Semi-flat minima and saddle points by embedding neural networks to overparameterization

Kenji Fukumizu†‡,† Shoichiro Yamaguchi‡
Yoh-ichi Mototake†
Mirai Tanaka†
†The Institute of Statistical Mathematics
Tachikawa, Tokyo 190-8562, Japan
{fukumizu, mototake, mirai}@ism.ac.jp
‡Preferred Networks, Inc.
Chiyoda-ku, Tokyo 100-0004, Japan
guguchi@preferred.jp

Abstract

We theoretically study the landscape of the training error for neural networks in overparameterized cases. We consider three basic methods for embedding a network into a wider one with more hidden units, and discuss whether a minimum point of the narrower network gives a minimum or saddle point of the wider one. Our results show that the networks with smooth and ReLU activation have different partially flat landscapes around the embedded point. We also relate these results to a difference of their generalization abilities in overparameterized realization.

1 Introduction

Deep neural networks (DNNs) have been applied to many problems with remarkable successes. On the theoretical understanding of DNNs, however, many problems are still unsolved. Among others, local minima are important issues on learning of DNNs; existence of many local minima is naturally expected by its strong nonlinearity, while people also observe that, with a large network and the stochastic gradient descent, training of DNNs may avoid this issue [8, 9]. For a better understanding of learning, it is essential to clarify the landscape of the training error.

This paper focuses on the error landscape in overparameterized situations, where the number of units is surplus to realize a function. This naturally occurs when a large network architecture is employed, and has been recently discussed in connection to optimization and generalization of neural networks ([14, 2, 1] to list a few). To formulate overparameterization rigorously, this paper introduces three basic methods, unit replication, inactive units, and inactive propagation, for embedding a network to a network of more units in some layer. We investigate especially the landscape of the training error around the embedded point, when we embed a minimizer of the error for a smaller model.

A relevant topic to this paper is flat minima [6, 7], which attract much attention in literature. Such flatness of minima is often observed empirically, and is connected to generalization performance [3, 8]. There are also some works on how to define flatness appropriately and its relations to generalization [15, 17]. Different from these works, this paper shows some embeddings cause semi-flat minima, at which a lower dimensional affine subset in the parameter space gives a constant value of error (see Sec. A). We will also discuss difference between smooth activation and Rectified Linear Unit (ReLU); at a semi-flat minimum obtained by embedding a network of zero training error, the ReLU networks have more flat directions. Using PAC-Bayes arguments [11], we relate this to the difference of generalization bounds between ReLU and smooth networks in overparameterized situations.

A neural network is a computational model inspired by the brain. It is composed of interconnected nodes, or neurons, which process information in a structured way. The key components of a neural network include an input layer, one or more hidden layers, and an output layer. The input layer receives data, the hidden layers process this data, and the output layer produces the result.

Quantum mechanics and quantum computing offer promising methods for solving complex problems, particularly those involving high-dimensional data. Quantum computers use quantum bits, or qubits, which can exist in multiple states simultaneously. This allows them to process large amounts of information more efficiently than classical computers.

2 Neural network and its embedding to a wider model

We discuss $J$ layer, fully connected neural networks that have an activation function $\varphi(z; w)$, where $z$ is the input to a unit and $w$ is a parameter vector. The output of the $i$-th unit $U^0_i$ in the $q$-th layer is recursively defined by $z^q_i = \varphi(z^{q-1}; U^0_i)$, where $U^0_i$ is the weight between the $U^0_i$ and the $(q-1)$-th layer. The activation function $\varphi(z; w)$ is any nonlinear function, which often takes the form $\varphi(w^{q}_{wgt}z-w_{bias})$ with $w = (w^{q}_{wgt}, w_{bias})$; typical examples are the sigmoidal function $\varphi(z; w) = \text{tanh}(w^{q}_{wgt}z-w_{bias})$ and ReLU $\varphi(z; w) = \max \{w^{q}_{wgt}z-w_{bias}, 0\}$. This paper assumes that there is $w^{(0)}_i$ such that $\varphi(x; w^{(0)}) = 0$ for any $x$. Focusing the $q$-th layer, with size of other layers fixed, the set of networks having $H$ units in the $q$-th layer is denoted by $\mathcal{N}_H$. With a parameter $\theta^{(H)} = (W_0, w_1, \ldots, w_H, v_1, \ldots, v_H, \nu_0)$, the function $f^{(H)}_\theta$ of $\mathcal{N}_H$ is defined by

