. Among many trained results, two are shown. The table shows the corresponding empirical probabilities from 1,000 independent simulations. The only difference is the initialization.

As a last example, we demonstrate the performance of the RAI on the MNIST dataset. For the training, we employ the cross-entropy loss and the mini-batch size of 100. The networks are trained using `Adam` (Kingma and Ba, 2015) with its default values. In Fig. 9, the convergence of the test accuracy is shown with respect to the number of epochs. On the left and right, we employ the ReLU network of depth 2 and width 1024 and of depth 50 and with 10, respectively. We see that when the shallow and wide network is employed, the RAI and the He initialization show similar generalization performance (test accuracy). However, when the deep and narrow network is employed, the RAI performs better than the He initialization. This indicates that the proposed RAI not only reduces the BDP of deep networks, but also has good generalization performance.

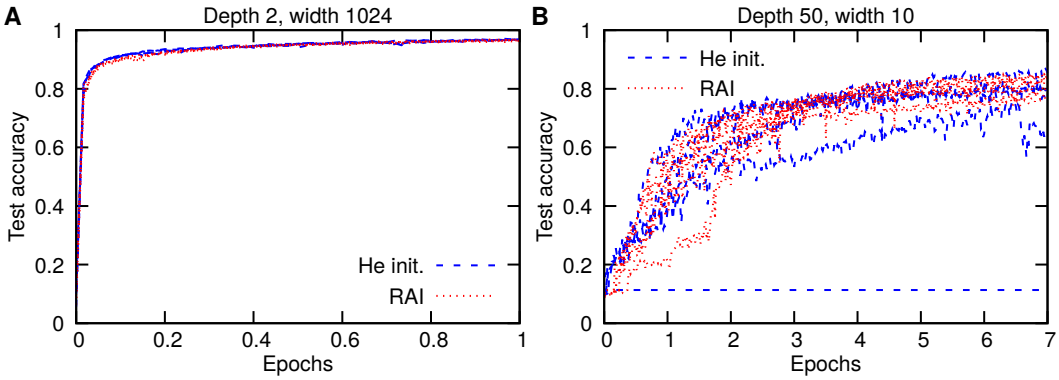


Figure 9: The test accuracy on the MNIST of five independent simulations are shown with respect to the number of epochs by the He initialization and the RAI. (Left) A shallow (depth 2, width 1024) ReLU network is employed. The He initialization and the RAI have similar performance. (Right) A deep (depth 50, width 10) ReLU network is employed. The RAI results in higher test accuracy than the He initialization.

6. Conclusion

In this paper, we establish, to the best of our knowledge, the first theoretical analysis on the dying ReLU. By focusing on the worst case of dying ReLU, we define 'the dying ReLU network' which refers to the problem when the ReLU network is dead. We categorize the dying process into two phases. One phase is the event where the ReLU network is initialized to be a constant function. We refer to this event as 'the network is born dead'. The other phase is the event where the ReLU network is collapsed after training. Certainly, the first phase implies the second, but not vice versa. We show that the probability that the network is born dead goes to 1 as the depth goes infinite. Also, we provide an upper and a lower bound of the dying probability in $d_{\text{in}} = 1$ when the standard symmetric initialization is used.

Furthermore, in order to overcome the dying ReLU networks, we propose a new initialization procedure, namely, a randomized asymmetric initialization (RAI). We show that the RAI has a smaller upper bound of the probability of NNs being born dead. By establishing the expected length map relation (second moment analysis), all parameters needed for the new method are theoretically designed. Numerical examples are provided to demonstrate the performance of our method. We observe that the RAI does not only overcome the dying ReLU but also improves the training and generalization performance.

Acknowledgments

This work received support by the DARPA EQUiPS grant N66001-15-2-4055, the AFOSR grant FA9550-17-1-0013, and the DARPA AIRA grant HR00111990025. The research of the third author was partially supported by the NSF of China 11771083 and the NSF of Fujian 2017J01556, 2016J01013.

References

A. F. Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

S. Amari, H. Park, and T. Ozeki. Singularities affect dynamics of learning in neuromanifolds. *Neural computation*, 18(5):1007–1065, 2006.

S. Arora, N. Cohen, N. Golowich, and W. Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

M. Chen, J. Pennington, and S. Schoenholz. Dynamical isometry and a mean field theory of rnns: Gating enables signal propagation in recurrent neural networks. In *International Conference on Machine Learning*, pages 872–881, 2018.

D. A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh, and B. Poczos. Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077, 2017.

S. S. Du, J. D. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018a.

S. S Du, J. D. Lee, Y. Tian, A Singh, and B Poczos. Gradient descent learns one-hidden-layer cnn: Dont be afraid of spurious local minima. In *International Conference on Machine Learning*, pages 1338–1347, 2018b.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

K. Fukumizu and S. Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural networks*, 13(3):317–327, 2000.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle pointsonline stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

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

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

X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

B. Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems*, pages 580–589, 2018.

B. Hanin and M. Sellke. Approximating continuous functions by relu nets of minimal width. *arXiv preprint arXiv:1710.11278*, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

G. Hinton. Overview of mini-batch gradient descent. http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf, 2014.

G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015.

C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1724–1732, 2017.

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

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pages 972–981, 2017.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–50. Springer, 1998.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436, 2015.

J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning*, volume 30, page 3, 2013.

D. Mishkin and J. Matas. All you need is a good init. In *International Conference on Learning Representations*, 2016.

V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2006.

B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems*, pages 3360–3368, 2016.

P. Ramachandran, B. Zoph, and Q. V. Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

I. Safran and O. Shamir. Spurious local minima are common in two-layer relu neural networks. In *International Conference on Machine Learning*, pages 4430–4438, 2018.

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

A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484, 2016.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1): 1929–1958, 2014.

R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *Advances in Neural Information Processing Systems*, pages 2377–2385, 2015.

Y. Sun, X. Wang, and X. Tang. Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2892–2900, 2015.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147, 2013.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

L. Trottier, P. Gigu, B. Chaib-draa, et al. Parametric exponential linear unit for deep convolutional neural networks. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 207–214. IEEE, 2017.

D. Ulyanov, A. Vedaldi, and V. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

C. Wu, J. Luo, and J. Lee. No spurious local minima in a two hidden unit relu network. In *International Conference on Learning Representations Workshop*, 2018.

Y. Wu and K. He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

C. Yun, S. Sra, and Jadbabaie A. Small nonlinearities in activation functions create bad local minima in neural networks. *arXiv preprint arXiv:1802.03487*, 2018.

M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Y. Zhou and Y. Liang. Critical points of neural networks: Analytical forms and landscape properties. *arXiv preprint arXiv:1710.11205*, 2017.

Appendix A. Proof of Theorem 1

The proof starts with the following lemma.

Lemma 10 *Let $\mathcal{N}^L(\boldsymbol{x})$ be a L-layer ReLU neural network with N_ℓ neurons at the ℓ-th layer. Suppose all weights are randomly independently generated from probability distributions satisfying $P(\boldsymbol{W}_j^\ell \boldsymbol{z} = \mathbf{0}) = 0$ for any nonzero vector $\boldsymbol{z} \in \mathbb{R}^{N_{\ell-1}}$ and any j-th row of $\boldsymbol{W}^\ell$. Then*

$$P(\mathcal{N}^L(\boldsymbol{x}) \text{ is born dead in } \mathcal{D}) = P(\exists\, \ell \in \{1, \dots, L-1\} \text{ such that } \phi(\mathcal{N}^\ell(\boldsymbol{x})) = \mathbf{0}\ \forall \boldsymbol{x} \in \mathcal{D}),$$

where $\mathcal{D} \subset B_r(\mathbf{0}) = \{\boldsymbol{x} \in \mathbb{R}^{d_{in}} | \|x\| < r\}$ for any $r > 0$.

