Unique Properties of Flat Minima in Deep Networks: Supplementary Material

Rotem Mulayoff [1] **Tomer Michaeli** [1]

This document contains supplementary material for the article 'Unique Properties of Flat Minima in Deep Networks', and includes the following parts:

I. Stability of Minima

It is well know that for a β-smooth function that is bounded from below, GD with constant step size $0 < \eta < 2/\beta$ converges to a stationary point. A twice continuously differentiable function f is β-smooth *if and only if* $\lambda_{\max}(\nabla^2 f(\boldsymbol{w})) \leq \beta$ for every point $\boldsymbol{w} \in \mathbb{R}^N$. Hence, convergence to a stationary point is guaranteed if $\lambda_{\max}(\nabla^2 f(\boldsymbol{w})) \leq 2/\eta$ for all $\boldsymbol{w} \in \mathbb{R}^N$. This seemingly stringent global requirement can in fact also be replaced by a local one, as shown by Wu et al. (2018). Specifically, they use the following.

Definition S1. *Let $\boldsymbol{w}^*$ be a stationary point of f. Consider the linearized dynamical system of GD, namely*

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta \nabla^2 f(\boldsymbol{w}^*)(\boldsymbol{w}_k - \boldsymbol{w}^*). \tag{S1}$$

Then $\boldsymbol{w}^*$ is said to be linearly stable *if there exists a constant $C \in \mathbb{R}$, such that $\|\boldsymbol{w}_k\| \leq C \|\boldsymbol{w}_0\|$ for all $k > 0$.*

In other words, $\boldsymbol{w}^*$ is linearly stable if once we have arrived near this critical point, we stay around it. In their paper, Wu et al. (2018) show that $\boldsymbol{w}^*$ is a linearly stable minimizer if

$$\lambda_{\max}\left(\left(\boldsymbol{I} - \eta \nabla^2 f(\boldsymbol{w}^*)\right)^2\right) \leq 1. \tag{S2}$$

Note that for all $i \in \{1, \ldots, N\}$

$$\begin{aligned}\lambda_i\left(\left(\boldsymbol{I} - \eta \nabla^2 f(\boldsymbol{w}^*)\right)^2\right) &= \lambda_i^2\left(\boldsymbol{I} - \eta \nabla^2 f(\boldsymbol{w}^*)\right) \\ &= \left(1 - \eta \lambda_i\left(\nabla^2 f(\boldsymbol{w}^*)\right)\right)^2 \\ &= 1 - \eta \lambda_i\left(\nabla^2 f(\boldsymbol{w}^*)\right)\left(2 - \eta \lambda_i\left(\nabla^2 f(\boldsymbol{w}^*)\right)\right),\end{aligned} \tag{S3}$$

[1]Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, Israel. Correspondence to: Rotem Mulayoff <rotem.mulayof@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

where λ_i is the ith largest eigenvalue. Since η and $\{\lambda_i\}$ are all nonnegative, it follows that (S2) is equivalent to

$$\lambda_{\max}\left(\nabla^2 f(\boldsymbol{w}^*)\right) \leq \frac{2}{\eta}. \tag{S4}$$

This results asserts that flat minima are stable solutions for GD. In their paper, they also provide a similar result for stochastic GD (SGD), which shows that the sharpness of a minimum should increase to ensure stability for SGD as well.

II. Proof of Lemma 1

Nar & Sastry (2018) showed that under the Lemma's conditions, the weight matrices converge at a linear rate to $\boldsymbol{W}_i = \boldsymbol{T}^{\frac{1}{m}}$ for all i. According to our Theorem 4, this solution is a flattest minimum, thus demonstrating (ii).

Arora et al. (2018) showed that gradient flow (GF) satisfies

$$\boldsymbol{W}_i(t)\boldsymbol{W}_i^T(t) = \boldsymbol{W}_{i+1}^T(t)\boldsymbol{W}_{i+1}(t), \quad \forall t \geq 0 \tag{S5}$$

in our setting. Denoting the SVD of $\boldsymbol{W}_i(t)$ by $\boldsymbol{U}_i(t)\boldsymbol{S}_i(t)\boldsymbol{V}_i^T(t)$, we thus have that $\boldsymbol{U}_i(t)\boldsymbol{S}_i^2(t)\boldsymbol{U}_i^T(t) = \boldsymbol{V}_{i+1}(t)\boldsymbol{S}_{i+1}^2(t)\boldsymbol{V}_{i+1}^T(t)$, which implies that[1]

$$\boldsymbol{U}_i(t) = \boldsymbol{V}_{i+1}(t), \quad \boldsymbol{S}_i(t) = \boldsymbol{S}_{i+1}(t), \quad \forall t \geq 0. \tag{S6}$$