$$f^{(H)}_\theta(x) := f^{(H)}(x; \theta^{(H)}) = \psi(\sum_{j=1}^{H} v_j \varphi(x; w_j, W_0; V_0),$$

where $\varphi(x; w_j, W_0)$ is the output of $U^0_j$ with a summarized parameter $W_0$ in the previous layers, and $\psi(z^{q+1}; V_0)$ is all the parts after $z^{q+1}$ with parameter $V_0$. Note that $v_j$ is a connection weight from the unit $U^0_j$ to the units in the $(q+1)$-th layer (we omit the bias term for simplicity). The number of units in the $(q-1)$-th and $(q+1)$-th layers are denoted by $D$ and $M$, respectively.

Embedding of a network refers to a map associating a narrower network in $\mathcal{N}_{H_0}$ ($H_0 < H$) with a network of a specific parameter in a wider model $\mathcal{N}_H$ to realize the same function, keeping other layers unchanged. For clarity, we use $(\zeta, u_i)$ instead of $(v_j, w_j)$ for the parameter $\theta^{(H_0)}$ of $\mathcal{N}_{H_0}$:

$$f^{(H_0)}_{\theta^{(H_0)}}(x) := f^{(H_0)}(x; \theta^{(H_0)}) = \psi(\sum_{i=1}^{H_0} v_i \varphi(x; u_i, W_0; V_0).$$

We consider minima and stationary points of the empirical risk (or training error)

$$L_H(\theta^{(H)}) := \sum_{n=1}^{N} \ell(y_n, f^{(H)}(x_n; \theta^{(H)})),$$

where $\ell(y, f)$ is a loss function to measure the discrepancy between a teacher $y$ and network output $f$, and $(x_1, y_1), \ldots, (x_n, y_n)$ are given training data. Typical examples of $\ell(y, f)$ include the square error $||y-f||^2/2$ and logistic loss $-y \log f - (1 - y) \log(1 - f)$ for $y \in \{0, 1\}$ and $f \in (0, 1)$. In the sequel, we assume the second order differentiability of $\ell(y, f)$ with respect to $f$ for each $y$.

2.1 Three embedding methods of a network

To formulate overparameterization, we introduce three basic methods for embedding $f^{(H_0)}_{\theta^{(H_0)}}$ into $\mathcal{N}_H$ so that it realizes exactly the same function as $f^{(H_0)}_{\theta^{(H_0)}}$. See Table 1 and Figure 1 for the definitions.

(I) Unit replication: We fix a unit, say the $H_0$-th unit $U^0_{H_0}$ in $\mathcal{N}_{H_0}$, and replicate it. Simply, $\theta^{(H)}$ has $H - H_0 + 1$ copies of $u_{H_0}$, and divides the weight $\zeta_{H_0}$ by $v_{H_0}, \ldots, v_H$, keeping the other
Table 1: Three methods of embedding

<table>
<thead>
<tr>
<th>Unit replication $\Pi_{\text{rep}}(\theta(H_0))$</th>
<th>Inactive units $\Pi_{\text{in}}(\theta(H_0))$</th>
<th>Inactive propagation $\Pi_{\text{prop}}(\theta(H_0))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_i = u_i \ (1 \leq i \leq H_0 - 1)$</td>
<td>$w_i = u_i \ (1 \leq i \leq H_0)$</td>
<td>$w_i = u_i \ (1 \leq i \leq H_0)$</td>
</tr>
<tr>
<td>$v_i = \zeta_i \ (1 \leq i \leq H_0 - 1)$</td>
<td>$v_i = \zeta_i \ (1 \leq i \leq H_0)$</td>
<td>$v_i = \zeta_i \ (1 \leq i \leq H_0)$</td>
</tr>
<tr>
<td>$w_H = \cdots = w_H = u_{H_0}$</td>
<td>$w_{H_0+1} = \cdots = w_H = w^{(o)}$</td>
<td>$w_{H_0+1}, \ldots, w_H$: arbitrary</td>
</tr>
<tr>
<td>$v_{H_0+1}, \ldots, v_H$: arbitrary</td>
<td>$v_{H_0+1} = \cdots = v_H = 0$</td>
<td>$v_{H_0+1} = \cdots = v_H = 0$</td>
</tr>
</tbody>
</table>

Figure 1: Embedding of a narrower network to a wider one.

