

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

A Definitions

Let Ω be an open domain in \mathbb{R}^d , and $f : \Omega \rightarrow \mathbb{R}$ be a differentiable function. $x_0 \in \Omega$ is called a *minimum* (global minimum) if $f(x) \geq f(x_0)$ for all $x \in \Omega$. x_0 is a *local minimum* if there is an open neighborhood U of x_0 such that x_0 is a minimum $f|_U$ in U . A point $x_0 \in \Omega$ is a saddle point if for any open neighborhood of x_0 has y and z such that $f(y) > f(x_0)$ and $f(z) < f(x_0)$. Some literature discuss *flat minima* [3, 4, 1, 5], which are observed to have link with generalization performance. In this paper, we introduce *semi-flat minima*, which is defined as an affine subset V of Ω such that $c = f(x)$ for any $x \in V$ and $f(z) \geq c$ for any $z \in \Omega$.

B Proof of Theorem 2

We show a proof using the original parameterization. We can also use the repameteriation introduced in Section 3.2, which may give other insights on the local properties, but we omit it here. See also Figure 3 for the meaning of parameters.

Recall that the gradients of L_H with respect to the parameters can be given by the *back-propagation*, which computes the derivatives with respect to the weight parameters successively from the output layer to the input. For simplicity we use the notation

$$\ell_\nu(\boldsymbol{\theta}^{(H)}) := \ell(\mathbf{y}_\nu, \mathbf{f}^{(H)}(\mathbf{x}_\nu; \boldsymbol{\theta}^{(H)})). \quad (14)$$

Let $\mathbf{z}^{k,\nu} = (z_1^{k,\nu}, \dots, z_{H_k}^{k,\nu})^T$ be the input to the H_k units in the k -th layer for \mathbf{x}_ν , i.e.,

$$z_i^{k,\nu} = \sum_{j=1}^{H_k} w_{ij}^k \phi(z_j^{k-1,\nu}),$$

where w_{ij}^k is the weight parameter connecting from \mathcal{U}_j^{k-1} to \mathcal{U}_i^k . Let

$$\delta_i^{k,\nu} := \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial z_i^k}.$$

Then, the back-propagation or generalized delta rule [9] computes the derivatives by

$$\delta_j^{k,\nu} = \sum_{i=1}^{H_{k+1}} w_{ij}^{k+1} \delta_i^{k+1} \phi'(z_j^k), \quad \frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial w_{ij}^k} = \sum_{\nu=1}^n \delta_i^{k,\nu} \phi(z_j^{k-1,\nu}). \quad (15)$$

Now consider the embedding using a unit in the q -th layer. Note that the output of any layer except q in $\mathbf{f}^{(H)}(\mathbf{x}; \boldsymbol{\theta}_\lambda^{(H)})$ is equal to that of $\mathbf{f}^{(H_0)}(\mathbf{x}; \boldsymbol{\theta}_*^{(H_0)})$, and the backpropagation of the both networks gives exactly the same $\delta_i^{k,\nu}$ to any $\mathcal{U}_{k,i}$ for $k > q$. It follows that

$$\left. \frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial V_0} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} = \left. \frac{\partial L_{H_0}(\boldsymbol{\theta}^{(H_0)})}{\partial V_0} \right|_{\boldsymbol{\theta}^{(H_0)} = \boldsymbol{\theta}_*^{(H_0)}} = O. \quad (16)$$

The derivatives of L_{H_0} with respect to $\boldsymbol{\zeta}_j$ and \mathbf{u}_j ($1 \leq j \leq H_0$) are given by

$$\frac{\partial L_{H_0}(\boldsymbol{\theta}^{(H_0)})}{\partial \boldsymbol{\zeta}_j} = \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H_0)})}{\partial \mathbf{z}^{q+1,\nu}} \frac{\partial \mathbf{z}^{q+1,\nu}}{\partial \boldsymbol{\zeta}_j} = \sum_{\nu=1}^n \boldsymbol{\delta}^{q+1,\nu} \varphi(\mathbf{x}_\nu; \mathbf{u}_j, W_0) \quad (17)$$

$$\frac{\partial L_{H_0}(\boldsymbol{\theta}^{(H_0)})}{\partial \mathbf{u}_j} = \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H_0)})}{\partial \mathbf{z}^{q+1,\nu}} \frac{\partial \mathbf{z}^{q+1,\nu}}{\partial \mathbf{u}_j} = \sum_{\nu=1}^n \boldsymbol{\delta}^{q+1,\nu T} \boldsymbol{\zeta}_j \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_j, W_0)}{\partial \mathbf{u}_j}, \quad (18)$$

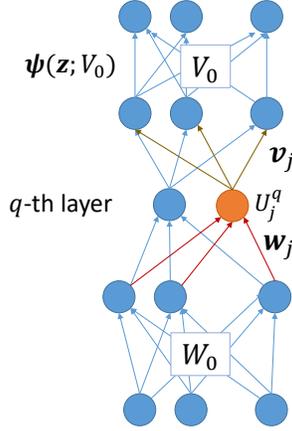


Figure 3: Function of neural networks

where $\delta^{q+1,\nu} = (\delta_1^{q+1,\nu}, \dots, \delta_M^{q+1,\nu})^T$.

In the same manner, for $1 \leq j \leq H_0 - 1$, the derivatives of L_H with respect to \mathbf{u}_j and \mathbf{w}_j are given by

$$\frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{v}_j} = \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}^{q+1,\nu}} \frac{\partial \mathbf{z}^{q+1,\nu}}{\partial \mathbf{v}_j} = \sum_{\nu=1}^n \delta^{q+1,\nu} \varphi(\mathbf{x}_\nu; \mathbf{w}_j, W_0) \quad (19)$$

$$\frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{w}_j} = \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}^{q+1,\nu}} \frac{\partial \mathbf{z}^{q+1,\nu}}{\partial \mathbf{w}_j} = \sum_{\nu=1}^n \delta^{q+1,\nu T} \mathbf{v}_j \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j, W_0)}{\partial \mathbf{w}_j}. \quad (20)$$

It is obvious that these derivatives at $\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}$ are equal to those of L_{H_0} at $\boldsymbol{\theta}_*^{(H_0)}$, and thus equal to zero.

For $H_0 \leq j \leq H$, by the definition of $\boldsymbol{\theta}_\lambda^{(H)}$, we have

$$\frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{v}_j} \Big|_{\boldsymbol{\theta}_\lambda^{(H)}} = \sum_{\nu=1}^n \delta^{q+1,\nu} \varphi(\mathbf{x}_\nu; \mathbf{w}_j, W_0) \Big|_{\boldsymbol{\theta}_\lambda^{(H)}} = \sum_{\nu=1}^n \delta_*^{q+1,\nu} \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0*}, W_{0*}) \quad (21)$$

$$\frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{w}_j} \Big|_{\boldsymbol{\theta}_\lambda^{(H)}} = \sum_{\nu=1}^n \delta^{q+1,\nu T} \mathbf{v}_j \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j, W_0)}{\partial \mathbf{w}_j} \Big|_{\boldsymbol{\theta}_\lambda^{(H)}} = \lambda_j \sum_{\nu=1}^n \delta_*^{q+1,\nu T} \boldsymbol{\zeta}_{H_0*} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0*}, W_{0*})}{\partial \mathbf{u}_{H_0}}, \quad (22)$$

which are zero from the stationary condition of $\boldsymbol{\theta}_*^{(H_0)}$. We have also

$$\begin{aligned} \frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial W_0} \Big|_{\boldsymbol{\theta}_\lambda^{(H)}} &= \sum_{\nu=1}^n \sum_{j=1}^H \delta^{q+1,\nu T} \mathbf{v}_j \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j, W_0)}{\partial W_0} \Big|_{\boldsymbol{\theta}_\lambda^{(H)}} \\ &= \sum_{\nu=1}^n \delta_*^{q+1,\nu T} \sum_{j=1}^{H_0-1} \boldsymbol{\zeta}_{j*} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{j*}, W_{0*})}{\partial W_0} + \sum_{\nu=1}^n \delta_*^{q+1,\nu T} \sum_{j=H_0}^H \lambda_j \boldsymbol{\zeta}_{H_0*} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0*}, W_{0*})}{\partial W_0} \\ &= \sum_{\nu=1}^n \delta_*^{q+1,\nu T} \sum_{j=1}^{H_0} \boldsymbol{\zeta}_{j*} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{j*}, W_{0*})}{\partial W_0} \end{aligned} \quad (23)$$

$$\begin{aligned} &= \frac{\partial L_{H_0}(\boldsymbol{\theta}^{(H_0)})}{\partial W_0} \Big|_{\boldsymbol{\theta}^{(H_0)} = \boldsymbol{\theta}_*^{(H_0)}} \\ &= O, \end{aligned} \quad (24)$$

which completes the proof.