Assume that GF converges to a global minimimum and let $\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ denote the SVD of $\boldsymbol{T}$. Since $\{\boldsymbol{S}_i(t)\}_{i=1}^m$ are identical, they converge to the same limit, $\bar{\boldsymbol{S}}$. Let $\boldsymbol{W}_i = \boldsymbol{U}_i\bar{\boldsymbol{S}}\boldsymbol{V}_i^T$ denote the limit of $\boldsymbol{W}_i(t)$. Then, from (S6), we have that $\boldsymbol{V}_{i+1}^T\boldsymbol{U}_i = \boldsymbol{I}$ for all i. Consequently,

$$\boldsymbol{W}_m\boldsymbol{W}_{m-1}\cdots\boldsymbol{W}_1 = \boldsymbol{U}_m\bar{\boldsymbol{S}}^m\boldsymbol{V}_1^T. \tag{S7}$$

But since the left hand side equals $\boldsymbol{T}$ by assumption, the right hand side must coincide with the SVD of $\boldsymbol{T}$. This means that $\bar{\boldsymbol{S}} = \boldsymbol{S}^{\frac{1}{m}}$. Again, by Theorem 4, this is a flattest minimum, thus demonstrating (i).

III. Scalar Networks

III.1. The Set of Flattest Minima

As mentioned in the main text, in the scalar case, the end-to-end function $f_{\boldsymbol{w}}(x)$ implemented by the network is given by

$$f_{\boldsymbol{w}}(x) = \prod_{j=1}^{m} w_j x, \tag{S8}$$

where $\boldsymbol{w} = [w_1, w_2, \ldots, w_m]^T$. In our analysis we consider a quadratic loss function, *i.e.*

$$\ell(\boldsymbol{w}) = \hat{\mathbb{E}}\left[\left(y - f_{\boldsymbol{w}}(x)\right)^2\right]. \tag{S9}$$

Our goal is to characterize the set of flattest minima of the loss w.r.t. $\boldsymbol{w}$. It is well known that the optimal coefficient for linear estimation is given by

$$\tau = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}, \tag{S10}$$

where $\hat{\sigma}_x^2 = \mathrm{E}[x^2]$ is the empirical second-order moment of x, and $\hat{\sigma}_{xy} = \mathrm{E}[xy]$ is the empirical cross second-order moment between x and y. Therefore, at any global minimum of $\ell(\boldsymbol{w})$, we have

$$\prod_{j=1}^{m} w_j = \tau. \tag{S11}$$

[1]The SVD can be non-unique, however there necessarily exists a decomposition satisfying $\boldsymbol{U}_i(t) = \boldsymbol{V}_{i+1}(t)$.

To compute the Hessian matrix of $\ell(\boldsymbol{w})$, we first calculate the partial derivative w.r.t. w_k,

$$\begin{aligned}\frac{\partial}{\partial w_k}\ell(\boldsymbol{w}) &= \frac{\partial}{\partial w_k}\hat{\mathbb{E}}\left[\left(y - f_{\boldsymbol{w}}(x)\right)^2\right] = -2\hat{\mathbb{E}}\left[\left(y - f_{\boldsymbol{w}}(x)\right)\frac{\partial}{\partial w_k}f_{\boldsymbol{w}}(x)\right] \\ &= -2\hat{\mathbb{E}}\left[\left(y - \prod_{j=1}^{m} w_j x\right)\prod_{j\neq k} w_j x\right] = 2\Big(\hat{\sigma}_x^2\prod_{j=1}^{m} w_j - \hat{\sigma}_{xy}\Big)\prod_{j\neq k} w_j \end{aligned} \tag{S12}$$

We now complete the derivation by differentiating (S12) w.r.t. w_q,

$$\begin{aligned}\frac{\partial^2}{\partial w_q \partial w_k}\ell(\boldsymbol{w}) &= \frac{\partial}{\partial w_q}\left[2\Big(\hat{\sigma}_x^2\prod_{j=1}^{m} w_j - \hat{\sigma}_{xy}\Big)\prod_{j\neq k} w_j\right] \\ &= 2\prod_{j\neq k} w_j \frac{\partial}{\partial w_q}\Big(\hat{\sigma}_x^2\prod_{j=1}^{m} w_j - \hat{\sigma}_{xy}\Big) + 2\Big(\hat{\sigma}_x^2\prod_{j=1}^{m} w_j - \hat{\sigma}_{xy}\Big)\frac{\partial}{\partial w_q}\prod_{j\neq k} w_j. \end{aligned} \tag{S13}$$