Proof Suppose $\mathcal{N}^L(\mathbf{x}) = \mathcal{N}^L(\mathbf{0})$ for all $\mathbf{x} \in \mathcal{D} \subset B_r(\mathbf{0})$. Then $\phi(\mathcal{N}^{L-1}(\mathbf{x})) = \phi(\mathcal{N}^{L-1}(\mathbf{0}))$ for all $\mathbf{x} \in \mathcal{D}$. If $\phi(\mathcal{N}^{L-1}(\mathbf{x})) = \phi(\mathcal{N}^{L-1}(\mathbf{0})) = \mathbf{0}$, we are done as $\ell = L-1$. If it is not the case, there exists j in $\{1, \cdots, N_{L-1}\}$ such that for all $\mathbf{x} \in \mathcal{D}$,

$$(\mathcal{N}^{L-1}(\mathbf{x}))_j = \boldsymbol{W}_j^{L-1}\phi(\mathcal{N}^{L-2}(\mathbf{x})) + \boldsymbol{b}_j^{L-1} = \boldsymbol{W}_j^{L-1}\phi(\mathcal{N}^{L-2}(\mathbf{0})) + \boldsymbol{b}_j^{L-1} = (\mathcal{N}^{L-1}(\mathbf{0}))_j > 0.$$

Thus we have $\boldsymbol{W}_j^{L-1}\left(\phi(\mathcal{N}^{L-2}(\mathbf{x})) - \phi(\mathcal{N}^{L-2}(\mathbf{0}))\right) = 0$ for all $\mathbf{x} \in \mathcal{D}$. Let consider the following events:

$$G_{L-1} := \{\boldsymbol{W}_j^{L-1}\phi(\mathcal{N}^{L-2}(\mathbf{x})) = \boldsymbol{W}_j^{L-1}\phi(\mathcal{N}^{L-2}(\mathbf{0})), \forall \mathbf{x} \in \mathcal{D}\},$$
$$R_{L-2} := \{\phi(\mathcal{N}^{L-2}(\boldsymbol{x})) = \phi(\mathcal{N}^{L-2}(\mathbf{0})), \forall \mathbf{x} \in \mathcal{D}\}.$$

Note that $P(G_{L-1}|R_{L-2}) = 1$. Also, since $P(\boldsymbol{W}^\ell \boldsymbol{z}) = 0$ for any nonzero vector $\boldsymbol{z}$, we have $P(G_{L-1}|R_{L-2}^c) = 0$. Therefore,

$$P(G_{L-1}) = P(G_{L-1}|R_{L-2})P(R_{L-2}) + P(G_{L-1}|R_{L-2}^c)P(R_{L-2}^c) = P(R_{L-2}).$$

Thus we can focus on $\phi(\mathcal{N}^{L-2}(\mathbf{x})) = \phi(\mathcal{N}^{L-2}(\mathbf{0})), \forall \mathbf{x} \in \mathcal{D}$. If $\phi(\mathcal{N}^{L-2}(\mathbf{x})) = \phi(\mathcal{N}^{L-2}(\mathbf{0})) = \mathbf{0}$, we are done as $\ell = L-2$. If it is not the case, it follows from the similar procedure that $\phi(\mathcal{N}^{L-3}(\mathbf{x})) = \phi(\mathcal{N}^{L-3}(\mathbf{0}))$ in $\mathcal{D}$. By repeating these, we conclude that

$$P(\mathcal{N}^\ell(\mathbf{x}) \text{ dies in } \mathcal{D}) = P(\exists\, \ell \in \{1, \dots, L-1\} \text{ such that } \phi(\mathcal{N}^\ell(\mathbf{x}))) = \mathbf{0}\ \forall \mathbf{x} \in \mathcal{D}).$$

■

Proof Let $\mathcal{D} \subset B_r(0) \subset \mathbb{R}^{d_{\text{in}}}$ be a training domain where r is any positive real number. We consider a probability space $(\Omega, \mathcal{F}, P)$ where all random weight matrices and bias vectors are defined on. For every $\ell \geq 1$, let $\mathcal{F}_\ell$ be a sub-σ-algebra of $\mathcal{F}$ generated by $\{\boldsymbol{W}^j, \boldsymbol{b}^j\}_{1 \leq j \leq \ell}$. Since $\mathcal{F}_k \subset \mathcal{F}_\ell$ for $k \leq \ell$, $(\mathcal{F}_\ell)$ is a filtration. Let us define the events of our interest $\{A_\ell\}_{2 \leq \ell}$ where

$$\begin{aligned} A_\ell &= \{\mathcal{N}^\ell(\mathbf{x}) \text{ is born dead in } \mathcal{D}\} \\ &\overset{a.s.}{=} \{\exists\, j \in \{1, \dots, \ell-1\} \text{ such that } \phi(\mathcal{N}^j(\mathbf{x})) = \mathbf{0}\ \forall \mathbf{x} \in \mathcal{D}\} \end{aligned} \tag{9}$$

where the second equality is from Lemma 10. Note that A_ℓ is measurable in $\mathcal{F}_{\ell-1}$. Here $\{\boldsymbol{b}^\ell\}_{1 \leq \ell}$ could be either 0 or random vectors. Since $\mathcal{N}^1(\mathbf{x}) = \boldsymbol{W}^1\mathbf{x} + \boldsymbol{b}^1$, $P(A_1) = 0$. To calculate $P(A_\ell)$ for $\ell \geq 2$, let consider another event $C_{\ell,k}$ on which exactly $(N_\ell - k)$-components of $\phi(\mathcal{N}^\ell(\mathbf{x}))$ are zero on $\mathcal{D}$. For notational completeness, we set $C_{1,k} = \emptyset$ for $0 \leq k < N_1$ and $C_{1,N_1} = \Omega$. Then since $C_{\ell-1,0} \subset A_\ell$, we have

$$P(A_\ell) \geq P(C_{\ell-1,0}). \tag{10}$$

We want to show $\lim_{\ell \to \infty} P(C_{\ell,0}) = 1$. Since $\{C_{\ell-1,k}\}_{0 \leq k \leq N_{\ell-1}}$ is a partition of Ω, by the total law of probability, we have

$$P(C_{\ell,s}) = \sum_{k=0}^{N_{\ell-1}} P(C_{\ell,s}|C_{\ell-1,k})P(C_{\ell-1,k}),$$

where $P(C_{\ell,0}|C_{\ell-1,0}) = 1$, and $P(C_{\ell,s}|C_{\ell-1,0}) = 0, \forall s \geq 1$. Since $\boldsymbol{W}_{ij}^\ell$ and $\boldsymbol{b}_j^\ell$ are independently initialized, we have

$$P(C_{\ell,0}|C_{\ell-1,k}) = \Big(P(\langle \boldsymbol{W}_j^\ell, \phi(\mathcal{N}^{\ell-1}(\mathbf{x}))\rangle + \boldsymbol{b}_j^\ell < 0 | C_{\ell-1,k})\Big)^{N_\ell} \geq p_k^{N_\ell} > 0,$$

where the second and the third inequalities hold from the assumption. Here $p_k > 0$ does not depend on ℓ. If $\boldsymbol{W}_{ij}^\ell, \boldsymbol{b}_j^\ell$ are randomly initialized from symmetric distributions around 0,

$$P(C_{\ell,0}|C_{\ell-1,k}) \geq \begin{cases} (2^{-k})^{N_\ell} & \text{if } \boldsymbol{b}^\ell = 0 \\ (2^{-(k+1)})^{N_\ell} & \text{if } \boldsymbol{b}^\ell \text{ is generated from a symmetric distribution} \end{cases}$$

Let define a transition matrix V_ℓ of size $(N_{\ell-1}+1) \times (N_\ell + 1)$ such that the $(i+1, j+1)$-component is defined to be

$$V_\ell(i+1, j+1) = P(C_{\ell,j}|C_{\ell-1,i}), \qquad \text{where} \quad 0 \leq j \leq N_\ell \quad \text{and} \quad 0 \leq i \leq N_{\ell-1}.$$

Then given

$$\pi_1 = [P(C_{1,0}), P(C_{1,1}), \cdots, P(C_{1,N_1})] = [0, \cdots, 0, 1] \in \mathbb{R}^{N_1+1},$$

we have

$$\pi_\ell = \pi_1 V_2 \cdots V_\ell = [P(C_{\ell,0}), P(C_{\ell,1}), \cdots, P(C_{\ell,N_\ell})], \qquad \ell \geq 2.$$

Suppose $N_\ell = N$ for all $\ell \geq 1$. Note that the first row of V_ℓ is $[1, 0, \cdots, 0]$ for all $\ell \geq 2$. Thus we have the following strictly increasing sequence $\{a_\ell\}_{\ell=1}^{\infty}$:

$$a_\ell := (\pi_\ell)_1 = P(C_{\ell,0}).$$

Since $a_\ell \leq 1$, it converges, say, $\lim_{\ell\to\infty} a_\ell = a \leq 1$. Suppose $a \neq 1$, i.e., $a - 1 < 0$ and let $p_* = \min_{1\leq k\leq N} p_k$. Then since $a_{k+1} = (\pi_k V_{k+1})_1$, we have

$$a_{k+1} = a_k + \sum_{j=1}^{N} P(C_{k+1,0}|C_{k,j})(\pi_k)_{j+1} \geq a_k + (1 - a_k)(p_*^{-(N+1)})^N.$$