C Embedding by inactive units and propagation for smooth networks

As in Eqs. (17) through (20), stationary conditions for L_{H_0} give, for $1 \leq i \leq H_0$,

$$\begin{aligned}\frac{\partial L_{H_0}(\boldsymbol{\theta}^{(H_0)})}{\partial \boldsymbol{\zeta}_i} &= \sum_{\nu=1}^n \boldsymbol{\delta}^{q+1,\nu} \varphi(\mathbf{x}_\nu; \mathbf{u}_i, W_0) = \mathbf{0} \\ \frac{\partial L_{H_0}(\boldsymbol{\theta}^{(H_0)})}{\partial \mathbf{u}_j} &= \sum_{\nu=1}^n \boldsymbol{\delta}^{q+1,\nu T} \boldsymbol{\zeta}_i \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_i, W_0)}{\partial \mathbf{u}_i} = \mathbf{0}.\end{aligned}\quad (25)$$

The derivatives of L_H with respect to \mathbf{v}_j and \mathbf{w}_j are given by

$$\frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{v}_j} = \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}^{q+1,\nu}} \frac{\partial \mathbf{z}^{q+1,\nu}}{\partial \mathbf{v}_j} = \sum_{\nu=1}^n \boldsymbol{\delta}^{q+1,\nu} \varphi(\mathbf{x}_\nu; \mathbf{w}_j, W_0) \quad (26)$$

$$\frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{w}_j} = \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}^{q+1,\nu}} \frac{\partial \mathbf{z}^{q+1,\nu}}{\partial \mathbf{w}_j} = \sum_{\nu=1}^n \boldsymbol{\delta}^{q+1,\nu T} \mathbf{v}_j \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j, W_0)}{\partial \mathbf{w}_j}. \quad (27)$$

In the case of inactive units, \mathbf{v}_j for $j \geq H_0 + 1$ is arbitrary and the $\frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}^{(0)}, W_0)}{\partial \mathbf{w}_j}$ is not necessarily zero, so that Eq. (25) does not necessarily imply that Eq. (27) is zero. In the case of inactive propagation, \mathbf{w}_j is arbitrary for $j \geq H_0 + 1$, which does not mean Eq. (26) is zero in general.

Consider the embedding by making both of units and propagation inactive; i.e.,

$$\begin{aligned}\mathbf{v}_i &= \boldsymbol{\zeta}_i \quad (1 \leq i \leq H_0) \\ \mathbf{w}_i &= \mathbf{u}_i \quad (1 \leq i \leq H_0) \\ \mathbf{v}_j &= \mathbf{0} \quad (H_0 + 1 \leq j \leq H) \\ \mathbf{w}_j &= \mathbf{w}^{(0)} \quad (H_0 + 1 \leq j \leq H).\end{aligned}\quad (28)$$

Then, for $j \geq H_0 + 1$, we have $\varphi(\mathbf{x}; \mathbf{w}_j, W_0) = 0$ at $\mathbf{w}_j = \mathbf{w}^{(0)}$ which means Eq. (26) is zero, and Eq. (27) vanishes from $\mathbf{v}_j = \mathbf{0}$. Therefore, the stationary point of L_{H_0} is embedded to a stationary point of L_H , but there is no flat direction for this stationary point in general.

D Proofs of Lemmas 3, 4, and Theorem 5 in Section 3

In the sequel, we repeatedly use the following relations.

$$\begin{aligned}\frac{\partial}{\partial \mathbf{b}} &= \sum_{j=H_0}^H \frac{\partial}{\partial \mathbf{w}_j}, & \frac{\partial}{\partial \boldsymbol{\eta}_c} &= \sum_{j=H_0}^H \alpha_{cj} \frac{\partial}{\partial \mathbf{w}_j}, \\ \frac{\partial}{\partial \mathbf{a}} &= \sum_{j=H_0}^H \lambda_j \frac{\partial}{\partial \mathbf{v}_j}, & \frac{\partial}{\partial \boldsymbol{\xi}_c} &= \sum_{k=H_0}^H \lambda_k \alpha_{ck} \frac{\partial}{\partial \mathbf{v}_k}.\end{aligned}\quad (29)$$

D.1 Proof of Lemma 3

It follows from Eq. (29) that

$$\begin{aligned}\left. \frac{\partial \mathbf{f}^{(H)}(\mathbf{x}; \boldsymbol{\theta}^{(H)})}{\partial \mathbf{b}} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} &= \sum_{j=H_0}^H \left. \frac{\partial \mathbf{f}^{(H)}(\mathbf{x}; \boldsymbol{\theta}^{(H)})}{\partial \mathbf{w}_j} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\ &= \sum_{j=H_0}^H \mathbf{v}_j \left. \frac{\partial \varphi(\mathbf{x}; \mathbf{w}_j)}{\partial \mathbf{w}_j} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\ &= \sum_{j=H_0}^H \lambda_j \boldsymbol{\zeta}_{H_0} \frac{\partial \varphi(\mathbf{x}; \mathbf{u}_{H_0})}{\partial \mathbf{u}_{H_0}} \\ &= \frac{\partial \mathbf{f}^{(H_0)}(\mathbf{x}; \boldsymbol{\theta}_*^{(1)})}{\partial \mathbf{u}_{H_0}},\end{aligned}$$

since $\sum_j \lambda_j = 1$. Also,

$$\begin{aligned} \frac{\partial \mathbf{f}^{(H)}(\mathbf{x}; \boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} &= \sum_{j=H_0}^H \alpha_{cj} \frac{\partial \mathbf{f}^{(H)}(\mathbf{x}; \boldsymbol{\theta}^{(H)})}{\partial \mathbf{w}_j} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\ &= \sum_{j=H_0}^H \alpha_{cj} \lambda_j \zeta_{H_0} \frac{\partial \varphi(\mathbf{x}; \mathbf{u}_{H_0})}{\partial \mathbf{u}_{H_0}} = 0, \end{aligned}$$

since $\sum_j \alpha_{cj} \lambda_j = 0$ by definition of A .

From Eq. (29), we have

$$\begin{aligned} \frac{\partial \mathbf{f}^{(H)}(\mathbf{x}; \boldsymbol{\theta}^{(H)})}{\partial \mathbf{a}} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} &= \sum_{j=H_0}^H \lambda_j \frac{\partial \mathbf{f}^{(H)}(\mathbf{x}; \boldsymbol{\theta}^{(H)})}{\partial \mathbf{v}_j} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\ &= \sum_{j=H_0}^H \lambda_j \varphi(\mathbf{x}; \mathbf{w}_j) I \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\ &= \varphi(\mathbf{x}; \mathbf{u}_{H_0, *}) I \\ &= \frac{\partial \mathbf{f}^{(H_0)}(\mathbf{x}; \boldsymbol{\theta}_*^{(H_0)})}{\partial \zeta_{H_0}}, \end{aligned}$$

and

$$\begin{aligned} \frac{\partial \mathbf{f}^{(H)}(\mathbf{x}; \boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\xi}_c} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} &= \sum_{k=H_0}^H \lambda_k \alpha_{ck} \frac{\partial \mathbf{f}^{(H)}(\mathbf{x}; \boldsymbol{\theta}^{(H)})}{\partial \mathbf{v}_k} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\ &= \sum_{k=H_0}^H \lambda_k \alpha_{ck} \varphi(\mathbf{x}; \mathbf{u}_{H_0}) I = 0. \end{aligned}$$

D.2 Proof of Lemma 4

We use the notation

$$\mathbf{z}_\nu = \mathbf{f}^{(H)}(\mathbf{x}_\nu; \boldsymbol{\theta}^{(H)}).$$

(i) First, we compute the blocks related to the derivative with respect to $\boldsymbol{\eta}$. We have

$$\frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c} = \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu} \frac{\partial \mathbf{z}_\nu}{\partial \boldsymbol{\eta}_c} = \sum_{\nu=1}^n \sum_{m=1}^M \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial z_{\nu, m}} \sum_{j=H_0}^H \alpha_{cj} v_{jm} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j}. \quad (30)$$

It follows from Eqs. (29) and (30) that

$$\begin{aligned} \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c \partial \mathbf{a}} &= \sum_{k=H_0}^H \lambda_k \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c \partial \mathbf{v}_k} \\ &= \sum_{\nu=1}^n \sum_{m=1}^M \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial z_\nu \partial z_{\nu, m}} \sum_{k=H_0}^H \lambda_k \varphi(\mathbf{x}_\nu; \mathbf{w}_k) \sum_{j=H_0}^H \alpha_{cj} v_{jm} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j} \\ &\quad + \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu} \sum_{k=H_0}^H \alpha_{ck} \lambda_k \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_k)}{\partial \mathbf{w}_k}. \end{aligned} \quad (31)$$

By inserting $\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}$, the first term is zero since $v_j = \lambda_j \zeta_*$ and $\sum_j \alpha_{cj} \lambda_j = 0$. The second term is also zero from $\sum_k \alpha_{ck} \lambda_k = 0$.