Eq. (S11) asserts that $\hat{\sigma}_x^2\prod_{j=1}^{m} w_j - \hat{\sigma}_{xy} = 0$ at global minima, therefore the second term in (S13) vanishes, and we obtain

$$\frac{\partial^2}{\partial w_q \partial w_k}L(\boldsymbol{w}) = 2\prod_{j\neq k} w_j \frac{\partial}{\partial w_q}\Big(\hat{\sigma}_x^2\prod_{j=1}^{m} w_j - \hat{\sigma}_{xy}\Big) = 2\hat{\sigma}_x^2\Big(\prod_{j\neq k} w_j\Big)\Big(\prod_{j\neq q} w_j\Big). \tag{S14}$$

Hence, using (S10), we can express the elements of the Hessian matrix $\boldsymbol{H}_{\boldsymbol{w}}$ as

$$\big(\boldsymbol{H}_{\boldsymbol{w}}\big)_{k,q} = 2\hat{\sigma}_x^2\tau^2\frac{1}{w_k w_q}. \tag{S15}$$

Let us define the vector $\boldsymbol{z} = [w_1^{-1}, w_2^{-1}, \ldots, w_m^{-1}]^T$, then the Hessian matrix can be equivalently written as

$$\boldsymbol{H}_{\boldsymbol{w}} = 2\hat{\sigma}_x^2\tau^2\boldsymbol{z}\boldsymbol{z}^T. \tag{S16}$$

This shows that the Hessian is a rank one matrix, which implies that it has only one nonzero eigenvalue, with a corresponding eigenvector $\boldsymbol{z}$. Therefore,

$$\lambda_{\max}\big(\boldsymbol{H}_{\boldsymbol{w}}\big)\boldsymbol{z} = \boldsymbol{H}_{\boldsymbol{w}}\boldsymbol{z} = 2\hat{\sigma}_x^2\big(\tau\big)^2\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{z} = 2\hat{\sigma}_x^2\big(\tau\big)^2\left\|\boldsymbol{z}\right\|^2\boldsymbol{z}, \tag{S17}$$

so that the eigenvalue is given by

$$\lambda_{\max}\big(\boldsymbol{H}_{\boldsymbol{w}}\big) = 2\hat{\sigma}_x^2\tau^2\|\boldsymbol{z}\|^2 = 2\hat{\sigma}_x^2\tau^2\sum_{j=1}^{m}\frac{1}{w_j^2}. \tag{S18}$$

To determine the sharpness of the flattest minima, we need to solve the problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^m}\ \lambda_{\max}\big(\boldsymbol{H}_{\boldsymbol{w}}\big) \qquad \text{s.t.} \qquad \prod_{j=1}^{m} w_j = \tau. \tag{S19}$$

By the inequality of the arithmetic and geometric means, we have that for any feasible point $\boldsymbol{w}$

$$\sum_{j=1}^{m}\frac{1}{w_j^2} \geq m \times \left(\prod_{j=1}^{m}\frac{1}{w_j^2}\right)^{\frac{1}{m}} = m \times \tau^{-\frac{2}{m}}. \tag{S20}$$

Therefore, for all feasible points,

$$\lambda_{\max}\big(\boldsymbol{H}_{\boldsymbol{w}}\big) \geq 2m\hat{\sigma}_x^2\tau^{2(1-\frac{1}{m})}. \tag{S21}$$

On the other hand, this inequality can be achieved by setting $|w_1| = |w_2| = \cdots = |w_m|$. This shows that the right-hand-side is precisely the sharpness of the flattest minimum, so that

$$\Omega_0 = \Big\{\boldsymbol{w}\in\mathbb{R}^m\ :\ \prod_{j=1}^{m}\operatorname{sgn}(w_j) = \operatorname{sgn}\big(\tau\big) \quad \text{and} \quad |w_j| = \sqrt[m]{|\tau|} \quad \forall j\Big\}. \tag{S22}$$

III.2. Proof of Lemma 2

In this section we examine the behavior of the loss function on a line connecting two minima.