For any $\theta(H) \in \Pi_{\text{rep}}(\theta(H_0)) \cup \Pi_{\text{in}}(\theta(H_0)) \cup \Pi_{\text{prop}}(\theta(H_0))$, we have $f(H) = f(H_0)$.

It is important to note that a network is not uniquely embedded in a wider model, in contrast to fixed bases models such as the polynomial model. This unidentifiability has been clarified for three-layer networks [10, 16]; in fact, for three layer networks of tanh activation, [16] shows that the three methods essentially cover all possible embedding. For three-layer networks of 1-dimensional input and smooth activation, [4] shows that this unidentifiable embedding causes minima or saddle points. The current paper extends this result to general networks with ReLU as well as smooth activation.

3 Embedding of smooth networks

This section assumes the second order differentiability of $\varphi(x; w)$ on $w$. The case of ReLU will be discussed in Sec. 4. Let $\theta^*_H$ be a stationary point of $L_H$, i.e., $\frac{\partial L_H}{\partial \theta_H} = 0$. We are interested in whether the embedding in Sec. 2 also gives a stationary point of $L_H$. More importantly, we wish to know if a minimum of $L_{H_0}$ is embedded to a minimum of $L_H$. A network can be embedded by any combination of the three methods, but we consider their effects separately for simplicity. The definition of minimum, saddle point, and related notions are given by Sec. A.

3.1 Stationary properties of embedding

To discuss the stationarity for the case (I) unit replication, we need to restrict $\Pi_{\text{rep}}(\theta(H_0))$ to a subset. For $\theta(H_0)$, define $\theta_{\lambda}^{(H)}$ for every $\lambda = (\lambda_{H_0}, \ldots, \lambda_H) \in \mathbb{R}^{H-H_0+1}$ with $\sum_{j=H_0}^H \lambda_j = 1$ by

$$w_i = u_i, \quad v_i = \zeta_i \ (1 \leq i \leq H_0 - 1),$$
$$w_H = \cdots = w_H = u_{H_0}, \quad v_j = \lambda_j \zeta_{H_0} \ (H_0 \leq j \leq H). \ (4)$$

Obviously, $\theta_{\lambda}^{(H)} \in \Pi_{\text{rep}}(\theta(H_0))$ so that $f^{(H)}(x; \theta_{\lambda}^{(H)}) = f(H_0)(x; \theta^{(H_0)})$. The next theorem tells that a stationary point of $\mathcal{N}_{H_0}$ is embedded to an $(H - H_0)$-dimensional stationary subset of $\mathcal{N}_H$. Let $\theta_{\lambda}^{(H_0)}$ be a stationary point of $L_{H_0}$. Then, for any $\lambda = (\lambda_{H_0}, \ldots, \lambda_H)$ with $\sum_{j=H_0}^H \lambda_j = 1$, the point $\theta_{\lambda}^{(H)}$ defined by Eq. (4) is a stationary point of $L_H$. Theorem 2.
We next consider the embedding \( \theta^{(H)}_\lambda \) of a minimum point \( \theta^{(H)}_* \) of \( L_{H_0} \). In the sequel, for readability, we discuss three-layer models \((J = 3)\) and linear output units. Note however that, for general \( J \), the derivatives and Hessian of \( L_H \) for the other parameters are exactly the same as those of \( L_{H_0} \), for the corresponding parameters. We omit the full description here. The two models are simply given by

\[
N_H: f^{(H)}(x; \theta^{(H)}) = \sum_{j=1}^{H} v_j \varphi(x; w_j) \quad \text{and} \quad N_{H_0}: f^{(H_0)}(x; \theta^{(H_0)}) = \sum_{i=1}^{H_0} \zeta_i \varphi(x; u_i). \tag{5}
\]

To simplify the Hessian for unit replication, we introduce a new parameterization of \( \tilde{N}_H \). Let \( \lambda \in \mathbb{R}^{H-H_0+1} \) be fixed such that \( \lambda_{H_0} + \cdots + \lambda_H = 1 \) and \( \lambda_j \neq 0 \). For such \( \lambda \), take an \((H-H_0+1) \times (H-H_0+1)\) matrix \( \Lambda = (\alpha_{cj}) \) \((H_0+1 \leq c \leq H, H_0 \leq j \leq H)\) that satisfies the two conditions:

\[
\begin{align*}
&\text{(A1)} \quad \begin{bmatrix} 1_{H-H_0+1}^T \end{bmatrix} \Lambda = 1, \\
&\text{(A2)} \quad \sum_{j=H_0}^{H} \alpha_{cj} \lambda_j = 0 \quad \text{for any } H_0+1 \leq c \leq H.
\end{align*}
\]

To find such \( \Lambda \), take \( \Lambda = (a_{H_0+1}, \ldots, a_H)^T \) so that \( \alpha_{cj}^T \lambda = 0 \). Then, if \( \sum_{c=H_0+1}^{H} s_c a_c = 1 \), for some scalars \( s_c \), taking the inner product with \( \lambda \) causes a contradiction.