Differentiation of Eq. (30) with \mathbf{b} gives

$$\begin{aligned} \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c \partial \mathbf{b}} &= \sum_{\nu=1}^n \sum_{m, m'=1}^M \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial z_{\nu, m} \partial z_{\nu, m'}} \sum_{k=H_0}^H v_{km'} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_k)}{\partial \mathbf{w}_k} \sum_{j=H_0}^H \alpha_{cj} v_{jm} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j} \\ &+ \delta_{jk} \sum_{\nu=1}^n \sum_{m=1}^M \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial z_{\nu, m}} \sum_{j=H_0}^H \alpha_{cj} v_{jm} \sum_{k=H_0}^H \frac{\partial^2 \varphi(\mathbf{x}_\nu; \mathbf{w}_k)}{\partial \mathbf{w}_k \partial \mathbf{w}_k}. \end{aligned} \quad (32)$$

At $\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}$, both the terms are zero for the same reason as Eq. (31).

Similarly, for $\mathbf{s}_i = \mathbf{v}_i$ or \mathbf{w}_i ($1 \leq i \leq H_0 - 1$),

$$\begin{aligned} \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c \partial \mathbf{s}_i} &= \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c \partial \mathbf{s}_i} \\ &= \sum_{\nu=1}^n \sum_{m=1}^M \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu \partial z_{\nu, m}} \frac{\partial \mathbf{z}_\nu}{\partial \mathbf{s}_i} \sum_{j=H_0}^H \alpha_{cj} v_{jm} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j}, \end{aligned}$$

which is zero at $\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}$ from $\sum_j \alpha_{cj} \lambda_j = 0$.

Next, from Eqs. (29) and (30), we have

$$\begin{aligned} \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c \partial \boldsymbol{\xi}_d} &= \sum_{\nu=1}^n \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu \partial \mathbf{z}_\nu} \sum_{k=H_0}^H \alpha_{dk} \lambda_k \varphi(\mathbf{x}_\nu; \mathbf{w}_k) \sum_{j=H_0}^H \alpha_{cj} \mathbf{v}_j \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j} \\ &+ \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu} \sum_{k=H_0}^H \alpha_{dk} \alpha_{ck} \lambda_k \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_k)}{\partial \mathbf{w}_k}. \end{aligned} \quad (33)$$

At $\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}$, the first term vanishes and the second term reduces to

$$\frac{\partial^2 L_H(\boldsymbol{\theta}_\lambda^{(H)})}{\partial \boldsymbol{\eta}_c \partial \boldsymbol{\xi}_d} = (A \Lambda A^T)_{cd} \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}_*^{(H_0)})}{\partial \mathbf{z}_\nu} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0, *})}{\partial \mathbf{u}_{H_0}},$$

which is $(A \Lambda A^T)_{cd} F$.

The block $\frac{L_H(\boldsymbol{\theta}_\lambda^{(H)})}{\partial \boldsymbol{\eta}_c \partial \boldsymbol{\eta}_d}$ can be computed in a similar way to Eq. (33):

$$\begin{aligned} \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c \partial \boldsymbol{\eta}_d} &= \sum_{\nu=1}^n \sum_{m, m'=1}^M \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial z_{\nu, m} \partial z_{\nu, m'}} \sum_{j=H_0}^H \alpha_{cj} v_{jm} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j} \sum_{k=H_0}^H \alpha_{dk} v_{km'} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_k)}{\partial \mathbf{w}_k} \\ &+ \sum_{\nu=1}^n \sum_{m=1}^M \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_{\nu, m}} \sum_{j=H_0}^H \alpha_{cj} \alpha_{dj} v_{jm} \frac{\partial^2 \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j \partial \mathbf{w}_j}. \end{aligned}$$

By plugging $\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}$, the first term is zero, and the second term is reduced to

$$\sum_{j=H_0}^H \lambda_j \alpha_{cj} \alpha_{dj} \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu} \boldsymbol{\zeta}_{H_0, *}, \quad (34)$$

which is $(A \Lambda A^T)_{cd} G$.

(ii) Second, we will compute the remaining second derivatives including $\boldsymbol{\xi}_c$. From Eq. (29), the first derivative with respect to $\boldsymbol{\xi}_c$ is given by

$$\frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\xi}_c} = \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu} \sum_{j=H_0}^H \lambda_j \alpha_{cj} \varphi(\mathbf{x}_\nu; \mathbf{w}_j). \quad (35)$$

From this expression,

$$\begin{aligned} \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\xi}_c \partial \mathbf{v}_k} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} &= \sum_{\nu=1}^n \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}_\lambda^{(H)})}{\partial \mathbf{z}_\nu \partial \mathbf{z}_\nu} \sum_{j=H_0}^H \lambda_j \alpha_{cj} (\varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0,*}))^2 \\ &= 0, \end{aligned}$$

which means $\frac{\partial^2 L_H(\boldsymbol{\theta}_\lambda^{(H)})}{\partial \boldsymbol{\xi}_c \partial \boldsymbol{\xi}_d}$ and $\frac{\partial^2 L_H(\boldsymbol{\theta}_\lambda^{(H)})}{\partial \boldsymbol{\xi}_c \partial \mathbf{a}}$ are zero.

It follows from Eqs. (35) and (29) that

$$\begin{aligned} &\frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\xi}_c \partial \mathbf{b}} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\ &= \sum_{\nu=1}^n \sum_{m=1}^M \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}_\lambda^{(H)})}{\partial \mathbf{z}_\nu \partial z_{\nu,m}} \sum_{j=H_0}^H \lambda_j \alpha_{cj} \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0,*}) \sum_{k=H_0}^H v_{km} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0,*})}{\partial \mathbf{u}_{H_0}} \\ &\quad + \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu} \sum_{j=H_0}^H \lambda_j \alpha_{cj} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0,*})}{\partial \mathbf{u}_{H_0}}, \end{aligned}$$

which is zero from $\sum_j \alpha_{cj} \lambda_j = 0$.

It is also easy to see that for $\mathbf{s}_i = \mathbf{v}_i$ or \mathbf{w}_i ($1 \leq i \leq H_0 - 1$)

$$\frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\xi}_c \partial \mathbf{s}_i} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} = \mathbf{0}.$$

(III) We compute the upper-left four blocks. We have

$$\frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{a}} = \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu} \sum_{j=H_0}^H \lambda_j \varphi(\mathbf{x}_\nu; \mathbf{w}_j), \quad (36)$$

from which

$$\frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{a} \partial \mathbf{a}} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} = \sum_{\nu=1}^n \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}_*^{(H_0)})}{\partial \mathbf{z}_\nu \partial \mathbf{z}_\nu} \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0,*})^2 = \frac{\partial^2 L_{H_0}(\boldsymbol{\theta}_*^{(H_0)})}{\partial \boldsymbol{\zeta}_{H_0} \partial \boldsymbol{\zeta}_{H_0}}$$

and

$$\begin{aligned} &\frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{a} \partial \mathbf{b}} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\ &= \sum_{\nu=1}^n \sum_{m=1}^M \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu \partial z_{\nu,m}} \sum_{j=H_0}^H \lambda_j \varphi(\mathbf{x}_\nu; \mathbf{w}_j) \sum_{k=H_0}^H v_{km} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_k)}{\partial \mathbf{w}_k} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\ &\quad + \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu} \sum_{j=H_0}^H \lambda_j \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\ &= \sum_{\nu=1}^n \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}_*^{(H_0)})}{\partial \mathbf{z}_\nu \partial \mathbf{z}_\nu} \boldsymbol{\zeta}_{H_0,*} \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0,*}) \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0,*})}{\partial \mathbf{u}_{H_0}} + \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}_*^{(H_0)})}{\partial \mathbf{z}_\nu} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0,*})}{\partial \mathbf{u}_{H_0}} \\ &= \frac{\partial^2 L_{H_0}(\boldsymbol{\theta}_*^{(H_0)})}{\partial \boldsymbol{\zeta}_{H_0} \partial \mathbf{u}_{H_0}}. \end{aligned}$$