Thus

$$0 \leq a - a_{k+1} \leq a - a_k + (a_k - 1)(p_*^{-(N+1)})^N.$$

By taking limit on the both sides, we have

$$0 \leq (a - 1)(p_*^{-(N+1)})^N < 0,$$

which leads a contradiction. Therefore, $a = \lim_{\ell\to\infty} P(C_{\ell,0}) = 1$. It then follows from Equation 10 that

$$\lim_{\ell\to\infty} P(\mathcal{N}^\ell(\mathbf{x}) \text{ is born dead in } \mathcal{D}) \geq \lim_{\ell\to\infty} P(C_{\ell-1,0}) = 1,$$

which completes the proof. ■

Appendix B. Proof of Theorem 2

Proof Based on Lemma 10, let consider

$$\begin{aligned}
A_\ell &= \{\exists\, j \in \{1, \ldots, \ell - 1\} \text{ such that } \phi(\mathcal{N}^\ell(\mathbf{x}))) = \mathbf{0}\ \forall \mathbf{x} \in \mathcal{D}\},\\
A_\ell^c &= \{\forall 1 \leq j < \ell \text{ there exists } \mathbf{x} \in \mathcal{D} \text{ such that } \phi(\mathcal{N}^j(\mathbf{x})) \neq \mathbf{0}\},\\
\tilde{A}_{\ell,\mathbf{x}}^c &= \{\forall 1 \leq j < \ell,\ \phi(\mathcal{N}^j(\mathbf{x})) \neq \mathbf{0}\},\\
\tilde{A}_{\ell,\mathbf{x}} &= \{\exists\, j \in \{1, \ldots, \ell - 1\} \text{ such that } \phi(\mathcal{N}^j(\mathbf{x})) = \mathbf{0}\}.
\end{aligned}$$

Then if $\mathbf{x} \in \mathcal{D}$, $\tilde{A}_{\ell,\mathbf{x}}^c \subset A_\ell^c$. Thus it suffices to compute $P(\tilde{A}_{\ell,\mathbf{x}}^c)$ as

$$P(A_\ell) = 1 - P(A_\ell^c) \leq 1 - P(\tilde{A}_{\ell,\mathbf{x}}^c).$$

For $\mathbf{x} \neq \mathbf{0}$, let consider

$$U_{j,\mathbf{x}} = \{\phi(\mathcal{N}^j(\mathbf{x})) = \mathbf{0}\}, \qquad U_{j,\mathbf{x}}^c = \{\phi(\mathcal{N}^j(\mathbf{x})) \neq \mathbf{0}\}.$$

Note that $\bigcup_{1\le j<\ell} U_{j,\mathbf{x}} = \tilde{A}_{\ell,\mathbf{x}}$ and $\tilde{A}^c_{\ell,\mathbf{x}} = \bigcap_{1\le j<\ell} U^c_{j,\mathbf{x}}$. Since $P(\tilde{A}^c_{j,\mathbf{x}}|\tilde{A}_{j-1,\mathbf{x}}) = 0$ for all j, we have

$$P(\tilde{A}^c_{\ell,\mathbf{x}}) = P(\tilde{A}^c_{\ell,\mathbf{x}}|\tilde{A}^c_{\ell-1,\mathbf{x}})P(\tilde{A}^c_{\ell-1,\mathbf{x}}) = \cdots = P(\tilde{A}^c_{1,\mathbf{x}})\prod_{j=2}^{\ell} P(\tilde{A}^c_{j,\mathbf{x}}|\tilde{A}^c_{j-1,\mathbf{x}}). \tag{11}$$

Note that $P(\tilde{A}^c_{1,\mathbf{x}}) = 1$. Also, note that since the rows of $\boldsymbol{W}^\ell$ and $\boldsymbol{b}^\ell_j$ are independent,

$$P(\tilde{A}_{j,\mathbf{x}}|\tilde{A}^c_{j-1,\mathbf{x}}) = \prod_{s=1}^{N_{j-1}} P\left(\boldsymbol{W}^{j-1}_s\phi(\mathcal{N}^{j-2}(\mathbf{x})) + \boldsymbol{b}^{j-1}_s \le 0|\tilde{A}^c_{j-1,\mathbf{x}}\right). \tag{12}$$

Since the weight and biases are randomly drawn from symmetry distribution around 0 and $P\left(\boldsymbol{W}^{j-1}_s\phi(\mathcal{N}^{j-2}(\mathbf{x})) + \boldsymbol{b}^{j-1}_s\right.$ 0, we obtain

$$P\left(\boldsymbol{W}^j_s\phi(\mathcal{N}^{j-1}(\mathbf{x})) + \boldsymbol{b}^j_s \le 0|\tilde{A}^c_{j-1,\mathbf{x}}\right) = \frac{1}{2}.$$

Therefore, $P(\tilde{A}_{j,\mathbf{x}}|\tilde{A}^c_{j-1,\mathbf{x}}) = 2^{-N_{j-1}}$ and thus, $P(\tilde{A}^c_{j,\mathbf{x}}|\tilde{A}^c_{j-1,\mathbf{x}}) = 1-2^{-N_{j-1}}$. It then follows from Equation 11 that

$$P(\tilde{A}^c_{\ell,\mathbf{x}}) = \prod_{j=1}^{\ell-1}(1-2^{-N_j}),$$

which completes the proof. ■

Appendix C. Proof of Theorem 3

Proof We now assume $d_{\text{in}} = 1$, $N_\ell = N$ and without loss of generality let $\mathcal{D} \subset [-r, r]$ be a training domain for any $r > 0$. Also we assume that all weights are initialized from continuous symmetric probability distributions around 0. And the biases are set to zeros.

Since $d_{\text{in}} = 1$, for each ℓ, there exist non-negative vectors $\boldsymbol{v}_\pm \in \mathbb{R}^{N_\ell}_+$ such that

$$\phi(\mathcal{N}^\ell(x)) = \boldsymbol{v}_+\phi(x) + \boldsymbol{v}_-\phi(-x). \tag{13}$$

Let $B_{\ell,0}$ be the event where $\phi(\mathcal{N}^\ell(x)) = 0$, and let $B_{\ell,1}$ be the event where

$$\phi(\mathcal{N}^\ell(x)) = \boldsymbol{v}_+\phi(x), \quad \text{or} \quad \boldsymbol{v}_-\phi(-x), \quad \text{or} \quad \boldsymbol{v}(\phi(x) + b\phi(-x))$$

for some $\boldsymbol{v}_\pm, \boldsymbol{v} \in \mathbb{R}^{N_\ell}_+$ and $b > 0$. Let $B_{\ell,2}$ be the event where

$$\phi(\mathcal{N}^\ell(x)) = \boldsymbol{v}_+\phi(x) + \boldsymbol{v}_-\phi(-x)$$

for some linearly independent vectors $\boldsymbol{v}_\pm$. Then it can be checked that $P(B_{\ell+1,k}|B_{\ell,s}) = 0$ for all $2 \ge k > s \ge 0$. Thus, it suffices to consider $P(B_{\ell+1,k}|B_{\ell,s})$ where $0 \le k \le s \le 2$. At $\ell = 1$, since $\phi(\mathcal{N}^1(\mathbf{x})) = \phi(\boldsymbol{W}^1\mathbf{x})$ and $\mathcal{D} \subset [-r, r]$, we have

$$P(B_{1,0}) = 0, \qquad P(B_{1,1}) = 2^{-N+1}, \qquad P(B_{1,2}) = 1-2^{-N+1}.$$

For $\ell > 1$, it can be checked that $P(B_{\ell,0}|B_{\ell-1,0}) = 1$, $P(B_{\ell,0}|B_{\ell-1,1}) = 2^{-N}$, and thus $P(B_{\ell,1}|B_{\ell-1,1}) = 1 - 2^{-N}$. For $P(B_{\ell,0}|B_{\ell-1,2})$ and $P(B_{\ell,1}|B_{\ell-1,2})$, we observe the followings. In $B_{\ell,2}$, we have

$$\phi(\langle \boldsymbol{w}, \phi(\mathcal{N}^\ell(x))\rangle) = \phi(\langle \boldsymbol{w}, \boldsymbol{v}_+\rangle)\phi(x) + \phi(\langle \boldsymbol{w}, \boldsymbol{v}_-\rangle)\phi(-x),.$$