Given such \( \lambda \) and \( A = (\alpha_{cj}) \), define a bijective linear transform from \((v_{H_0}, \ldots, v_H; w_{H_0}, \ldots, w_H)\) to \((a, \xi_{H_0+1}, \ldots, \xi_H; b, \eta_{H_0+1}, \ldots, \eta_H)\) by

\[
\begin{align*}
w_j &= b + \sum_{c=H_0+1}^{H} \alpha_{cj} \eta_c \quad \text{and} \quad v_j = \lambda_j a + \sum_{c=H_0+1}^{H} \alpha_{cj} \xi_c \quad (H_0 \leq j \leq H). \tag{6}
\end{align*}
\]

The parameter \( b \) serves as the direction that makes all the hidden units behave equally, and \( (\eta_j) \) define the remaining \( H-1 \) directions that differentiate them. The parameter \( b \) thus essentially plays the role of \( u_{H_0} \) for \( \tilde{N}_H \). Also, \( a \) works as \( \xi_{H_0} \) when all \( w_j \) are equal. The next lemma confirms this role of \((a, b)\) and shows that the directions \( \eta_c \) and \( \xi_c \) do not change the function \( f^{(H)} \) at \( \theta^{(H)}_\lambda \).

**Lemma 3.** Let \( \theta^{(H_0)} \) be any parameter of \( N_{H_0} \), and \( \theta^{(H)}_\lambda \) be its embedding defined by Eq. (4). Then,

\[
\begin{align*}
\frac{\partial f^{(H)}(x; \theta^{(H)})}{\partial b} \bigg|_{\theta^{(H)} = \lambda^{(H)}_{\lambda}} &= \frac{\partial f^{(H_0)}(x; \theta^{(H_0)})}{\partial u_{H_0}}, \\
\frac{\partial f^{(H)}(x; \theta^{(H)})}{\partial a} \bigg|_{\theta^{(H)} = \lambda^{(H)}_{\lambda}} &= \frac{\partial f^{(H_0)}(x; \theta^{(H_0)})}{\partial \xi_{H_0}}, \\
\frac{\partial f^{(H)}(x; \theta^{(H)})}{\partial \eta_c} \bigg|_{\theta^{(H)} = \lambda^{(H)}_{\lambda}} &= \frac{\partial f^{(H_0)}(x; \theta^{(H_0)})}{\partial \xi_{H_0}}, \\
\frac{\partial f^{(H)}(x; \theta^{(H)})}{\partial \xi_c} \bigg|_{\theta^{(H)} = \lambda^{(H)}_{\lambda}} &= 0.
\end{align*}
\]

From Lemma 3, the Hessian takes a simple form:

**Lemma 4.** Let \( \lambda \) and \( A \) be as above. Suppose \( \theta^{(H_0)}_* \) is a stationary point of \( N_{H_0} \) and \( \theta^{(H)}_\lambda \) is its embedding defined by Eq. (4). Then, the Hessian matrix of \( L_H \) with respect to \( \omega = (a, b, \xi_{H_0+1}, \ldots, \xi_H, \eta_{H_0+1}, \ldots, \eta_H) \) at \( \theta^{(H)}_\lambda \) is given by

\[
\frac{\partial^2 L_H(\theta^{(H)}_\lambda)}{\partial \omega \partial \omega} = \begin{bmatrix}
a & b & \xi_\delta & \xi_\xi & \xi_\eta \\
b & \xi_c & O & O & O \\
\xi_\xi & O & O & F & \tilde{F} \\
\xi_\eta & O & O & \tilde{F}^T & \tilde{G}
\end{bmatrix}.
\]