Claim S1. *Assume that $\tau > 0$ and let $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$ be minimizers of the objective (S9) in $\mathbb{R}_+^m$. Then, along the line connecting $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$, the loss function will appear sharper around $\boldsymbol{w}^{(1)}$ than around $\boldsymbol{w}^{(2)}$ if*

$$\sum_{i=1}^{m} \frac{w_i^{(2)}}{w_i^{(1)}} > \sum_{i=1}^{m} \frac{w_i^{(1)}}{w_i^{(2)}}. \tag{S23}$$

Proof. The direction vector of the connecting line is $\boldsymbol{\alpha} = (\boldsymbol{w}^{(1)} - \boldsymbol{w}^{(2)})/\|\boldsymbol{w}^{(1)} - \boldsymbol{w}^{(2)}\|$. Along this direction, the behavior of the loss function around $\boldsymbol{w}^{(i)}$ is given by

$$\ell(\boldsymbol{w}^{(i)} + \eta\boldsymbol{\alpha}) \approx \ell(\boldsymbol{w}^{(i)}) + \eta\boldsymbol{\alpha}^T \nabla\ell(\boldsymbol{w}^{(i)}) + \frac{\eta^2}{2}\boldsymbol{\alpha}^T \boldsymbol{H}(\boldsymbol{w}^{(i)})\boldsymbol{\alpha}. \tag{S24}$$

Since $\nabla\ell(\boldsymbol{w}^{(1)}) = \nabla\ell(\boldsymbol{w}^{(2)}) = \boldsymbol{0}$ and $\ell(\boldsymbol{w}^{(1)}) = \ell(\boldsymbol{w}^{(2)})$, the loss function will appear sharper around $\boldsymbol{w}^{(1)}$ than around $\boldsymbol{w}^{(2)}$, if $\boldsymbol{\alpha}^T \boldsymbol{H}(\boldsymbol{w}^{(1)})\boldsymbol{\alpha} > \boldsymbol{\alpha}^T \boldsymbol{H}(\boldsymbol{w}^{(2)})\boldsymbol{\alpha}$. From (S16), this condition is equivalent to $\|\boldsymbol{\alpha}^T \boldsymbol{z}^{(1)}\|^2 > \|\boldsymbol{\alpha}^T \boldsymbol{z}^{(2)}\|^2$, or more explicitly,

$$\left(\left(\boldsymbol{w}^{(1)} - \boldsymbol{w}^{(2)}\right)^T \boldsymbol{z}^{(1)}\right)^2 > \left(\left(\boldsymbol{w}^{(1)} - \boldsymbol{w}^{(2)}\right)^T \boldsymbol{z}^{(2)}\right)^2. \tag{S25}$$

Since $\boldsymbol{z} = [w_1^{-1}, w_2^{-1}, \ldots, w_m^{-1}]^T$, this inequality can be written as

$$\left|\sum_{i=1}^{m} \frac{w_i^{(2)}}{w_i^{(1)}} - m\right| > \left|\sum_{i=1}^{m} \frac{w_i^{(1)}}{w_i^{(2)}} - m\right|. \tag{S26}$$

Note that

$$\sum_{i=1}^{m} \frac{w_i^{(2)}}{w_i^{(1)}} \geq m \sqrt[m]{\prod_{i=1}^{m} \frac{w_i^{(2)}}{w_i^{(1)}}} = m \sqrt[m]{\frac{\prod_{i=1}^{m} w_i^{(2)}}{\prod_{j=1}^{m} w_j^{(1)}}} = m \sqrt[m]{\frac{\tau}{\tau}} = m. \tag{S27}$$

Similarly, $\sum_{i=1}^{m} \frac{w_i^{(1)}}{w_i^{(2)}} \geq m$. Therefore, (S26) can be reduced to

$$\sum_{i=1}^{m} \frac{w_i^{(2)}}{w_i^{(1)}} > \sum_{i=1}^{m} \frac{w_i^{(1)}}{w_i^{(2)}}. \tag{S28}$$