Since $\boldsymbol{v}_\pm$ are nonzero vectors in $\mathbb{R}_+^{N_\ell}$ and thus satisfy $\langle \boldsymbol{v}_+, \boldsymbol{v}_-\rangle \geq 0$, by the assumption, we obtain

$$\begin{aligned}
&P(\langle \boldsymbol{w}, \boldsymbol{v}_+\rangle < 0, \langle \boldsymbol{w}, \boldsymbol{v}_-\rangle < 0 | \boldsymbol{v}_\pm) = \frac{1}{2} - p_{\boldsymbol{v}_\pm} \geq 1/4,\\
&P(\langle \boldsymbol{w}, \boldsymbol{v}_+\rangle > 0, \langle \boldsymbol{w}, \boldsymbol{v}_-\rangle > 0 | \boldsymbol{v}_\pm) = \frac{1}{2} - p_{\boldsymbol{v}_\pm} \geq 1/4,\\
&P(\langle \boldsymbol{w}, \boldsymbol{v}_+\rangle > 0, \langle \boldsymbol{w}, \boldsymbol{v}_-\rangle < 0 | \boldsymbol{v}_\pm) = P(\langle \boldsymbol{w}, \boldsymbol{v}_+\rangle < 0, \langle \boldsymbol{w}, \boldsymbol{v}_-\rangle > 0 | \boldsymbol{v}_\pm) = p_{\boldsymbol{v}_\pm} \leq \frac{1}{4}.
\end{aligned}$$

Thus we have

$$P(B_{\ell,0}|B_{\ell-1,2}) = \mathbb{E}\left[\left(\frac{1}{2} - p_{\boldsymbol{v}_\pm}\right)^N\right] \geq (1/4)^N$$

where the expectation is taken under $p_{\boldsymbol{v}_\pm}$. For $P(B_{\ell,1}|B_{\ell-1,2})$, there are only three ways to move from $B_{\ell-1,2}$ to $B_{\ell,1}$. That is

$$\begin{aligned}
&B_{\ell,1}^{(A)}|B_{\ell-1,2} : \phi(\mathcal{N}^{\ell-1}(x)) \to \phi(\mathcal{N}^\ell(x)) = \boldsymbol{v}_+\phi(x),\\
&B_{\ell,1}^{(B)}|B_{\ell-1,2} : \phi(\mathcal{N}^{\ell-1}(x)) \to \phi(\mathcal{N}^\ell(x)) = \boldsymbol{v}_-\phi(-x),\\
&B_{\ell,1}^{(C)}|B_{\ell-1,2} : \phi(\mathcal{N}^{\ell-1}(x)) \to \phi(\mathcal{N}^\ell(x)) = \boldsymbol{v}(\phi(x) + b\phi(-x)).
\end{aligned}$$

Thus $P(B_{\ell,1}|B_{\ell-1,2}) = P(B_{\ell,1}^{(A)}|B_{\ell-1,2}) + P(B_{\ell,1}^{(B)}|B_{\ell-1,2}) + P(B_{\ell,1}^{(C)}|B_{\ell-1,2})$. Note that due to the symmetry, $P(B_{\ell,1}^{(A)}|B_{\ell-1,2}) = P(B_{\ell,1}^{(B)}|B_{\ell-1,2})$. Thus we have

$$\begin{aligned}
P(B_{\ell,1}|B_{\ell-1,2}) &= 2\left(\sum_{j=1}^{N}\binom{N}{j}\mathbb{E}\left[\left(\frac{1}{2} - p_{\boldsymbol{v}_\pm}\right)^{N-j}\left(p_{\boldsymbol{v}_\pm}\right)^j\right]\right) + \binom{N}{1}\mathbb{E}\left[\left(\frac{1}{2} - p_{\boldsymbol{v}_\pm}\right)^N\right]\\
&= 2^{-N+1} + (N-2)P(B_{\ell,0}|B_{\ell-1,2})\\
&\geq 2^{-N+1} + (N-2)4^{-N}.
\end{aligned}$$

Note that

$$P(B_{\ell,0}|B_{\ell-1,2}) + P(B_{\ell,1}|B_{\ell-1,2}) = 2^{-N+1} + (N-1)P(B_{\ell,0}|B_{\ell-1,2}).$$

Since $P(A_{\ell+1}) = P(B_{\ell,0})$ where A_ℓ is defined in Equation 9, we aim to estimate $P(B_{\ell,0})$. Let V_ℓ be the transition matrix whose $(i+1, j+1)$-component is $P(B_{\ell,j}|B_{\ell-1,i})$. Then $P(B_{\ell,0}) = \pi_1 V_2 \cdots V_\ell$ where $(\pi_1)_j = P(B_{1,j-1})$. By letting

$$\gamma_\ell = 1 + \frac{2^{-N} - P(B_{\ell,0}|B_{\ell-1,2})}{2^{-N} + (N-1)P(B_{\ell,0}|B_{\ell-1,2})}.$$

we obtain

$$V_\ell = Q_\ell D_\ell Q_\ell^{-1}, \qquad Q_\ell = \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 1/\sqrt{3} & \frac{1}{\sqrt{1+\gamma_\ell^2}} & 0 \\ 1/\sqrt{3} & \frac{\gamma_\ell}{\sqrt{1+\gamma_\ell^2}} & 1 \end{bmatrix}, \quad Q_\ell^{-1} = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ -\sqrt{1+\gamma_\ell^2} & \sqrt{1+\gamma_\ell^2} & 0 \\ -(1-\gamma_\ell) & -\gamma_\ell & 1 \end{bmatrix}$$

where $D_\ell = \text{diag}(V_\ell)$.

To find a lower bound of $P(B_{\ell,0})$, we consider the following transition matrix $\mathcal{P}$ of size 3×3 which is defined to be

$$\mathcal{P} = \begin{bmatrix} 1 & 0 & 0 \\ (1/2)^N & 1-(1/2)^N & 0 \\ (1/4)^N & (1/2)^{N-1} + (N-2)(1/4)^N & 1-(1/2)^{N-1} - (N-1)(1/4)^N \end{bmatrix}.$$

It can be checked that

$$\mathcal{P} = QDQ^{-1}, \qquad Q = \begin{bmatrix} 1/\sqrt{3} & 0 & 0 \\ 1/\sqrt{3} & \frac{1}{\sqrt{1+\gamma^2}} & 0 \\ 1/\sqrt{3} & \frac{\gamma}{\sqrt{1+\gamma^2}} & 1 \end{bmatrix}, \quad Q^{-1} = \begin{bmatrix} \sqrt{3} & 0 & 0 \\ -\sqrt{1+\gamma^2} & \sqrt{1+\gamma^2} & 0 \\ -(1-\gamma) & -\gamma & 1 \end{bmatrix}$$

where $\gamma = \frac{\mathcal{P}_{32}}{\mathcal{P}_{22}-\mathcal{P}_{33}} = \frac{2^{-N+1}+(N-2)4^{-N}}{2^{-N}+(N-1)4^{-N}} = 1 + \frac{2^{-N}-4^{-N}}{2^{-N}+(N-1)4^{-N}}$ and $D = \text{diag}(\mathcal{P})$. Thus we have

$$\mathcal{P}^\ell = \begin{bmatrix} 1 & 0 & 0 \\ 1-(\mathcal{P}_{22})^\ell & (\mathcal{P}_{22})^\ell & 0 \\ 1-(\mathcal{P}_{22})^\ell - (\gamma-1)((\mathcal{P}_{22})^\ell - (\mathcal{P}_{33})^\ell) & \gamma((\mathcal{P}_{22})^\ell - (\mathcal{P}_{33})^\ell) & (\mathcal{P}_{33})^\ell \end{bmatrix}.$$

Similarly, we obtain

$$V_2 \cdots V_{\ell+1} = \begin{bmatrix} 1 & 0 & 0 \\ 1-(\mathcal{P}_{22})^\ell & (\mathcal{P}_{22})^\ell & 0 \\ \xi_{\ell,31} & \xi_{\ell,32} & \xi_{\ell,33} \end{bmatrix}$$

where $g(x) = 1 - 2^{-N+1} - (N-1)x$, $p_\ell = (V_\ell)_{31} = P(B_{\ell,0}|B_{\ell-1,2})$, $\gamma_{\ell,i} = \gamma_i$ for $1 \le i \le \ell$, $\gamma_{\ell,\ell+1} = 1$,

$$\xi_{\ell,31} = (\mathcal{P}^\ell)_{31} - (\gamma_{\ell,1}-1)(g(p_1))^\ell + \sum_{i=1}^{\ell} (\gamma_{\ell,i} - \gamma_{\ell,i+1})(\mathcal{P}_{22})^{\ell-i} \prod_{j=1}^{i} g(p_j),$$

$$\xi_{\ell,33} = \prod_{j=1}^{\ell} g(p_j), \qquad \xi_{\ell,32} = 1 - \xi_{\ell,31} - \xi_{\ell,33}.$$