The lower-right block \( \tilde{G} := (\partial^2 L_H(\theta^{(H)}_\lambda))_{cd} \), which is a symmetric matrix of \((H-H_0) \times D \) dimension, is given by \((\Lambda \Lambda^T) \otimes G \) with \( \Lambda = \text{Diag}(\lambda_{H_0}, \ldots, \lambda_H) \) and \( G := \sum_{j=1}^{n} \frac{\partial f^{(H)}(x; \theta^{(H)})}{\partial z} \zeta_{H_0} \frac{\partial \varphi(x; u_{H_0})}{\partial u_{H_0}} \) and \( \tilde{F} := (\partial^2 L_H(\theta^{(H)}_\lambda))_{cd} \), which is of size \((H-H_0) \times M \) dimension, is given by \((\Lambda \Lambda^T) \otimes F \) with \( F := \sum_{j=1}^{n} \frac{\partial f^{(H)}(x; \theta^{(H)})}{\partial z} \zeta_{H_0} \frac{\partial \varphi(x; u_{H_0})}{\partial u_{H_0}} \).
Lemma 4 shows that, with the reparametrization, the Hessian at the embedded stationary point $\theta^{(H)}_\lambda$ contains the Hessian of $L_{H_0}$ with $a, b,$ and that the cross blocks between $(a, b)$ and $(\xi, \eta)$ are zero. Note that the $\xi$ block is zero, which is important when we prove Theorem 5.

**Theorem 5.** Consider a three layer network given by Eq. (5). Suppose that the output dimension $\phi$ is positive or negative definite, and $F \neq O$. See Sec. D.3 in Supplements for a complete proof. The assumption $M \geq 2$ is necessary for the condition (i) to happen. In fact, [4] discussed the case of $M = 1$, in which $F = O$ is derived. The paper also gave a sufficient condition that the embedded point $\theta^{(H)}_\lambda$ is a local minimum when $G$ is positive (or negative) definite. See Sec. E for more details on the special case of $M = 1$.

Suppose that $\theta^{(H_0)}$ attains zero training error. Then, $\theta^{(H)}_\lambda$ can never be a saddle point but a global minimum. Therefore, the situation (ii) can never happen. In that case, if $G$ is invertible, it must be positive definite and $F = O$. We will discuss this case further in Sec. 5.1.

## 4 Semi-flat minima by embedding of ReLU networks

This section discusses networks with ReLU. Its special shape causes different results. Let $\phi(t)$ be the ReLU function: $\phi(t) = \max\{t, 0\}$, which is used very often in DNNs to prevent vanishing gradients [12, 5]. The activation is given by $\varphi(x; w) = \phi(w^T \tilde{x})$ with $w^T \tilde{x} := w_{\text{bias}} x - w_{\text{bias}}$. It is important to note that the ReLU function satisfies positive homogeneity; i.e., $\phi(\alpha t) = \alpha \phi(t)$ for any $\alpha \geq 0$. This causes special properties on $\varphi$, that is, (a) $\varphi(x; rw) = r \varphi(x; w)$ for any $r \geq 0$, (b) $\frac{\partial \varphi(x; w)}{\partial w} \bigg|_{w=rw} = \frac{\partial \varphi(x; w)}{\partial w} \bigg|_{w=w}$ if $r > 0$, $w^T \tilde{x} \neq 0$, and (c) $\frac{\partial^2 \varphi(x; w)}{\partial w \partial w} = 0$ if $w^T \tilde{x} \neq 0$.

From the positive homogeneity, effective parameterization needs some normalization of $v_j$ or $w_j$. However, this paper uses the redundant parameterization. In our theoretical arguments, no problem is caused by the redundancy, while it gives additional flat directions in the parameter space.

### 4.1 Embeddings of ReLU networks

Reflecting the above special properties, we introduce modified versions for embeddings of $\theta^{(H_0)}_\gamma$.

**(I)\text{R Unit replication:** Fix $L^{(H)}_{H_0}$, and take $\gamma = (\gamma_{H_0}, \ldots, \gamma_{H}) \in \mathbb{R}^{H-H_0+1}$ and $\beta = (\beta_{H_0}, \ldots, \beta_{H})$ such that $\beta_j > 0$ ($H_0 \leq j \leq H$) and $\sum_{j=H_0}^{H} \gamma_j \beta_j = 1$. Define $\theta^{(H)}_\gamma$ by

$$
\begin{align*}
  w_i &= u_i, & v_i &= \zeta_i \ (1 \leq i \leq H_0 - 1), \\
  w_j &= \beta_j u_{H_0}, & v_j &= \gamma_j \zeta_{H_0} \quad (H_0 \leq j \leq H).
\end{align*}
$$