Finally, using

$$\frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{b}} = \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{z}_\nu} \frac{\partial \mathbf{z}_\nu}{\partial \mathbf{b}} = \sum_{\nu=1}^n \sum_{m=1}^M \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial z_{\nu,m}} \sum_{j=H_0}^H v_{jm} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j},$$

we have

$$\begin{aligned}
& \left. \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{b} \partial \mathbf{b}} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\
&= \sum_{\nu=1}^n \sum_{m, m'=1}^M \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial z_{\nu, m} \partial z_{\nu, m'}} \sum_{j=H_0}^H v_{jm} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j} \sum_{k=1}^H v_{km'} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}_k)}{\partial \mathbf{w}_k} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\
&\quad + \sum_{\nu=1}^n \sum_{m=1}^M \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial z_{\nu, m}} \sum_{j=H_0}^H v_{jm} \frac{\partial^2 \varphi(\mathbf{x}_\nu; \mathbf{w}_j)}{\partial \mathbf{w}_j \partial \mathbf{w}_j} \Big|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} \\
&= \sum_{\nu=1}^n \left(\zeta_{H_0, *}^T \frac{\partial^2 \ell_\nu(\boldsymbol{\theta}_*^{(H_0)})}{\partial \mathbf{z} \partial \mathbf{z}} \zeta_{H_0, *} \right) \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0, *})}{\partial \mathbf{u}_{H_0}} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0, *})}{\partial \mathbf{u}_{H_0}} \\
&\quad + \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}^{(H)})}{\partial z_\nu} \zeta_{H_0, *} \frac{\partial^2 \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0, *})}{\partial \mathbf{u}_{H_0} \mathbf{u}_{H_0}} \\
&= \frac{\partial^2 L_{H_0}(\boldsymbol{\theta}_*^{(H_0)})}{\partial \mathbf{u}_{H_0} \partial \mathbf{u}_{H_0}}.
\end{aligned}$$

(iv) Finally, it is similarly proved that for $\mathbf{s}_i = \mathbf{v}_i$ or \mathbf{w}_i ($1 \leq i \leq H_0 - 1$)

$$\begin{aligned}
\left. \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{a} \partial \mathbf{s}_i} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} &= \frac{\partial^2 L_{H_0}(\boldsymbol{\theta}_*^{(H_0)})}{\partial \zeta_{H_0} \partial \mathbf{s}_i} \\
\left. \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \mathbf{b} \partial \mathbf{s}_i} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\lambda^{(H)}} &= \frac{\partial^2 L_{H_0}(\boldsymbol{\theta}_*^{(H_0)})}{\partial \mathbf{u}_{H_0} \partial \mathbf{s}_i}.
\end{aligned}$$

This completes the proof.

D.3 Proof of Theorem 5

Let $\tilde{F} := (A\Lambda A^T) \otimes F$ and $\tilde{G} := (A\Lambda A^T) \otimes G$. Since $\lambda_j \neq 0$ ($\forall j$) and A is of full rank, $(A\Lambda A^T)$ is of full rank. (i) Under the assumption, \tilde{G} is invertible. Then, the lower-right four blocks of the Hessian has the expression

$$\begin{pmatrix} I & -\tilde{F}\tilde{G}^{-1} \\ O & I \end{pmatrix} \begin{pmatrix} O & \tilde{F} \\ \tilde{F}^T & \tilde{G} \end{pmatrix} \begin{pmatrix} I & O \\ -\tilde{F}\tilde{G}^{-1} & I \end{pmatrix} = \begin{pmatrix} -\tilde{F}^T \tilde{G}^{-1} \tilde{F} & O \\ O & \tilde{G} \end{pmatrix}. \quad (37)$$

If G is positive definite, so is \tilde{G} , and thus $-\tilde{F}^T \tilde{G}^{-1} \tilde{F}$ has negative eigenvalues for $F \neq O$. The Hessian of L_H at $\boldsymbol{\theta}_\lambda^{(H)}$ has both of positive and negative eigenvalues, which implies $\boldsymbol{\theta}_\lambda^{(H)}$ is a saddle point. The case of negative definite G is similar. (ii) If G has positive and negative definite, so does \tilde{G} . This means that the Hessian of L_H at $\boldsymbol{\theta}_\lambda^{(H)}$ has positive and negative eigenvalues. \square

Remark. If F is of full rank, which is $r = (H - H_0) \min\{D, M\}$, then the matrix $\tilde{F}^T \tilde{G}^{-1} \tilde{F}$ has r positive eigenvalues. Thus, the number of positive and negative eigenvalues of the matrix in Eq. (37) are $(H - H_0) \times D$ and r , respectively. When F has positive and negative eigenvalues, the index depends on the eigenvectors of \tilde{F} and \tilde{G} , and not easy to tell.

E Local minima for smooth networks of 1-dimensional output

The special property of $M = 1$ is caused by vanishing \tilde{F} in the Hessian. In fact, the stationarity condition $\frac{\partial L_{H_0}(\boldsymbol{\theta}_*^{(H_0)})}{\partial \mathbf{u}_{H_0}} = 0$ implies

$$\zeta_{H_0, *} \sum_{\nu=1}^n \frac{\partial \ell_\nu(\boldsymbol{\theta}_*^{(H_0)})}{\partial z_\nu} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0, *})}{\partial \mathbf{u}_{H_0}} = 0.$$

Note that ζ_{H_0} is a scalar, and if we assume $\zeta_{H_0*} \neq 0$, the above condition implies $F = 0$. Then the corresponding part of the Hessian takes the form

$$\begin{pmatrix} O & O \\ O & \tilde{G} \end{pmatrix},$$

which does not have negative eigenvalues if G is non-negative definite. The zero blocks of the Hessian correspond to the directions ξ_c ($c = H_0 + 1, \dots, H$), which make an affine subspace of $\Pi_{\text{repl}}(\theta_\lambda^{(H_0)})$ having the same value $L_H(\theta^{(H)}) = L_{H_0}(\theta_*^{(H_0)})$. Therefore, only the Hessian in the directions $(\mathbf{a}, \mathbf{b}, \boldsymbol{\eta}_{H_0+1}, \dots, \boldsymbol{\eta}_H)$ matters to determine if $\theta_\lambda^{(H)}$ is a minimum or saddle point. Note also that for $M \geq 2$ the stationarity condition gives

$$\sum_{\nu=1}^n \sum_{m=1}^M \frac{\partial \ell_\nu(\theta_*^{(H_0)})}{\partial z_{\nu,m}} \zeta_{H_0,m*} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0,*})}{\partial \mathbf{u}_{H_0}} = 0,$$

which does not necessary mean $F = O$.

The following theorem is a slight extension of Fukumizu and Amari [2, Theorem 3], in which only the case $H = H_0 + 1$ is discussed.

Theorem 11. *Suppose that the dimension of the output is 1 and $\theta_*^{(H_0)}$ is a minimum of L_{H_0} with positive definite Hessian matrix. In the following, the matrix G and the parameter $\theta_\lambda^{(H)}$ are used in the same meaning as in Lemma 4.*

(1) *Assume that the matrix G is positive definite.*

- (a) $\theta_\lambda^{(H)}$ with $\sum_{j=H_0}^H \lambda_j = 1$ and $\lambda_j > 0$ ($\forall j$) is a minimum of L_H .
- (b) $\theta_\lambda^{(H)}$ with $\sum_{j=H_0}^H \lambda_j = 1$ and $\lambda_j < 0$ for some j is a saddle point of L_H .

(2) *Assume that the matrix G is negative definite.*

- (a) If $\sum_{j=H_0}^H \lambda_j = 1$ and there is only one i_0 such that $\lambda_{i_0} > 0$ and $\lambda_j < 0$ ($\forall j \neq i_0$), $\theta_\lambda^{(H)}$ is a minimum of L_H .
- (b) If $\sum_{j=H_0}^H \lambda_j = 1$ and $\lambda_j > 0$ for at least two indices, $\theta_\lambda^{(H)}$ is a saddle point of L_H .

(3) *If the matrix G has both of positive and negative eigenvalues, $\theta_\lambda^{(H)}$ is a saddle point for any $\boldsymbol{\lambda}$ with $\sum_{a=H_0}^H \lambda_a = 1$ and $\lambda_a \neq 0$ ($\forall a$).*

Proof. For notational simplicity, the proof is given only for $H_0 = 1$; ζ_1 and \mathbf{u}_1 are written by ζ and \mathbf{u} , respectively. Extension to a general H_0 is easy and we omit it. In the proof, let $\tilde{A}^T := (\mathbf{1}_H A^T)$, which is invertible by assumption. Note also that ζ, v_j are scalar parameters in the case of $M = 1$.