We want to show that

$$(\pi_1 \mathcal{P}^\ell)_1 \le (\pi_1 V_2 \cdots V_{\ell+1})_1 = P(B_{\ell+1,0})$$

where

$$\pi_1 = [P(B_{1,0}), P(B_{1,1}), P(B_{1,2})] = [0, 2^{-N+1}, 1-2^{-N+1}].$$

Let denote $\gamma_{\ell,i}^{-} := \gamma_{\ell,i} - 1$ and $g_i := g(p_i)$. It then suffices to show that

$$\mathcal{J} := \sum_{i=1}^{\ell} (\gamma_{\ell,i}^{-} - \gamma_{\ell,i+1}^{-})(\mathcal{P}_{22})^{\ell-i} \prod_{j=1}^{i} g_j - \gamma_{\ell,1}^{-}(g_1)^{\ell} \geq 0.$$

Note that $4^{-N} = p_1 \leq p_j < 2^{-N}$, $\mathcal{P}_{22} > g(p_j)$, and thus $\mathcal{P}_{22}^{\ell} > (g(p_1))^{\ell} \geq \prod_{j=1}^{\ell} g(p_j)$. Also,

$$\mathcal{P}_{22} - g(p_i) = 2^{-N} + (N-1)p_i, \qquad \gamma_{\ell,i}^{-} = \frac{2^{-N} - p_i}{2^{-N} + (N-1)p_i} = \frac{2^{-N} - p_i}{\mathcal{P}_{22} - g(p_i)}.$$

Thus we have

$$\begin{aligned}
\mathcal{J} &= \gamma_{\ell,1}^{-}(g_1 \mathcal{P}_{22}^{\ell-1} - g_1^{\ell}) - \sum_{i=2}^{\ell} \gamma_{\ell,i}^{-}(\mathcal{P}_{22} - g_i)\mathcal{P}_{22}^{\ell-i} \prod_{j=1}^{i-1} g_j \\
&\geq \gamma_{\ell,1}^{-}(\mathcal{P}_{22}^{\ell} - g_1^{\ell}) - \sum_{i=1}^{\ell} \gamma_{\ell,i}^{-}(\mathcal{P}_{22} - g_i)\mathcal{P}_{22}^{\ell-i} g_1^{i-1} \\
&\geq \gamma_{\ell,1}^{-}(\mathcal{P}_{22}^{\ell} - g_1^{\ell}) - \gamma_{\ell,1}^{-}(\mathcal{P}_{22} - g_1)\frac{\mathcal{P}_{22}^{\ell} - g_1^{\ell}}{\mathcal{P}_{22} - g_1} = 0.
\end{aligned}$$

Therefore,

$$(\mathcal{P}^{\ell})_{31} \leq (V_2 \cdots V_{\ell+1})_{31},$$

which implies that

$$(\pi_1 \mathcal{P}^{\ell})_1 \leq (\pi_1 V_2 \cdots V_{\ell+1})_1 = P(B_{\ell+1,0}) = P(A_{\ell+2}).$$

Furthermore, it can be checked that

$$(\pi \mathcal{P}^{\ell})_1 = 1 - (\mathcal{P}_{22})^{\ell} - \left(\frac{(1-2^{-N})(1-2^{-N+1})}{1+(N-1)2^{-N}} \right) ((\mathcal{P}_{22})^{\ell} - (\mathcal{P}_{33})^{\ell}).$$

■

Appendix D. Proof of Theorem 4

Proof Since $\mathcal{N}^L(\mathbf{x})$ is a constant function, it follows from Lemma 10 that with probability 1, there exists ℓ such that $\phi(\mathcal{N}^{\ell}(\mathbf{x})) \equiv \mathbf{0}$. Then the gradients of the loss function with respect to the weights and biases in the $1, \cdots, \ell$-th layers vanish. Hence, the weights and biases in layers $1, \ldots, \ell$ will not change when a gradient based optimizer is employed. This implies that $\mathcal{N}^L(\mathbf{x})$ always remains to be a constant function as $\phi(\mathcal{N}^{\ell}(\mathbf{x})) \equiv \mathbf{0}$. Furthermore, the gradient method changes only the weights and biases in layers $\ell+1, \ldots, L$. Therefore, the ReLU NN can only be optimized to a constant function, which minimizes the loss function $\mathcal{L}$. ■

Appendix E. Proof of Theorem 6

Lemma 11 *Let $\boldsymbol{v} \in \mathbb{R}^{n+1}$ be a vector such that $\boldsymbol{v}_k \sim P$ and $\boldsymbol{v}_{-k} \sim N(0, \sigma^2 \boldsymbol{I}_n)$ where $\boldsymbol{v}_{-k}$ is defined in Equation 5. For any nonzero vector $\boldsymbol{x} \in \mathbb{R}^{n+1}$ whose k-th element is positive (i.e., $\boldsymbol{x}_k > 0$), let*

$$\|\tilde{\boldsymbol{x}}_{-k}\|^2 = \frac{\sum_{j \neq k} \boldsymbol{x}_j^2}{\boldsymbol{x}_k^2}, \qquad \tilde{\sigma}^2 = \|\tilde{\boldsymbol{x}}_{-k}\|^2 \sigma^2.$$

Then

$$P\left(\langle \boldsymbol{v}, \boldsymbol{x} \rangle > 0\right) = \begin{cases} \frac{1}{2} + \int_0^M (1 - F_P(t)) \frac{1}{\sqrt{2\pi}\tilde{\sigma}} e^{-\frac{t^2}{2\tilde{\sigma}^2}} dt & \text{if } \tilde{\sigma}^2 > 0 \text{ and } \boldsymbol{x}_k > 0 \\ 1/2 & \text{if } \|\boldsymbol{x}_{-k}\|^2 > 0 \text{ and } \boldsymbol{x}_k = 0 \\ 1 & \text{if } \tilde{\sigma}^2 = 0 \text{ and } \boldsymbol{x}_k > 0, \end{cases} \tag{14}$$

where $F_P(t)$ is the cdf of P.

Proof We first recall some properties of the normal distribution. Let $Y_1, \cdots, Y_n$ be i.i.d. random variables from $N(0, \sigma^2)$ and let $\boldsymbol{Y}_n = (Y_1, \cdots, Y_n)$. Then for any vector $\boldsymbol{a} \in \mathbb{R}^n$,

$$\langle \boldsymbol{Y}_n, \boldsymbol{a} \rangle = \sum_{i=1}^n \boldsymbol{a}_i Y_i \sim N(0, \|\boldsymbol{a}\|^2 \sigma^2).$$

Suppose $\boldsymbol{v}$ is a random vector generated in the way described in Subsection 4.1. Then for any $\boldsymbol{x} \in \mathbb{R}^{n+1}$,

$$\langle \boldsymbol{v}, \boldsymbol{x} \rangle = \sum_{i \neq k} \boldsymbol{v}_i \boldsymbol{x}_i + \boldsymbol{v}_k \boldsymbol{x}_k = \boldsymbol{x}_k \left(Z + \boldsymbol{v}_k\right),$$

where $\tilde{\sigma}^2 = \|\tilde{\boldsymbol{x}}_{-k}\|^2 \sigma^2$ and $Z \sim N(0, \tilde{\sigma}^2)$. If $\boldsymbol{x}_k > 0$ and $\|\mathbf{x}_{-k}\|^2 > 0$, we have

$$\langle \boldsymbol{v}, \boldsymbol{x} \rangle > 0 \quad \iff \quad \boldsymbol{v}_k > -Z \overset{d}{=} Z.$$

Therefore, it suffices to compute $P(\boldsymbol{v}_k > Z)$. Let $f_Z(z)$ be the pdf of Z and $f_{\mathrm{P}}(x)$ be the pdf of $\boldsymbol{v}_k \sim \mathrm{P}$. Then

$$\begin{aligned} P(\boldsymbol{v}_k > Z) &= \int_{-\infty}^{\infty} \int_z^M f_{\mathrm{P}}(x) dx f_Z(z) dz \\ &= \int_{-\infty}^{0} \int_0^M f_{\mathrm{P}}(x) dx f_Z(z) dz + \int_0^M \int_z^M f_{\mathrm{P}}(x) dx f_Z(z) dz \\ &= \frac{1}{2} + \int_0^M (1 - F_{\mathrm{P}}(z)) f_Z(z) dz, \end{aligned}$$

which completes the proof. ■

Proof For each ℓ and s, let k_s^ℓ be randomly uniformly chosen in $\{1, \cdots, N_{\ell-1} + 1\}$. Let $\boldsymbol{V}_s^\ell = [\boldsymbol{W}_s^\ell, \boldsymbol{b}_s^\ell]$ and $\mathbf{n}^\ell(\mathbf{x}) = \left[\phi(\mathcal{N}^\ell(\mathbf{x})), 1\right]$. Recall that for a vector $\mathbf{v} \in \mathbb{R}^{N+1}$,

$$\mathbf{v}_{-k} := [v_1, \cdots, v_{k-1}, v_{k+1}, \cdots, v_{N+1}].$$

To emphasize the dependency of k_s^ℓ, we denote $\boldsymbol{V}_s^\ell$ whose k-th component is generated from P by $\boldsymbol{V}_s^\ell(k)$.