■

Notice that the loss function is symmetric in a sense that if we flip the sign of two scalar layers, then it remains the same. Therefore, without loss of generality, we can restrict our analysis to a single orthant. Let $\tau > 0$, and $\boldsymbol{w}^{(1)}$ be the flattest minimum in $\mathbb{R}_+^m$, *i.e.* $w_i^{(1)} = \tau^{1/m}$ for all $i \in \{1, \ldots, m\}$. Given a second minimum $\boldsymbol{w}^{(2)} \in \mathbb{R}_+^m$ for which the connecting line between $\boldsymbol{w}^{(1)}$ and $\boldsymbol{w}^{(2)}$ is loyal to the true sharpness, we can construct a third solution $\boldsymbol{w}^{(3)}$ that will appear deceivingly flatter than $\boldsymbol{w}^{(1)}$ over their connecting line. Specifically, let us set

$$w_i^{(3)} = \frac{\left(w_i^{(1)}\right)^2}{w_i^{(2)}}. \tag{S29}$$

Clearly, $\boldsymbol{w}^{(3)}$ is a global minimum as $\prod_{i=1}^{m} w_i^{(3)} = \tau$. Since $\boldsymbol{w}^{(2)}$ appears sharper than $\boldsymbol{w}^{(1)}$ along their connecting line, then according to Claim S1 we have

$$\sum_{i=1}^{m} \frac{w_i^{(1)}}{w_i^{(2)}} > \sum_{i=1}^{m} \frac{w_i^{(2)}}{w_i^{(1)}}. \tag{S30}$$

Thus,

$$\sum_{i=1}^{m}\frac{w_i^{(3)}}{w_i^{(1)}}=\sum_{i=1}^{m}\frac{w_i^{(1)}}{w_i^{(2)}}>\sum_{i=1}^{m}\frac{w_i^{(2)}}{w_i^{(1)}}=\sum_{i=1}^{m}\frac{w_i^{(1)}}{w_i^{(3)}}. \tag{S31}$$

Therefore, by Claim S1, $\boldsymbol{w}^{(1)}$ appears sharper than $\boldsymbol{w}^{(3)}$ along their connecting line.

In the special case of two layer networks ($m=2$), we have that for any minimizer, $w_2^{(i)}=\tau/w_1^{(i)}$. Hence,

$$\frac{w_1^{(1)}}{w_1^{(2)}}+\frac{w_2^{(1)}}{w_2^{(2)}}=\frac{w_2^{(2)}}{w_2^{(1)}}+\frac{w_1^{(2)}}{w_1^{(1)}}. \tag{S32}$$

This means that the minima will appear equally sharp.

IV. Proof of Lemma 3

In this section we derive the Hessian matrix defined in (18) at a global minimum point, *i.e.* for $\boldsymbol{w}\in\Omega$. Throughout this section we will be using the following properties of the Kronecker product. For any matrices $\boldsymbol{M}_1,\boldsymbol{M}_2,\boldsymbol{M}_3,\boldsymbol{M}_4$,

$$\mathrm{vec}\left(\boldsymbol{M}_1\boldsymbol{M}_2\boldsymbol{M}_3\right)=\left(\boldsymbol{M}_3^T\otimes\boldsymbol{M}_1\right)\mathrm{vec}\left(\boldsymbol{M}_2\right), \tag{P1}$$

$$\left(\boldsymbol{M}_1\otimes\boldsymbol{M}_2\right)^T=\left(\boldsymbol{M}_1^T\otimes\boldsymbol{M}_2^T\right), \tag{P2}$$

$$\left(\boldsymbol{M}_1\otimes\boldsymbol{M}_2\right)\left(\boldsymbol{M}_3\otimes\boldsymbol{M}_4\right)=\left(\boldsymbol{M}_1\boldsymbol{M}_3\right)\otimes\left(\boldsymbol{M}_2\boldsymbol{M}_4\right). \tag{P3}$$

Let us start the computation of $\boldsymbol{H}_{\boldsymbol{w}}$ by rearranging the loss function so as to simplify the differentiation w.r.t. $\boldsymbol{w}_k$. Specifically, we have that

$$\begin{aligned}
\ell(\boldsymbol{w})&=\hat{\mathbb{E}}\left[\left\|y-\prod_{j=1}^{m}\boldsymbol{W}_j x\right\|^2\right]\\
&=\hat{\mathbb{E}}\left[\left\|y-\Big(\prod_{i=k+1}^{m}\boldsymbol{W}_i\Big)\boldsymbol{W}_k\Big(\prod_{j=1}^{k-1}\boldsymbol{W}_j x\Big)\right\|^2\right]\\
&=\hat{\mathbb{E}}\left[\left\|y-\Big(\prod_{j=1}^{k-1}\boldsymbol{W}_j x\Big)^T\otimes\Big(\prod_{i=k+1}^{m}\boldsymbol{W}_i\Big)\boldsymbol{w}_k\right\|^2\right]\\
&=\hat{\mathbb{E}}\left[\left\|y-\Big[x^T\Big(\prod_{j=1}^{k-1}\boldsymbol{W}_j\Big)^T\Big]\otimes\Big[\boldsymbol{I}\Big(\prod_{i=k+1}^{m}\boldsymbol{W}_i\Big)\Big]\boldsymbol{w}_k\right\|^2\right]\\
&=\hat{\mathbb{E}}\left[\left\|y-\Big[x\otimes\boldsymbol{I}\Big]^T\Big[\Big(\prod_{j=1}^{k-1}\boldsymbol{W}_j\Big)^T\otimes\Big(\prod_{i=k+1}^{m}\boldsymbol{W}_i\Big)\Big]\boldsymbol{w}_k\right\|^2\right],
\end{aligned} \tag{S33}$$