(1-a). We first show that if G is positive definite, the lower-right block of the Hessian, $\frac{\partial^2 L_H(\theta_\lambda^{(H)})}{\partial \boldsymbol{\eta} \partial \boldsymbol{\eta}} = (A \Lambda A^T) \otimes G$, is positive definite. This can be proved if $A \Lambda A^T$ is positive definite, since the eigenvalues of the tensor product is given by the products of respective eigenvalues of $A \Lambda A^T$ and G . By the assumptions, $A \Lambda A^T$ is non-negative definite. Suppose $A \Lambda A^T \mathbf{s} = 0$ for $\mathbf{s} \in \mathbb{R}^{H-1} \setminus \{0\}$. Then, $A^T \mathbf{s} = 0$, and this implies $\tilde{A}^T \tilde{\mathbf{s}} = 0$ for $\tilde{\mathbf{s}} = (\mathbf{s}^T, 0)^T \in \mathbb{R}^H$. This is impossible by the invertible assumption of \tilde{A} .

Now consider the Hessian $\nabla^2 L_H(\theta_\lambda^{(H)})$ in Lemma 4. It is obvious that this Hessian is non-negative definite, but not positive definite, as the blocks corresponding to $(\xi_j)_{j=2}^H$ are zero. Let $\Pi_{\theta_*^{(H_0)}}$ be the $(H - 1)$ dimensional affine plane in the parameter space of \mathcal{N}_H such that

$$\Pi_{\theta_*^{(H_0)}} := \{(a, \xi_2, \dots, \xi_H; \mathbf{b}, \boldsymbol{\eta}_2, \dots, \boldsymbol{\eta}_H) \mid a = \zeta_*, \mathbf{b} = \mathbf{u}_*, \boldsymbol{\eta}_2 = \dots = \boldsymbol{\eta}_H = 0\}.$$

This plane includes $\theta_\lambda^{(H)}$, and is parallel to the subspace spanned by ξ_j axes. The function L_H takes the same value as $L_1(\theta_*^{(1)})$ on the whole of $\Pi_{\theta_*^{(H_0)}}$. Thus, $\theta_\lambda^{(H)}$ is a minimum of L_H if the Hessian

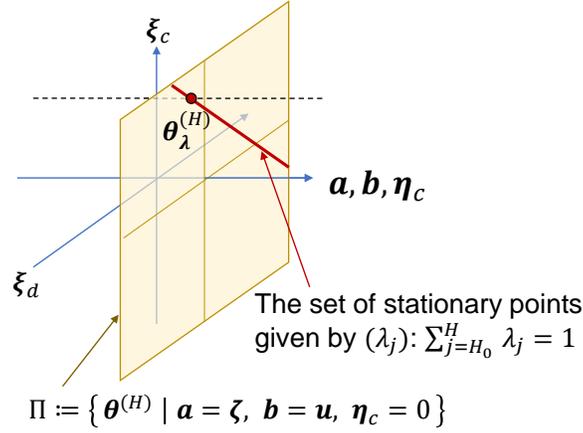


Figure 4: All the parameters on the affine subspace Π has the same function as $f^{(H)}(x; \theta_\lambda^{(H)})$, and the affine subspace (in red) is a set of stationary points of $L_H(\theta^{(H)})$. The local behavior of L_H around $\theta_\lambda^{(H)}$ is determined by the second derivative along the $\mathbf{a}, \mathbf{b}, \eta_c$ directions.

is positive definite along the directions compliment to $\Pi_{\theta_*^{(H_0)}}$ (see Figure 4). From Lemma 4, the Hessian at $\theta_\lambda^{(H)}$ along the directions $(\mathbf{a}, \mathbf{b}, \eta_j)$ is given by

$$\begin{pmatrix} \frac{\partial^2 L_1(\theta_*^{(1)})}{\partial \theta_*^{(H_0)} \partial \theta_*^{(H_0)}} & O \\ O & (A\Lambda A^T) \otimes G \end{pmatrix},$$

which is positive definite. This completes the proof of (1-a).

(1-b) From $A\lambda = 0$, it is easy to see that

$$\tilde{A}\Lambda\tilde{A}^T = \begin{pmatrix} 1 & 0 \\ 0^T & A\Lambda A^T \end{pmatrix}.$$

Thus, the eigenvalues of $\tilde{A}\Lambda\tilde{A}^T$ is the eigenvalues of $A\Lambda A^T$ and 1. By Sylvester's law of inertia, the signature (the pair of the number of positive eigenvalues and that of negative ones) of $\tilde{A}\Lambda\tilde{A}^T$ coincides with the signature of Λ . Since some λ_i are negative by the assumption, $A\Lambda A^T$ has a negative eigenvalue. Thus, under the assumption that G is positive definite, $(A\Lambda A^T) \otimes G$ has a negative eigenvalue. Since $\frac{\partial^2 L_H(\theta_\lambda^{(H)})}{\partial \mathbf{a} \partial \mathbf{a}}$ is positive definite, the Hessian of $L_H(\theta^{(H)})$ at $\theta^{(H)} = \theta_\lambda^{(H)}$ has positive and negative eigenvalues, which means $\theta_\lambda^{(H)}$ is a saddle point.

(2-a) It suffices to show that $A\Lambda A^T$ is negative definite. Then, $(A\Lambda A^T) \otimes G$ is positive definite, and the assertion is proved by the same argument as (1-a). Without loss of generality, we can assume that $\lambda_j < 0$ for $1 \leq j \leq H-1$ and $\lambda_H > 0$. Let $A = (A_0, \mathbf{h})$ where A_0 is an invertible matrix of size $H-1$, and let $\boldsymbol{\lambda}^T = (\boldsymbol{\lambda}_0^T, \lambda_H)$ with $\boldsymbol{\lambda}_0 \in \mathbb{R}^{H-1}$. The elements of $\boldsymbol{\lambda}_0$ are all negative by assumption. It follows that

$$A_0 \boldsymbol{\lambda}_0 + \lambda_H \mathbf{h} = \mathbf{0}, \quad \sum_{j=1}^H \lambda_j = 1.$$

A simple computation using $\mathbf{h} = -\frac{1}{\lambda_H} A_0 \boldsymbol{\lambda}_0$ provides

$$A\Lambda A^T = A_0 \left(\Lambda_0 + \frac{1}{\lambda_H} \boldsymbol{\lambda}_0 \boldsymbol{\lambda}_0^T \right) A_0^T,$$

where $\Lambda_0 = \text{Diag}(\lambda_1, \dots, \lambda_{H-1})$. It is then sufficient to show that $B_0 := \Lambda_0 + \frac{1}{\lambda_H} \boldsymbol{\lambda}_0 \boldsymbol{\lambda}_0^T$ is negative definite. If $\mathbf{s} \in \mathbb{R}^{H-1} \setminus \{0\}$ is orthogonal to $\boldsymbol{\lambda}_0$, we have $\mathbf{s}^T B_0 \mathbf{s} = \mathbf{s}^T \Lambda_0 \mathbf{s} < 0$. Additionally,

$$\begin{aligned}
\boldsymbol{\lambda}_0^T B_0 \boldsymbol{\lambda}_0 &= \sum_{j=1}^{H-1} \lambda_j^3 + \frac{1}{\lambda_H} \left(\sum_{j=1}^{H-1} \lambda_j^2 \right)^2 \\
&= \frac{1}{\lambda_H} \left\{ \left(1 - \sum_{j=1}^{H-1} \lambda_j \right) \left(\sum_{j=1}^{H-1} \lambda_j^3 \right) + \left(\sum_{j=1}^{H-1} \lambda_j^2 \right)^2 \right\} \\
&= \frac{1}{\lambda_H} \left\{ \left(\sum_{j=1}^{H-1} \lambda_j^3 \right) + \sum_{i \neq j} \lambda_i^2 \lambda_j^2 - \sum_{i \neq j} \lambda_i \lambda_j^3 \right\} \\
&= \frac{1}{\lambda_H} \left\{ \left(\sum_{j=1}^{H-1} \lambda_j^3 \right) + \sum_{i \neq j} \lambda_i^2 \lambda_j^2 - \sum_{i \neq j} \lambda_i \lambda_j \frac{\lambda_i^2 + \lambda_j^2}{2} \right\} \\
&= \frac{1}{\lambda_H} \left\{ \left(\sum_{j=1}^{H-1} \lambda_j^3 \right) - \sum_{i \neq j} \frac{1}{2} \lambda_i \lambda_j (\lambda_i - \lambda_j)^2 \right\},
\end{aligned}$$

which is negative as well. This proves the assertion.

(2-b) If there are two positive eigenvalues, the corresponding eigenspaces of at least two dimensions must intersect with the $H - 1$ dimensional subspace spanned by the row vectors of A . Thus, $A \Lambda A^T$ has at least one positive eigenvalue, which means $(A \Lambda A^T) \otimes G$ has negative eigenvalues. The remaining proof is similar to (1-b).