The proof can be started from Equation 12. It suffices to compute

$$P(\tilde{A}_{j,\mathbf{x}}|\tilde{A}^c_{j-1,\mathbf{x}}) = \prod_{s=1}^{N_{j-1}} P\left(\boldsymbol{W}_s^{j-1}\phi(\mathcal{N}^{j-2}(\mathbf{x})) + \boldsymbol{b}_s^{j-1} \leq 0|\tilde{A}^c_{j-1,\mathbf{x}}\right).$$

Note that for each s,

$$P\left(\langle \boldsymbol{V}_s^j, \mathbf{n}^{j-1}\rangle \leq 0|\tilde{A}^c_{j-1,\mathbf{x}}\right) = \frac{1}{N_{j-1}+1}\sum_{k=1}^{N_{j-1}+1} P\left(\langle \boldsymbol{V}_s^j(k), \mathbf{n}^{j-1}\rangle \leq 0|\tilde{A}^c_{j-1,\mathbf{x}}\right).$$

Also, from Lemma 11, we have

$$P\left(\langle \boldsymbol{V}_s^j(k), \mathbf{n}^{j-1}\rangle \leq 0|\tilde{A}^c_{j-1,\mathbf{x}}\right) = \frac{1}{2} - \int_0^M (1-F_P(z))\frac{1}{\sqrt{2\pi}\tilde{\sigma}_{\ell,k}} e^{-\frac{z^2}{2\tilde{\sigma}^2_{\ell,k}}} dz$$

where

$$\tilde{\sigma}^2_{\ell,k} = \frac{\|\mathbf{n}^{\ell-1}_{-k}(\mathbf{x}))\|^2\sigma_\ell^2}{\left(\mathbf{n}_k^{\ell-1}(\mathbf{x})\right)^2}.$$

Note that if $\mathbf{n}_k^{\ell-1}(\mathbf{x}) = 0$, the above probability is simply $1/2$. Also, since the event $\tilde{A}^c_{j-1,\mathbf{x}}$ is given, we know that $\phi(\mathcal{N}^\ell(\mathbf{x})) \neq 0$. Thus,

$$P\left(\langle \boldsymbol{V}_s^j, \mathbf{n}^{j-1}\rangle \leq 0|\tilde{A}^c_{j-1,\mathbf{x}}\right) < \frac{1}{2}.$$

We thus denote

$$\left(\frac{1}{2} - \delta_{\ell-1,\mathbf{x}}(\omega)\right)^{N_{\ell-1}} := P(\tilde{A}_{\ell,\mathbf{x}}|\tilde{A}^c_{\ell-1,\mathbf{x}}(\omega)),$$

where $\delta_{\ell-1,\mathbf{x}}$ is $\mathcal{F}_{\ell-2}$ measurable whose value lies in $(0, 0.5]$ and it depends on $\mathbf{x}$. It then follows from Equation 11 that

$$P(\tilde{A}^c_{\ell,\mathbf{x}}) = \int \prod_{j=1}^{\ell-1}\left(1 - \left(\frac{1}{2} - \delta_{j,\mathbf{x}}(\omega)\right)^{N_j}\right) dP(\omega)$$

where P is the probability distribution with respect to $\{\boldsymbol{W}^j, \boldsymbol{b}^j\}_{1\leq j\leq \ell}$. Note that since the weight matrix and the bias vector of the first layer is initialized from a symmetric distribution, $\delta_{1,\mathbf{x}} = 0$. Let $\delta^*_{1,\mathbf{x}} = 0$. By the mean value theorem, there exists some numbers $\delta^*_{2,\mathbf{x}}, \cdots, \delta^*_{\ell-1,\mathbf{x}} \in (0, 0.5]$ such that

$$P(\tilde{A}^c_{\ell,\mathbf{x}}) = \prod_{j=1}^{\ell-1}\left(1 - \left(\frac{1}{2} - \delta^*_{j,\mathbf{x}}\right)^{N_{j-1}}\right).$$

Let $\boldsymbol{\delta}^*_{\mathbf{x}} = (\delta^*_{1,\mathbf{x}}, \cdots, \delta^*_{\ell-1,\mathbf{x}})$. By setting

$$\boldsymbol{\delta}^* = \boldsymbol{\delta}^*_{\mathbf{x}^*}, \qquad \text{where} \quad \mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}\in\mathcal{D}\setminus\{\mathbf{0}\}} \prod_{j=1}^{\ell-1}\left(1 - \left(\frac{1}{2} - \delta^*_{j,\mathbf{x}}\right)^{N_{j-1}}\right),$$

the proof is completed. ■

Appendix F. Proof of Theorem 8

Let $X_\ell \sim \mathrm{P}_\ell$ whose pdf is denoted by $f_{P_\ell}(x)$. Suppose $0 < \mu'_{\ell,i} = E[X_\ell^i] < \infty$ for $i = 1, 2$. We then define three probability distribution functions:

$$f^{(2)}_{P_\ell}(x) = \frac{x^2}{\mu'_{\ell,2}} f_{P_\ell}(x), \qquad f^{(1)}_{P_\ell}(x) = \frac{x}{\mu'_{\ell,1}} f_{P_\ell}(x), \qquad f^{(0)}_{P_\ell}(x) = f_{P_\ell}(x).$$

Also we denote its corresponding cdfs by $F^{(i)}_{P_\ell}(t) = \int_0^t f^{(i)}_{P_\ell}(x)dx$. For notational convenience, let us assume that $\mathrm{P}_\ell = \mathrm{P}$ and $M_\ell = M$ for all ℓ and define

$$\mathbf{n}^\ell(\mathbf{x}) := \Big[\phi(\mathcal{N}^\ell(\mathbf{x})), 1\Big] \in \mathbb{R}_+^{N_\ell+1}.$$

We denote the pdf of normal distributions by $f^{\ell,k}_Y(y)$ where

$$f^{\ell,k}_Y(y) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}_{\ell,k}} e^{-\frac{y^2}{2\tilde{\sigma}^2_{\ell,k}}}, \qquad \tilde{\sigma}^2_{\ell,k} = \frac{\|\mathbf{n}^{\ell-1}_{-k}(\mathbf{x}))\|^2\sigma^2_\ell}{\left(\mathbf{n}^{\ell-1}_k(\mathbf{x})\right)^2}, \qquad \sigma^2_\ell = \frac{\sigma^2_w}{N_{\ell-1}}$$

where $1 \le k \le N_{\ell-1}+1$. Recall that for a vector $\mathbf{v} \in \mathbb{R}^{N+1}$,

$$\mathbf{v}_{-k} := [v_1, \cdots, v_{k-1}, v_{k+1}, \cdots, v_{N+1}].$$

For any probability distribution P defined on $[0, M]$ whose pdf is denoted by $f_P(x)$, let

$$\mathcal{J}^{\ell,k}_{\mathrm{P}} := \int_0^M \int_y^M f_P(x)dx f^{\ell,k}_Y(y)dy. \tag{15}$$

Let $Y_{\ell,k} \sim N(0, \tilde{\sigma}^2_{\ell,k})$ and $X \sim \mathrm{P}$. Then

$$\begin{aligned} P\left(X > Y_{\ell,k} | \tilde{\sigma}^2_{\ell,k}\right) &= \int_{-\infty}^{\infty} \int_0^M \mathbb{1}_{\{X > Y_{\ell,k}\}} f_P(x) f^{\ell,k}_Y(y)dxdy \\ &= \int_{-\infty}^{0} \int_0^M f_P(x)dx f^{\ell,k}_Y(y)dy + \int_0^M \int_y^M f_P(x)dx f^{\ell,k}_Y(y)dy \\ &= \frac{1}{2} + \mathcal{J}^{\ell,k}_{\mathrm{P}}. \end{aligned}$$

Therefore, $0 \le \mathcal{J}^{\ell,k}_{\mathrm{P}} \le 0.5$.

The proof of Theorem 8 starts with the following lemma.