**(II)\text{R Inactive units:** Define a parameter $\hat{\theta}^{(H)}$ by

$$

w_i = u_i,  \quad v_i = \zeta_i \ (1 \leq i \leq H_0),  \quad v_j : \text{arbitrary } \quad (H_0 + 1 \leq j \leq H)
$$

$$
\begin{align*}
  w_j &\text{ such that } w_j^T \tilde{x}_\nu < 0 \quad (\forall \nu, H_0 + 1 \leq j \leq H).
\end{align*}
$$

Note that the condition (II)\text{R} is different from the smooth activation case. The last condition is easily satisfied if $w_{\text{bias}}$ is large. Note also that $\varphi(x; w_j) = 0$ for each $\nu$, but $\varphi(x; w_j) \neq 0$ in general. Since a small change of $w_j$ ($H_0 + 1 \leq j \leq H$) does not alter $\varphi(x; w_j) = 0$, the function $L_H$ is constant locally on $v_j$ and $w_j$ ($H_0 + 1 \leq j \leq H$) at $\theta^{(H)}$. This is clear from the smooth case, where changing $w_j$ from $w^{(0)}$ may cause a different function.

**(III)\text{R Inactive propagation:** The inactive propagation is exactly the same as the smooth activation case. The embedded point is denoted by $\hat{\theta}^{(H)}$.

The following proposition is obvious from the definitions.
We first consider the embedding of a minimum by inactive units. Let \( \mathbf{w}_j, \mathbf{u}_j \) be any parameter of the ReLU network \( \mathcal{N}_{H_0} \), and \( \theta_{\gamma, \beta} \) be defined by Eq. (9). Assume that \( \mathbf{w}_j, \mathbf{u}_j \neq 0 \) for all \( j \). Then, there is \( \delta > 0 \) such that

\[
f^{(H)}(\mathbf{x}_\nu; \theta_{\gamma, \beta}) = f^{(H_0)}(\mathbf{x}_\nu; \theta_{\gamma, \beta}^{H_0}) \quad (\forall \mathbf{x}_\nu \in B_\delta(\theta_{\gamma, \beta}), \forall \nu = 1, \ldots, n).
\]

Next, for the unit replication of ReLU networks, the piecewise linearity of ReLU causes additional flat directions. To see this, for a fixed \( (\gamma, \beta) \) with \( \sum_j \gamma_j \beta_j = 1 \), we introduce a parameterization in a similar manner to the smooth case. Let \( A = (0) \) be an \( (H - H_0) \times (H - H_0 + 1) \) matrix such that \( \sum_{j=H_0}^H \alpha_j \gamma_j \beta_j = 0 \) (\( \forall \nu \)) and \( (I_{H - H_0 + 1} - A) \) is invertible. Fix such \( A \) and define \( (a, \xi_{H_0+1}, \ldots, \xi_H; b, \eta_{H_0+1}, \ldots, \eta_H) \) by Eq. (6). The next proposition shows that a small change of \( (\eta_j)_{j=H_0+1}^H \) does not alter the value \( L_H(\theta^{(H)}_{\gamma, \beta}) = L_{H_0}(\theta^{(H_0)}_{\gamma, \beta}) \). Let \( B^\gamma_{\delta}(\theta^{(H)}) \) denote the intersection of the ball of radius \( \delta > 0 \) at \( \theta^{(H)} \) and the affine subspace spanned by \( \eta_{H_0+1}, \ldots, \eta_H \) at \( \theta^{(H)} \).

**Proposition 8.** Let \( \{\mathbf{x}_\nu\}_{\nu=1}^n \) be any data set, \( \theta^{(H)}_{\gamma, \beta} \) be any parameter of the ReLU network \( \mathcal{N}_{H_0} \), and \( \theta_{\gamma, \beta} \) be defined by Eq. (9). Assume that \( \mathbf{u}_{H_0}^{T} \mathbf{x}_\nu \neq 0 \) for all \( \nu \). Then, there is \( \delta > 0 \) such that

\[
f^{(H)}(\mathbf{x}_\nu; \theta^{(H)}_{\gamma, \beta}) = f^{(H_0)}(\mathbf{x}_\nu; \theta_{\gamma, \beta}^{H_0}) \quad (\forall \theta^{(H)}_{\gamma, \beta} \in B^\gamma_{\delta}(\theta^{(H)}_{\gamma, \beta}), \forall \nu = 1, \ldots, n).
\]

See Sec. F.1 for the proof. The situation \( \mathbf{u}_{H_0}^{T} \mathbf{x}_\nu \neq 0 \) may easily occur in practice (Fig. 2(a)).