(3) $A \Lambda A^T$ is of full rank, and thus $(A \Lambda A^T) \otimes G$ has both of positive and negative eigenvalues. The assertion is proved by the same argument as the case (1-b). \square

F Proof of Proposition 8 and Theorem 10 in Section 4

F.1 Proof of Proposition 8

First, note that, from $\mathbf{u}_{H_0,*}^T \mathbf{x}_\nu \neq 0 (\forall \nu)$, there is $\delta > 0$ such that for each \mathbf{x}_ν the sign of $(\mathbf{u}_{H_0,*} + \sum_{c=H_0+1}^H \alpha_{cj} \boldsymbol{\eta}_c)^T \mathbf{x}_\nu$ equals to that of $\mathbf{u}_{H_0,*}^T \mathbf{x}_\nu$ for any $j = H_0, \dots, H$ and $(\boldsymbol{\eta}_c)_{c=H_0+1}^H$ such that $\|(\boldsymbol{\eta}_{H_0+1}, \dots, \boldsymbol{\eta}_H)\| \leq \delta$.

Fix \mathbf{x}_ν , and assume first $\mathbf{u}_{H_0,*}^T \mathbf{x}_\nu > 0$. Then, $(\mathbf{u}_{H_0,*} + \sum_{c=H_0+1}^H \alpha_{cj} \boldsymbol{\eta}_c)^T \mathbf{x}_\nu > 0$ holds for $(\boldsymbol{\eta}_c)_c$ with $\|(\boldsymbol{\eta}_c)_c\| \leq \delta$. With the notation

$$\mathcal{F}_{H_0} := \sum_{i=1}^{H_0-1} \mathbf{v}_i \varphi(\mathbf{x}_\nu; \mathbf{w}_i) = \sum_{i=1}^{H_0-1} \zeta_{i,*} \varphi(\mathbf{x}_\nu; \mathbf{u}_{i,*}), \quad (38)$$

for any $\boldsymbol{\theta}^{(H)} \in B_\delta^\eta(\boldsymbol{\theta}_{\gamma,\beta}^{(H)})$, we have

$$\begin{aligned}
\mathbf{f}^{(H)}(\mathbf{x}_\nu; \boldsymbol{\theta}^{(H)}) &= \mathcal{F}_{H_0} + \sum_{j=H_0}^H \gamma_j \zeta_{H_0,*} \varphi \left(\beta_j \left(\mathbf{u}_{H_0,*} + \sum_{c=H_0+1}^H \alpha_{cj} \boldsymbol{\eta}_c \right)^T \mathbf{x}_\nu \right) \\
&= \mathcal{F}_{H_0} + \sum_{j=H_0}^H \gamma_j \zeta_{H_0,*} \beta_j \left(\mathbf{u}_{H_0,*} + \sum_{c=H_0+1}^H \alpha_{cj} \boldsymbol{\eta}_c \right)^T \mathbf{x}_\nu \\
&= \mathcal{F}_{H_0} + \sum_{j=H_0}^H \gamma_j \beta_j \zeta_{H_0,*} \mathbf{u}_{H_0,*}^T \mathbf{x}_\nu + \zeta_{H_0,*} \sum_{c=H_0+1}^H \sum_{j=H_0}^H \alpha_{cj} \gamma_j \beta_j \boldsymbol{\eta}_c^T \mathbf{x}_\nu \\
&= \mathcal{F}_{H_0} + \zeta_{H_0,*} \mathbf{u}_{H_0,*}^T \mathbf{x}_\nu \\
&= \mathbf{f}^{(H_0)}(\mathbf{x}_\nu; \boldsymbol{\theta}_*^{(H_0)}),
\end{aligned}$$

where we used $\sum_j \gamma_j \beta_j = 1$ and $\sum_j \alpha_{cj} \gamma_j \beta_j = 0$.

Next, if $\mathbf{u}_{H_0, *}^T \mathbf{x}_\nu < 0$, we have

$$\mathbf{f}^{(H_0)}(\mathbf{x}_\nu; \boldsymbol{\theta}_*^{(H_0)}) = \mathcal{F}_{H_0},$$

and

$$\mathbf{f}^{(H)}(\mathbf{x}_\nu; \boldsymbol{\theta}^{(H)}) = \mathcal{F}_{H_0} + \sum_{j=H_0}^H \gamma_j \zeta_{H_0, *} \varphi \left(\beta_j (\mathbf{u}_{H_0, *} + \sum_{c=H_0+1}^H \alpha_{cj} \boldsymbol{\eta}_c)^T \mathbf{x}_\nu \right) = \mathcal{F}_{H_0},$$

which completes the proof.

F.2 Proof of Theorem 10

We use the same reparameterization $(\mathbf{v}_1, \dots, \mathbf{v}_{H_0-1}, \mathbf{a}, \mathbf{w}_1, \dots, \mathbf{w}_{H_0-1}, \mathbf{b}, \boldsymbol{\xi}_{H_0+1}, \dots, \boldsymbol{\xi}_H, \boldsymbol{\eta}_{H_0+1}, \dots, \boldsymbol{\eta}_H)$ as in Section 4.1 with $A\boldsymbol{\gamma} = 0$. We focus on the behavior of L_H for a change of $\boldsymbol{\xi}_c, \boldsymbol{\eta}_c$ with the others fixed at the values of $\boldsymbol{\theta}_\gamma^{(H)}$. Note that, by the assumption $\mathbf{u}_{H_0}^T \mathbf{x}_\nu \neq 0$ for any ν , $L_H(\boldsymbol{\theta}^{(H)})$ is differentiable at $\boldsymbol{\theta}_\gamma^{(H)}$ with respect to $\boldsymbol{\xi}_c, \boldsymbol{\eta}_c$. By the same manner as Lemma 3, we have

$$\left. \frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\gamma^{(H)}} = O, \quad \left. \frac{\partial L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\xi}_c} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\gamma^{(H)}} = O,$$

which means L_H is stationary at $\boldsymbol{\theta}_\gamma^{(H)}$ as a function of $\boldsymbol{\eta}_c$ and $\boldsymbol{\xi}_c$.

From Lemma 4, we have

$$\left. \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\xi}_c \partial \boldsymbol{\xi}_d} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\gamma^{(H)}} = O$$

and

$$\left. \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\xi}_c \partial \boldsymbol{\eta}_d} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\gamma^{(H)}} = (A\Lambda A^T)_{cd} \sum_{\nu: \mathbf{u}_{H_0, *}^T \mathbf{x}_\nu > 0} \frac{\partial \ell_\nu(\boldsymbol{\theta}_\gamma^{(H)})}{\partial \mathbf{z}_\nu} \mathbf{x}_\nu^T.$$

Using the fact $\frac{\partial^2 \varphi(\mathbf{x}_\nu; \mathbf{u}_{H_0, *})}{\partial \mathbf{u}_{H_0} \partial \mathbf{u}_{H_0}} = 0$, we have

$$\left. \frac{\partial^2 L_H(\boldsymbol{\theta}^{(H)})}{\partial \boldsymbol{\eta}_c \partial \boldsymbol{\eta}_d} \right|_{\boldsymbol{\theta}^{(H)} = \boldsymbol{\theta}_\gamma^{(H)}} = O.$$

Therefore, the Hessian of L_H at $\boldsymbol{\theta}_\gamma^{(H)}$ with respect to $\boldsymbol{\xi}_a, \boldsymbol{\eta}_b$ is given by

$$\begin{pmatrix} O & \tilde{F} \\ \tilde{F}^T & O \end{pmatrix}$$

where $\tilde{F} = (A\Lambda A^T) \otimes F$. Under the assumption that $F \neq O$, the eigenvalues of the above Hessian are $\{\delta_i, -\delta_i\}_{i=1}^r$, where $\{\delta_i\}_{i=1}^r$ is the singular values of \tilde{F} . This means there are increasing directions and decreasing directions of L_H around $\boldsymbol{\theta}_\gamma^{(H)}$, and thus it is a saddle point.

G PAC-Bayesian bound of generalization

G.1 Brief summary of general PAC-Bayes bound

The PAC-Bayesian framework [6, 7] has been developed for bounding generalization performance of learning models. It has been recently applied also to analysis of generalization of neural networks [8]. The following form of the bound is taken from [6].