Lemma 12 *Suppose $Y \sim N(0, \tilde{\sigma}^2_{\ell,k})$, $X \sim P_\ell$ and $\mu'_{\ell,1} \ge M/2$. Then*

$$\int_0^M \int_y^M (x-y)^2 f_{P_\ell}(x)dx f^{\ell,k}_Y(y)dy \le \mu'_{\ell,2}/2.$$

Proof Let $I_{\ell,k}$ be the quantity of our interest:

$$I_{\ell,k} := \int_0^M \int_y^M (x-y)^2 f_{P_\ell}(x) dx f_Y^{\ell,k}(y) dy.$$

Without loss of generality, we can assume $M = 1$. This is because

$$\begin{aligned}
I_{\ell,k} &= \int_0^M \int_y^M (x-y)^2 f_{P_\ell}(x) f_Y^{\ell,k}(y) dx dy \\
&= \int_0^M \int_{y/M}^1 (Mu-y)^2 f_{P_\ell}(Mu) f_Y^{\ell,k}(y) M du dy \\
&= \int_0^1 \int_s^1 M^2 (u-s)^2 (M f_{P_\ell}(Mu))(M f_Y^{\ell,k}(Ms)) du ds \\
&= M^2 \int_0^1 \int_s^1 (u-s)^2 f_U(u) f_S^{\ell,k}(s) du ds \\
&= M^2 \tilde{I}_{\ell,k}
\end{aligned}$$

where $u = x/M$ and $s = y/M$. Here $MU \sim \mathrm{P}_\ell$ and $S \sim N(0, \tilde{\sigma}_{\ell,k}^2/M^2)$.

Let us first consider the inner integral. Let $G(y) = \int_y^1 (x^2 - 2xy + y^2) f_{P_\ell}(x) dx$. It then can be written as follow:

$$\begin{aligned}
G(y) &= \int_y^1 (x^2 - 2xy + y^2) f_{P_\ell}(x) dx = \int_y^1 (x^2 f_{P_\ell}(x) - 2yx f_{P_\ell}(x) + y^2 f_{P_\ell}(x)) dx \\
&= \int_y^1 \left(\mu'_{\ell,2} f_{P_\ell}^{(2)}(x) - 2\mu'_{\ell,1} y f_{P_\ell}^{(1)}(x) + y^2 f_{P_\ell}^{(0)}(x) \right) dx \\
&= \mu'_{\ell,2}(1 - F_{P_\ell}^{(2)}(y)) - 2y\mu'_{\ell,1}(1 - F_{P_\ell}^{(1)}(y)) + y^2(1 - F_{P_\ell}^{(0)}(y)) \\
&= -2\mu'_{\ell,1} y + y^2 + \mu'_{\ell,2} \int_y^1 f_{P_\ell}^{(2)}(x) dx + 2\mu'_{\ell,1} y F_{P_\ell}^{(1)}(y) - y^2 F_{P_\ell}^{(0)}(y).
\end{aligned}$$

Note that since $0 \le y \le 1$, and thus $y^2 \le y$, we have

$$\begin{aligned}
G(y) &= y^2(1 - F_{P_\ell}^{(0)}(y)) - 2\mu'_{\ell,1} y(1 - F_{P_\ell}^{(1)}(y)) + \mu'_{\ell,2} \int_y^1 f_{P_\ell}^{(2)}(x) dx \\
&\le y(1 - F_{P_\ell}^{(0)}(y)) - 2\mu'_{\ell,1} y(1 - F_{P_\ell}^{(1)}(y)) + \mu'_{\ell,2} \int_y^1 f_{P_\ell}^{(2)}(x) dx \\
&= -y(2\mu'_{\ell,1} - 1) \left[1 - \left(\frac{2\mu'_{\ell,1} F_{P_\ell}^{(1)}(y) - F_{P_\ell}^{(0)}(y)}{2\mu'_{\ell,1} - 1} \right) \right] + \mu'_{\ell,2} \int_y^1 f_{P_\ell}^{(2)}(x) dx.
\end{aligned}$$

Since $2\mu'_{\ell,1} \ge 1$ and

$$1 - \left(\frac{2\mu'_{\ell,1} F_{P_\ell}^{(1)}(y) - F_{P_\ell}^{(0)}(y)}{2\mu'_{\ell,1} - 1} \right) \ge 0, \qquad \forall y \in [0,1],$$

we have $G(y) \le \mu'_{\ell,2} \int_y^1 f_{P_\ell}^{(2)}(x)dx$. Therefore,

$$I_{\ell,k} \le \mu'_{\ell,2} \int_0^1 \int_y^1 f_{P_\ell}^{(2)}(x)dx f_Y^{\ell,k}(y)dy = \mu'_{\ell,2} \mathcal{J}_{f_{P_\ell}^{(2)}}^{\ell,k}$$

where $\mathcal{J}_{\mathrm{P}}^{\ell,k}$ is defined in Equation 15. Since $0 \le \mathcal{J}_{\mathrm{P}}^{\ell,k} \le 0.5$, we conlcude that $I_{\ell,k} \le \mu'_{\ell,2}/2$. ■

Proof It follows from

$$\mathcal{N}_j^\ell(\mathbf{x}) = \boldsymbol{W}_j^\ell \phi(\mathcal{N}^{\ell-1}(\mathbf{x})) + \boldsymbol{b}_j^\ell, \qquad 1 \le j \le N_\ell$$

that we have

$$E\left[\mathcal{N}_j^\ell(\mathbf{x})^2 | \mathcal{N}^{\ell-1}(\mathbf{x})\right] = \sigma_\ell^2 \sum_{t \ne k_j^\ell} \phi(\mathcal{N}_t^{\ell-1}(\mathbf{x}))^2 + \mu'_{\ell,2} \phi(\mathcal{N}_{k_j^\ell}^{\ell-1}(\mathbf{x}))^2.$$

Since k_j^ℓ is randomly uniformly selected, by taking the expectation with respect to k_j^ℓ, we have

$$\mathbb{E}\left[\mathcal{N}_j^\ell(\mathbf{x})^2 | \mathcal{N}^{\ell-1}(\mathbf{x})\right] = \frac{\sigma_\ell^2 N_\ell + \mu'_{\ell,2}}{N_\ell + 1} \left(1 + \|\phi(\mathcal{N}^{\ell-1}(\mathbf{x}))\|^2\right).$$

Since the above is independent of j, and $q^\ell(\mathbf{x}) = \|\mathcal{N}^\ell(\mathbf{x})\|^2/N_\ell$,

$$\mathbb{E}\left[q^\ell(\mathbf{x}) | \mathcal{N}^{\ell-1}(\mathbf{x})\right] = \frac{\sigma_\ell^2 N_\ell + \mu'_{\ell,2}}{N_\ell + 1} \left(1 + \|\phi(\mathcal{N}^{\ell-1}(\mathbf{x}))\|^2\right). \tag{16}$$

At a fixed ℓ and for each $t = 1, \cdots, N_\ell$, let $\mathbf{v}^{\ell,t} = [\mathbf{W}_t^\ell, \mathbf{b}_t^\ell] \in \mathbb{R}^{N_{\ell-1}+1}$ be a random vector such that $\mathbf{v}_{-k_t}^{\ell,t} \sim N(0, \sigma_\ell^2 \boldsymbol{I}_{N_{\ell-1}})$ and $\mathbf{v}_{k_t}^{\ell,t} \sim \mathrm{P}_\ell$. Then $\mathcal{N}_t^\ell(\mathbf{x}) = \langle \mathbf{v}^{\ell,t}, \mathbf{n}^\ell(\mathbf{x}) \rangle$. Given $\mathcal{N}^{\ell-1}(\mathbf{x})$, assuming $(\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x})) > 0$, one can view $\mathcal{N}_t^\ell(\mathbf{x})$ as

$$\mathcal{N}_t^\ell(\mathbf{x}) = \mathbf{n}_{k_t}^{\ell-1}(\mathbf{x}) \left(\sigma_{\ell,k_t} Z + X_\ell\right), \qquad Z \sim N(0,1) \text{ and } X_\ell \sim \mathrm{P}_\ell.$$

Thus $\phi(\mathcal{N}_t^\ell(\mathbf{x}))^2 = \left(\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x})\right)^2 \phi\left(\sigma_{\ell,k_t} Z + X_\ell\right)^2$ and we obtain

$$\begin{aligned}
&\frac{1}{(\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x}))^2} E\left[\phi(\mathcal{N}_t^\ell(\mathbf{x}))^2 | \mathcal{N}^{\ell-1}(\mathbf{x})\right] = E\left[\phi\left(\sigma_{\ell,k_t} Z + X_\ell\right)^2 | \mathcal{N}^{\ell-1}(\mathbf{x})\right] \\
&= \int_{-\infty}^M \int_y^M (x-y)^2 dF_P(x) f_Y^{\ell,k_t}(y)dy \\
&= \int_{-\infty}^0 \int_0^M (x-y)^2 dF_P(x) f_Y^{\ell,k_t}(y)dy + \int_0^M \int_y^M (x-y)^2 dF_P(x) f_Y^{\ell,k_t}(y)dy \\
&= \int_{-\infty}^0 \int_0^M \left(x^2 + y^2 - 2xy\right) dF_P(x) f_Y^{\ell,k_t}(y)dy + \int_0^M \int_y^M (x-y)^2 dF_P(x) f_Y^{\ell,k_t}(y)dy \\
&= \int_{-\infty}^0 \left(\mu'_{\ell,2} + y^2 - 2\mu'_{\ell,1} y\right) f_Y^{\ell,k_j}(y)dy + \int_0^M \int_y^M (x-y)^2 dF_P(x) f_Y^{\ell,k_t}(y)dy \\
&= \frac{1}{2}\left(\mu'_{\ell,2} + \tilde{\sigma}_{\ell,k_t}^2\right) + \mu'_{\ell,1}\sqrt{\frac{2}{\pi}}\tilde{\sigma}_{\ell,k_t} + \int_0^M \int_y^M (x-y)^2 dF_P(x) f_Y^{\ell,k_t}(y)dy.
\end{aligned}$$