### 4.2 Embedding a local minimum of ReLU networks

We first consider the embedding of a minimum by inactive units. Let \( \hat{\theta}^{(H)} \) be an embedding of \( \theta^{(H_0)} \) by Eq. (10). From Proposition 7, \( L_H(\hat{\theta}^{(H)}_\gamma) \) does not depend on \( (\mathbf{v}_j, \mathbf{w}_j)_{j=H_0+1}^H \) around \( \hat{\theta}^{(H)}_\gamma \) but takes the same value as \( L_{H_0}(\theta^{(H_0)}_{\gamma, \beta}) \) with \( \theta^{(H_0)}_{\gamma, \beta} = (\mathbf{v}_j, \mathbf{w}_j)_{j=1}^{H_0} \). We have thus the following theorem.

**Theorem 9.** Assume that \( \theta^{(H_0)}_{\gamma, \beta} \) is a minimum of \( L_{H_0} \). Then, the embedded point \( \hat{\theta}^{(H)}_\gamma \) defined by Eq. (10) (inactive units) is a minimum of \( L_H \).

Theorem 9 and Proposition 7 imply that there is an \( (H - H_0) \times (M + D) \) dimensional affine subset that gives local minima, and in those directions \( L_H \) is flat.

Next, we consider the embedding by unit replication, which needs further restriction on \( \gamma \) and \( \beta \). Let \( \theta^{(H_0)}(\gamma, \beta) \) be a parameter of \( \mathcal{N}_{H_0} \), and \( \gamma = (\gamma_j)_{j=H_0}^H \) satisfy \( \sum_{j=H_0}^H \gamma_j > 0 \). Define \( \theta^{(H)}(\gamma, \beta) \) by replacing \( \mathbf{w}_j = \beta_j \mathbf{u}_j \) in Eq. (9) with \( \mathbf{w}_j = \mathbf{u}_{H_0} / \sum_{k=H_0}^H \gamma_k \) (\( H_0 \leq j \leq H \)). If we assume \( \mathbf{u}_{H_0}^{T} \mathbf{x}_\nu \neq 0 \) (\( \forall \nu \)), the function \( L_H \) is differentiable on \( \eta_\nu \), \( \xi_\nu \), and for the same reason as Theorem 5, the derivatives are zero. By restricting the function on those directions around \( \theta^{(H)}(\gamma, \beta) \), from the fact \( \frac{\partial^2 \nu}{\partial \eta_\nu \partial \xi_\nu} = 0 \), we can see that the Hessian has the form \( \left( \begin{array}{cc} F & \mathbf{p} \\ \mathbf{p}^T & 0 \end{array} \right) \), which includes a positive and negative eigenvalue unless \( F = 0 \). This derives the following theorem. (See Sec. F.2 for a complete proof.)

**Theorem 10.** Suppose that \( \theta^{(H_0)}(\gamma, \beta) \) is a minimum point of \( L_{H_0} \). Assume that \( \mathbf{u}_{H_0}^{T} \mathbf{x}_\nu \neq 0 \) for any \( \nu = 1, \ldots, n \), and that \( F \neq 0 \) where \( F \) is given by Lemma 4. Then, for any \( \gamma \in \mathbb{R}^{H - H_0+1} \) such that \( \sum_{j=H_0}^H \gamma_j > 0 \), the embedded parameter \( \theta^{(H)}(\gamma, \beta) \) is a saddle point of \( L_H \).
5 Discussions

5.1 Minimum of zero error

In using a very large network with more parameters than the data size, the training error may reach zero. Assume \( \ell(y, z) \geq 0 \) and that a narrower model attains \( L_{H_0}(\theta_s^{(H_0)}) = 0 \) without redundant units, i.e., any deletion of a unit will increase the training error. We investigate overparameterized realization of such a global minimum by embedding in a wider network \( \mathcal{N}_H \). Note that by any methods the embedded parameter is a minimum. This causes special local properties on the embedded point.