Let $f(\mathbf{x}; \boldsymbol{\theta})$ be a real-valued function of \mathbf{x} with parameter $\boldsymbol{\theta} \in \Theta$. We consider the case that the loss function $\ell(\mathbf{y}; \mathbf{z})$ is bounded, and without loss of generality assume $\ell(\mathbf{y}, \mathbf{z}) \in [0, 1]$. Training data $(\mathbf{x}_1, \mathbf{y}_1), \dots, (\mathbf{x}_n, \mathbf{y}_n)$ is an i.i.d. sample from a distribution \mathcal{D} on (\mathbf{x}, \mathbf{y}) . Given function $f(\mathbf{x}; \boldsymbol{\theta})$, the training error (or empirical risk) is evaluated by

$$\hat{L}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{\nu=1}^n \ell(f(\mathbf{x}_\nu, \boldsymbol{\theta}), \mathbf{y}_\nu)$$

and the generalization error (or risk) is defined by

$$L(\boldsymbol{\theta}) = E_{\mathcal{D}}[\ell(\mathbf{f}(\mathbf{x}_\nu, \boldsymbol{\theta}), \mathbf{y}_\nu)].$$

In PAC-Bayes bound, we introduce a "prior" distribution P on the parameter space with an assumption that P does not depend on the training sample, and an arbitrary probability distribution Q on Θ . The distribution Q may depend on the training sample. Then, for any $\delta > 0$, the inequality

$$E_Q[L(\boldsymbol{\theta})] \leq E_Q[\hat{L}(\boldsymbol{\theta})] + 2\sqrt{\frac{2(KL(Q||P) + \ln \frac{n}{\delta})}{n-1}} \quad (39)$$

holds for sufficiently large n with probability greater than $1 - \delta$.

First, we can see that, if the distribution of Q is concentrated on a parameter set that gives very close values to $L(\hat{\boldsymbol{\theta}})$ or $\hat{L}(\hat{\boldsymbol{\theta}})$ at a parameter $\hat{\boldsymbol{\theta}}$ obtained by learning, then we have

$$E_Q[L(\boldsymbol{\theta})] \approx L(\hat{\boldsymbol{\theta}}), \quad E_Q[\hat{L}(\boldsymbol{\theta})] \approx \hat{L}(\hat{\boldsymbol{\theta}}).$$

In such cases, Eq. (39) shows the behavior of generalization error by its upper bound involving the approximate training error and the complexity term, which is expressed by the KL-divergence.

G.2 Generalization error bounds of embedded networks

The difference of the semi-flatness between networks of the smooth and ReLU activation can be related to the different generalization abilities of these models through the PAC-Bayes bound Eq. (39).

G.2.1 Choice in general cases

First we consider the general problem of choosing P and Q appropriately when the minimum of $\hat{L}(\boldsymbol{\theta})$ is sharp (non-flat) and can be approximated locally by a quadratic function around $\hat{\boldsymbol{\theta}}$, which is a minimum of $\hat{L}(\boldsymbol{\theta}^{(H)})$. The prior P should be non-informative, and thus if $\Theta = \mathbb{R}^d$, a normal distribution $N(0, \sigma^2 I_d)$ with a large σ is a reasonable choice. To relate the PAC-Bayes bound Eq. (39) to the generalization error at $\hat{\boldsymbol{\theta}}$, the distribution Q (posterior) should distribute on parameters that do not change the empirical risk values so much from the values given by $\hat{\boldsymbol{\theta}}$. Under the assumption that $\hat{L}(\boldsymbol{\theta})$ is well approximated by a quadratic function, we set Q by a normal distribution $N(0, \tau^2 \mathcal{H}^{-1})$ where \mathcal{H} is the Hessian

$$\mathcal{H} := \nabla^2 \hat{L}(\hat{\boldsymbol{\theta}})$$

with a small value of τ . Using the variance-covariance matrices based on the inverse Hessian is confirmed as follows. Suppose we set Q by $N(\hat{\boldsymbol{\theta}}, \Sigma)$ with a general Σ such that $\Sigma \ll \sigma^2$. Then, the Taylor series approximation of $\hat{L}(\boldsymbol{\theta}^{(H)})$ gives

$$E_Q[\hat{L}(\boldsymbol{\theta}^{(H)})] \approx \hat{L}(\hat{\boldsymbol{\theta}}^{(H)}) + \frac{1}{2} \text{Tr}[\mathcal{H}\Sigma],$$

and thus the right hand side of Eq. (39) is approximated by

$$\hat{L}(\hat{\boldsymbol{\theta}}^{(H)}) + \frac{1}{2} \text{Tr}[\mathcal{H}\Sigma] + 2\sqrt{\frac{2(KL(Q||P) + \ln \frac{n}{\delta})}{n-1}}. \quad (40)$$

It is well known that $KL(Q||P)$ with P and Q normal distributions is given by

$$KL(Q||P) = \frac{1}{2} \left[\log \frac{|\sigma^2 I_d|}{|\Sigma|} + \text{Tr}[\sigma^{-2} \Sigma] + \frac{\|\hat{\boldsymbol{\theta}}\|^2}{\sigma^2} - d \right]$$

To minimize Eq. (40) with respect to Σ , the differentiation provides the stationary condition

$$\mathcal{H} + \lambda(-\Sigma^{-1} + \sigma^{-2} I_d) = O$$

with some positive constant λ . From the assumption $\sigma^2 \gg \Sigma$, by neglecting $\sigma^{-2} I_d$, an approximate solution is given by

$$\Sigma_{opt} \approx \tau^2 \mathcal{H}^{-1},$$

where $\tau > 0$ is a scalar. Plugging this to Eq. (40) provides

$$\hat{L}(\hat{\boldsymbol{\theta}}^{(H)}) + \frac{\tau^2}{2}d + 2\sqrt{\frac{2\{d \log \frac{\sigma^2}{\tau^2} + \log \det \mathcal{H} + \frac{\tau^2}{\sigma^2} \text{Tr}[\mathcal{H}^{-1}] + \frac{\|\hat{\boldsymbol{\theta}}\|^2}{\sigma^2} - d\} + 2 \ln \frac{n}{\delta}}{n-1}}.$$

The second term is linear to τ^2 , and the main factor in the third term is $(d \log \frac{\sigma^2}{\tau^2})^{1/2} n^{-1/2}$ when $\sigma \gg 1$ and $\tau \ll 1$.

G.2.2 The case of inactive units

We now discuss the embedding of the smooth and ReLU networks by inactive units when the training error achieves zero error. As discussed in Section 5.1, some of the parameters give flat-directions, which requires some modification of the arguments in Section G.2.1.

As notations, $\boldsymbol{\theta}_{sm}^{(H)} \in \mathbb{R}^{d_{sm}}$ and $\boldsymbol{\theta}_{rl}^{(H)} \in \mathbb{R}^{d_{rl}}$ are used for the parameters of networks with smooth and ReLU activation, respectively, and they are decomposed as $\boldsymbol{\theta}_{sm}^{(H)} = (\boldsymbol{\theta}_{sm,0}^{(H)}, \boldsymbol{\theta}_{sm,1}^{(H)}, \boldsymbol{\theta}_{sm,2}^{(H)})$ and $\boldsymbol{\theta}_{rl}^{(H)} = (\boldsymbol{\theta}_{rl,0}^{(H)}, \boldsymbol{\theta}_{rl,1}^{(H)}, \boldsymbol{\theta}_{rl,2}^{(H)})$, corresponding to the components of a copy of $\boldsymbol{\theta}^{(H_0)}$, $(\mathbf{v}_j)_{j=H_0+1}^H$, and $(\mathbf{w}_j)_{j=H_0+1}^H$. Note that the both models have the same number of surplus parameters, i.e. $\dim(\boldsymbol{\theta}_{sm,1}^{(H)}) = \dim(\boldsymbol{\theta}_{sm,2}^{(H)}) =: d_1$ and $\dim(\boldsymbol{\theta}_{rl,1}^{(H)}) = \dim(\boldsymbol{\theta}_{rl,2}^{(H)}) =: d_2$. Different choices of P and Q are employed in the smooth and ReLU networks: we use P_{sm}, Q_{sm} for the smooth networks and P_{rl}, Q_{rl} for the ReLU case.

For the smooth activation, as in Section G.2.1, a non-informative prior

$$P_{sm} : N(0, \sigma^2 I)$$

is used with $\sigma \gg 1$. For the distribution Q_{sm} , we reflect the Hessian at the embedding by inactive units. By the definition, the directions of $(\mathbf{v}_j)_{j=H_0+1}^H$ give flat surface to L_H . The Hessian with respect to $(\mathbf{v}_j, \mathbf{w}_j)_{j=H_0+1}^H$ is thus given in the form

$$\begin{pmatrix} O & O \\ O & S \end{pmatrix},$$

where S is an $(H - H_0) \times D$ dimensional symmetric matrix given by

$$S_{jk} = \sum_{\nu=1}^n \mathbf{v}_j^T \frac{\partial^2 \ell_\nu(\hat{\boldsymbol{\theta}})}{\partial \mathbf{z} \partial \mathbf{z}} \mathbf{v}_k \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}^{(0)})}{\partial \mathbf{w}_j} \frac{\partial \varphi(\mathbf{x}_\nu; \mathbf{w}^{(0)})}{\partial \mathbf{w}_k} + \delta_{jk} \sum_{\nu=1}^n \frac{\partial \ell_\nu(\hat{\boldsymbol{\theta}})}{\partial \mathbf{z}} \mathbf{v}_j \frac{\partial^2 \varphi(\mathbf{x}_\nu; \mathbf{w}^{(0)})}{\partial \mathbf{w}_j \partial \mathbf{w}_k}.$$

For the flat directions of $(\mathbf{v}_j)_{j=H_0+1}^H$, the same distribution as P is optimal for the upper bound. Reflecting this, we set

$$Q_{sm} : N(\hat{\boldsymbol{\theta}}_{sm,0}^{(H)}, \tau^2 \mathcal{H}_{sm}^{-1}) \times N(\hat{\boldsymbol{\theta}}_{sm,1}^{(H)}, \sigma^2 I_{d^1}) \times N(\hat{\boldsymbol{\theta}}_{sm,2}^{(H)}, \tau^2 S^{-1}),$$

where $\hat{\boldsymbol{\theta}}_{sm}^{(H)}$ is the embedded point and $\mathcal{H}_{sm} := \nabla^2 L_{H_0}(\boldsymbol{\theta}_{*,sm}^{(H_0)})$ is the Hessian of the narrower network.