By multiplying $(\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x}))^2$ in the both sides, we have

$$E[\phi(\mathcal{N}_t^\ell(\mathbf{x}))^2|\mathcal{N}^{\ell-1}(\mathbf{x})] = \frac{1}{2}\left((\mu'_{\ell,2}-\sigma_\ell^2)(\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x}))^2 + (\|\phi(\mathcal{N}^{\ell-1}(\mathbf{x}))\|^2+1)\sigma_\ell^2\right) + (\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x}))^2\int_0^M\int_y^M (x-y)^2 dF_P(x) f_Y^{\ell,k_t}(y)dy + \mu'_{\ell,1}\sigma_\ell\sqrt{\frac{2}{\pi}}\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x})\|\mathbf{n}_{-k_t}^{\ell-1}(\mathbf{x})\|.$$

Since k_t is randomly selected, by taking expectation w.r.t. k_j, we have

$$\mathbb{E}[\phi(\mathcal{N}_t^\ell(\mathbf{x}))^2|\mathcal{N}^{\ell-1}(\mathbf{x})] = \frac{(1+\|\phi(\mathcal{N}^{\ell-1}(\mathbf{x}))\|^2)}{2}\left(\frac{\mu'_{\ell,2}-\sigma_\ell^2}{N_{\ell-1}+1}+\sigma_\ell^2\right) + \sum_{k_t=1}^{N_{\ell-1}+1}\frac{(\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x}))^2}{N_{\ell-1}+1}\int_0^M\int_y^M (x-y)^2 dF_P(x) f_Y^{\ell,k_t}(y)dy + \mu'_{\ell,1}\sqrt{\frac{2}{\pi}}\sigma_\ell\sum_{k_t=1}^{N_{\ell-1}+1}\frac{\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x})\|\mathbf{n}_{-k_t}^{\ell-1}(\mathbf{x})\|}{N_{\ell-1}+1}. \quad (17)$$

By the Cauchy-Schwarz inequality, the third term in the right hand side of Equation 17 can be bounded by

$$\mu'_{\ell,1}\sqrt{\frac{2}{\pi}}\sigma_\ell\sum_{k_t=1}^{N_{\ell-1}+1}\frac{\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x})\|\mathbf{n}_{-k_t}^{\ell-1}(\mathbf{x})\|}{N_{\ell-1}+1} \le \mu'_{\ell,1}\sqrt{\frac{2}{\pi}}\frac{\sigma_w(1+\|\phi(\mathcal{N}^{\ell-1}(\mathbf{x}))\|^2)}{N_{\ell-1}+1}.$$

Let $I_{\ell,k_t} = \int_0^M\int_y^M (x-y)^2 dF_P(x) f_Y^{\ell,k_t}(y)dy$. It then follows from Lemma 12 that $I_{\ell,k} \le \mu'_{\ell,2}/2$ for any k. Thus the second term in the right hand side of Equation 17 can be bounded by

$$\sum_{k_t=1}^{N_{\ell-1}+1}\frac{(\mathbf{n}_{k_t}^{\ell-1}(\mathbf{x}))^2}{N_{\ell-1}+1}\int_0^M\int_y^M (x-y)^2 dF_P(x) f_Y^{\ell,k_t}(y)dy \le \frac{\mu'_{\ell,2}}{2}\frac{(1+\|\phi(\mathcal{N}^{\ell-1}(\mathbf{x}))\|^2)}{N_{\ell-1}+1}$$

Therefore, we obtain

$$E[\phi(\mathcal{N}_t^\ell(\mathbf{x}))^2|\mathcal{N}^{\ell-1}(\mathbf{x})] \le \mathcal{C}_\ell\frac{(1+\|\phi(\mathcal{N}^{\ell-1}(\mathbf{x}))\|^2)}{2}$$

where

$$\mathcal{C}_\ell = \frac{\mu'_{\ell,2}-\sigma_\ell^2}{N_{\ell-1}+1}+\sigma_\ell^2+\frac{2\sqrt{2}\mu'_{\ell,1}\sigma_w}{(N_{\ell-1}+1)\sqrt{\pi}}+\frac{\mu'_{\ell,2}}{N_{\ell-1}+1}.$$

Since $\mathcal{C}_\ell$ is independent of t,

$$E[\|\phi(\mathcal{N}^\ell(\mathbf{x}))\|^2|\mathcal{N}^{\ell-1}(\mathbf{x})] \le N_\ell\mathcal{C}_\ell\frac{(1+\|\phi(\mathcal{N}^{\ell-1}(\mathbf{x}))\|^2)}{2}.$$

It then follows from Equation 16,

$$1+\|\phi(\mathcal{N}^{\ell-1}(\mathbf{x}))\|^2 = \frac{N_\ell+1}{\sigma_\ell^2 N_\ell+\mu'_{\ell,2}}E\left[q^\ell(\mathbf{x})|\mathcal{N}^{\ell-1}(\mathbf{x})\right]$$

that

$$E[\|\phi(\mathcal{N}^\ell(\mathbf{x}))\|^2|\mathcal{N}^{\ell-1}(\mathbf{x})] \leq \frac{\mathcal{C}N_\ell(N_\ell+1)}{2(\sigma_\ell^2 N_\ell + \mu'_{\ell,2})} E\left[q^\ell(\mathbf{x})|\mathcal{N}^{\ell-1}(\mathbf{x})\right].$$

Thus we have

$$\begin{aligned} E\left[q^{\ell+1}(\mathbf{x})|\mathcal{N}^{\ell-1}(\mathbf{x})\right] &= \frac{\sigma_{\ell+1}^2 N_{\ell+1} + \mu'_{\ell+1,2}}{N_{\ell+1}+1}\left(1 + E[\|\phi(\mathcal{N}^\ell(\mathbf{x}))\|^2|\mathcal{N}^{\ell-1}(\mathbf{x})]\right) \\ &\leq \sigma_{b,\ell}^2 + \frac{\mathcal{A}_{\ell,upp}}{2} E\left[q^\ell(\mathbf{x})|\mathcal{N}^{\ell-1}(\mathbf{x})\right] \end{aligned}$$

where

$$\sigma_{b,\ell}^2 = \frac{\sigma_{\ell+1}^2 N_{\ell+1} + \mu'_{\ell+1,2}}{N_{\ell+1}+1}, \qquad \mathcal{A}_{\ell,upp} = \frac{\sigma_{b,\ell}^2 N_\ell(N_\ell+1)}{\sigma_\ell^2 N_\ell + \mu'_{\ell,2}}\mathcal{C}_\ell.$$

By taking expectation with respect to $\mathcal{N}^{\ell-1}(\mathbf{x})$, we obtain

$$E[q^{\ell+1}(\mathbf{x})] \leq \frac{\mathcal{A}_{\ell,upp}}{2} E[q^\ell(\mathbf{x})] + \sigma_{b,\ell}^2$$

The lower bound can be obtained by dropping the second and the third terms in Equation 17. Thus

$$\frac{\mathcal{A}_{\ell,low}}{2} E[q^\ell(\mathbf{x})] + \sigma_{b,\ell}^2 \leq E[q^{\ell+1}(\mathbf{x})] \leq \frac{\mathcal{A}_{\ell,upp}}{2} E[q^\ell(\mathbf{x})] + \sigma_{b,\ell}^2$$

where

$$\mathcal{A}_{\ell,low} = \frac{\sigma_{\ell+1}^2 N_{\ell+1} + \mu'_{\ell+1,2}}{N_{\ell+1}+1} \frac{N_\ell+1}{\sigma_\ell^2 N_\ell + \mu'_{\ell,2}} \left(\frac{\mu'_{\ell,2} N_\ell - \sigma_w^2}{N_{\ell-1}+1} + \sigma_w^2\right).$$

■