For simplicity, we assume three-layer networks and \( \|x_v\| \leq 1 \) (\( \forall \nu \)). First, consider the unit replication for the smooth activation. As discussed in the last part of Sec. 3.2, the Hessian takes the form

\[
\text{Smooth: } \nabla^2 L_H(\theta^{(H)}_{\chi}) = \eta_{\varepsilon} \left[ \begin{array}{ccc} \nabla^2 L_{H_0}(\theta^{(H_0)}_s) & 0 & 0 \\ 0 & O & O \\ 0 & O & \tilde{G} \end{array} \right],
\]

where \( \tilde{G} \) is non-negative definite. It is not difficult to see (Sec. G.2.2) that, in the case of inactive propagation, the case of inactive propagation is similar.

For ReLU activation, assume \( \theta^{(H_0)}_s \) is a differentiable point of \( L_{H_0} \). From Proposition 7, the Hessian at the embedding \( \tilde{\theta}^{(H)} \) by inactive units is given by

\[
\text{ReLU: } \nabla^2 L_H(\tilde{\theta}^{(H)}) = \left[ \begin{array}{ccc} \theta^{(H_0)}_s & (v_j, w_j) \\ \nabla^2 L_{H_0}(\theta^{(H_0)}_s) & O \\ O & O \end{array} \right].
\]

Similarly to the smooth case, the Hessian for the unit replication \( \theta^{(H)}_s \) takes the same form as Eq. (12).

5.2 Generalization error bounds of embedded networks

Based on the results in Sec. 5.1, here we compare the embedding between ReLU and smooth activation. The results suggest that the ReLU networks can have an advantage in generalization error when zero training error is realized by some type of overparameterized models.

Suppose that the smooth model \( \mathcal{N}_{H_0,v} \) and ReLU model \( \mathcal{N}_{H_0,r} \) attain zero training error without redundant units. They are embedded by the method of inactive units into \( \mathcal{N}_{H_s} \) and \( \mathcal{N}_{H_r} \), respectively, so that \( H_s - H_0,v = H_r - H_0,r (=: E) \) (the same number of surplus units). The dimensionality of the parameters of \( \mathcal{N}_{H_0,s} \) and \( \mathcal{N}_{H_0,r} \) are denoted by \( d_{sm}^0 \) and \( d_{r}^0 \), respectively.

The major difference of the local properties in Eqs. (11) and (12) is the existence of matrix \( S \) or \( \tilde{G} \) in the smooth case. The ReLU network has a flat error surface \( L_H \) in both the directions of \( w_j \) and \( v_j \). In this sense, the embedded minimum is flatter in the ReLU network. We relate this difference of semi-flatness to the generalization ability of the networks through the PAC-Bayes bounds, which has been already used for discussing deep learning [13]. Our motivation here is to consider the difference of the activation functions. We give a summary here and defer the details in Sec. G, Supplements.

Let \( \mathcal{D} \) be a probability distribution of \( (x, y) \) and \( L_H(\theta^{(H)}_s) := E_{\mathcal{D}}[\ell(y, f(x; \theta^{(H)}_s))] \) be the generalization error (or risk). Training data \( (x_1, y_1), \ldots, (x_n, y_n) \) are i.i.d. sample with distribution \( \mathcal{D} \).

Then, with a trained parameter \( \theta \), the PAC-Bayes bound tells

\[
\mathcal{L}_H(\theta) \leq \frac{1}{n} L_H(\tilde{\theta}) + 2 \sqrt{\frac{2(KL(Q\|P) + \ln \frac{2d}{n})}{n-1}},
\]

where \( P \) is a prior distribution which does not depend on the training data, and \( Q \) is any distribution such that it distributes on parameters that do not change the value of \( L_H \) so much from \( L_H(\tilde{\theta}) \).

We focus on the embedding by inactive units here. See Sec. G.2.3, Supplements, for the other cases. The essential factor of the PAC-Bayes bound is the KL-divergence \( KL(Q\|P) \), which is to be small. We use different choices of \( P \) and \( Q \) for the smooth and ReLU networks (see Sec. G for details). For the smooth networks, \( F_{sm} \) is a non-informative normal distribution \( N(0, \sigma^2 I_{d_{sm}}) \) with
Abundance of semi-flat minima in ReLU networks. Theorems 9 and 10 discuss three layer models.

For a better theoretical understanding of the error landscape, this paper has discussed three methods for embedding a network to a wider model, and studied overparameterized realization of a function and its local properties. From the difference of the properties between smooth and ReLU networks, our results suggest that ReLU may have an advantage in realizing zero errors with better generalization. The current analysis reveals some nontrivial geometry of the error landscape, and its implications to dynamics of learning will be within important future works.
References