For the ReLU networks, we first fix $K > 1$ as a constant. Since in the direction of $(\mathbf{w}_j)_{j=H_0+1}^H$ we can presume the existence of the bonded flat subset $B_K^{H-H_0}$, we define the prior P_{rl} by

$$P_{rl} : N(0, \sigma^2 I_{d^0}) \times N(0, \sigma^2 I_{d^1}) \times \text{Unif}_{B_K^{H-H_0}}.$$

Reflecting the flat directions, the posterior Q_{rl} is defined by

$$Q_{rl} : N(\hat{\boldsymbol{\theta}}_{rl,0}^{(H)}, \tau^2 \mathcal{H}_{rl}^{-1}) \times N(\hat{\boldsymbol{\theta}}_{rl,1}^{(H)}, \sigma^2 I_{d^1}) \times \text{Unif}_{B_K^{H-H_0}},$$

where $\mathcal{H}_{rl} := \nabla^2 L_{H_0}(\boldsymbol{\theta}_{*,rl}^{(H_0)})$ is the Hessian of the narrower network.

With these choices, the KL divergence of the smooth case is given by

$$\begin{aligned} KL(Q_{sm} || P_{sm}) &= \frac{1}{2} \left[d_{sm}^0 \log \frac{\sigma^2}{\tau^2} + d^1 \log \frac{\sigma^2}{\tau^2} + \log \det \mathcal{H}_{sm} + \log \det S \right. \\ &\quad \left. + \text{Tr} \left[\frac{\tau^2}{\sigma^2} (\mathcal{H}_{sm}^{-1} + S^{-1}) \right] + \frac{\|\hat{\boldsymbol{\theta}}_{sm}\|^2}{\sigma^2} - d_{sm}^0 + d^1 \right], \end{aligned}$$

while in the case of ReLU networks,

$$KL(Q_{rl}||P_{rl}) = \frac{1}{2} \left[d_{rl}^0 \log \frac{\sigma^2}{\tau^2} + \log \det \mathcal{H}_{rl} + \text{Tr} \left[\frac{\tau^2}{\sigma^2} \mathcal{H}_{rl}^{-1} \right] + \frac{\|\hat{\boldsymbol{\theta}}_{rl}\|^2}{\sigma^2} - d_{rl}^0 \right].$$

With $\sigma \gg 1$ and $\tau \ll 1$, the major difference between these divergences comes from the term

$$d^1 \log \frac{\sigma^2}{\tau^2}$$

in the smooth networks. This suggests the advantage of the ReLU network in the overparameterized realization of zero training error in terms of the PAC-Bayesian upper bound of generalization error.

G.2.3 The Hessian for the zero error cases

We summarize the Hessian matrix for the embedding of a global minimum that attains zero training error. For simplicity, we write only the four blocks corresponding to the surplus units.

Smooth activation

(I) Unit replication: As discussed in Sections 3.2 and 5.1, the the part of the Hessian is given by

$$\begin{pmatrix} O & O \\ O & \tilde{G} \end{pmatrix}. \quad (41)$$

(II) Inactive units: The part of the Hessian is given by

$$\begin{pmatrix} O & O \\ O & S_1 \end{pmatrix}, \quad (42)$$

where

$$(S_1)_{jk} = \frac{\partial^2 L_H(\hat{\boldsymbol{\theta}})}{\partial \mathbf{w}_j \partial \mathbf{w}_k} = \sum_{\nu=1}^n \mathbf{v}_j^T \frac{\partial^2 \ell_{\nu}(\hat{\boldsymbol{\theta}})}{\partial \mathbf{z}_{\nu} \partial \mathbf{z}_{\nu}} \mathbf{v}_k \frac{\partial \varphi(\mathbf{x}_{\nu}; \mathbf{w}^{(0)})}{\partial \mathbf{w}} \frac{\partial \varphi(\mathbf{x}_{\nu}; \mathbf{w}^{(0)})^T}{\partial \mathbf{w}} \\ + \delta_{jk} \sum_{\nu=1}^n \frac{\partial \ell_{\nu}(\hat{\boldsymbol{\theta}})}{\partial \mathbf{z}_{\nu}} \mathbf{v}_j \frac{\partial^2 \varphi(\mathbf{x}_{\nu}; \mathbf{w}^{(0)})}{\partial \mathbf{w} \partial \mathbf{w}}.$$

(III) Inactive propagations: The part of the Hessian is given by

$$\begin{pmatrix} S_2 & O \\ O & O \end{pmatrix}, \quad (43)$$

where

$$(S_2)_{jk} = \frac{\partial^2 L_H(\hat{\boldsymbol{\theta}})}{\partial \mathbf{v}_j \partial \mathbf{v}_k} = \sum_{\nu=1}^n \frac{\partial^2 \ell_{\nu}(\hat{\boldsymbol{\theta}})}{\partial \mathbf{z}_{\nu} \partial \mathbf{z}_{\nu}} \varphi(\mathbf{x}_{\nu}; \mathbf{w}_j) \varphi(\mathbf{x}_{\nu}; \mathbf{w}_k).$$

We see that in all of the three cases the part of the Hessian for the surplus parameters contains a non-zero block.

ReLU

(I)_R Unit replication: As discussed in Sections 4.2, the the part of the Hessian is given by $\begin{pmatrix} O & \tilde{F} \\ \tilde{F}^T & O \end{pmatrix}$. Since the embedded point must not be a saddle, we have $\tilde{F} = O$. As a result, the part of the Hessian is constant zero.

(II)_R Inactive units: As discussed in Section 5.1, the part of the Hessian is zero.

(III)_R Inactive propagations: In this case, the part of the Hessian is given by

$$\begin{pmatrix} S_3 & O \\ O & O \end{pmatrix}, \quad (44)$$

where

$$(S_2)_{jk} = \frac{\partial^2 L_H(\hat{\boldsymbol{\theta}})}{\partial \mathbf{v}_j \partial \mathbf{v}_k} = \sum_{\nu=1}^n \frac{\partial^2 \ell_{\nu}(\hat{\boldsymbol{\theta}})}{\partial \mathbf{z}_{\nu} \partial \mathbf{z}_{\nu}} \varphi(\mathbf{x}_{\nu}; \mathbf{w}_j) \varphi(\mathbf{x}_{\nu}; \mathbf{w}_k)$$

which is not necessarily zero unless $\varphi(\mathbf{x}_{\nu}; \mathbf{w}_j) = 0$ for all ν .

We can see that the embedding by inactive units and unit replication give zero matrix for the part of Hessian, while the inactive propagation does not necessarily has zero matrix.

References

- [1] P. Chaudhari, A. Choromanska, S. Soatto, Y. LeCun, C. Baldassi, C. Borgs, J. T. Chayes, L. Sagun, and R. Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *CoRR*, abs/1611.01838, 2017.
- [2] K. Fukumizu and S. Amari. Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks*, 13(3):317–327, 2000.
- [3] S. Hochreiter and J. Schmidhuber. Simplifying neural nets by discovering flat minima. In *Advances in Neural Information Processing Systems 7*, pages 529–536. MIT Press, 1995.
- [4] S. Hochreiter and J. Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997. doi: 10.1162/neco.1997.9.1.1.
- [5] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836, 2017.
- [6] D. McAllester. Simplified PAC-Bayesian margin bounds. In *Learning Theory and Kernel Machines. Lecture Notes in Computer Science*, volume 2777, pages 203–215, 2003.
- [7] D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, Dec 1999.
- [8] B. Neyshabur, S. Bhojanapalli, and N. Srebro. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.
- [9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP Research Group, editors, *Parallel distributed processing*, volume 1, pages 318–362. MIT Press, Cambridge, 1986.