where in the third equality we used property (P1), and in the last we used properties (P2) and (P3). To simplify expressions, we define the following matrices

$$\boldsymbol{U}_k\triangleq\Big(\prod_{j=1}^{k-1}\boldsymbol{W}_j\Big)^T\otimes\Big(\prod_{i=k+1}^{m}\boldsymbol{W}_i\Big)\qquad\text{and}\qquad\boldsymbol{X}\triangleq x\otimes\boldsymbol{I}. \tag{S34}$$

Thus, the loss function (2) is given by

$$\ell(\boldsymbol{w})=\hat{\mathbb{E}}\left[\left\|y-\boldsymbol{X}^T\boldsymbol{U}_k\boldsymbol{w}_k\right\|^2\right]. \tag{S35}$$

Now we are ready to calculate the partial derivative of $\ell(\boldsymbol{w})$ w.r.t $\boldsymbol{w}_k$. Notice that $\boldsymbol{U}_k$ is not a function of $\boldsymbol{w}_k$, therefore

$$\frac{\partial}{\partial\boldsymbol{w}_k}\hat{\mathbb{E}}\left[\left\|y-\boldsymbol{X}^T\boldsymbol{U}_k\boldsymbol{w}_k\right\|^2\right]=-2\hat{\mathbb{E}}\left[\boldsymbol{U}_k^T\boldsymbol{X}\left(y-\boldsymbol{X}^T\boldsymbol{U}_k\boldsymbol{w}_k\right)\right]=2\boldsymbol{U}_k^T\left(\hat{\mathbb{E}}\left[\boldsymbol{X}\boldsymbol{X}^T\right]\boldsymbol{U}_k\boldsymbol{w}_k-\hat{\mathbb{E}}\left[\boldsymbol{X}y\right]\right). \tag{S36}$$

Furthermore,

$$\hat{\mathbb{E}}\left[\boldsymbol{X}\boldsymbol{X}^T\right] = \hat{\mathbb{E}}\left[(x\otimes\boldsymbol{I})(x\otimes\boldsymbol{I})^T\right] = \hat{\mathbb{E}}\left[(xx^T\otimes\boldsymbol{I})\right] = \left(\hat{\mathbb{E}}\left[xx^T\right]\right)\otimes\boldsymbol{I} = \hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}, \tag{S37}$$

where in the second equality we used properties (P2) and (P3), and in the third equality we used the linearity of the Kronecker product. Additionally,

$$\hat{\mathbb{E}}\left[\boldsymbol{X}y\right] = \hat{\mathbb{E}}\left[(x\otimes\boldsymbol{I})y\right] = \hat{\mathbb{E}}\left[\mathrm{vec}\left(yx^T\right)\right] = \mathrm{vec}\left(\hat{\boldsymbol{\Sigma}}_{yx}\right), \tag{S38}$$

where in the second step we used (P1). Overall we have that

$$\frac{\partial}{\partial\boldsymbol{w}_k}\ell(\boldsymbol{w}) = 2\boldsymbol{U}_k^T\left[\left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\boldsymbol{U}_k\boldsymbol{w}_k - \mathrm{vec}\left(\hat{\boldsymbol{\Sigma}}_{yx}\right)\right]. \tag{S39}$$

Next we prepare Eq. (S39) for differentiation w.r.t $\boldsymbol{w}_q$. First, for all k

$$\boldsymbol{U}_k\boldsymbol{w}_k = \left[\left(\prod_{j=1}^{k-1}\boldsymbol{W}_j\right)^T\otimes\left(\prod_{i=k+1}^{m}\boldsymbol{W}_i\right)\right]\mathrm{vec}\left(\boldsymbol{W}_k\right) = \mathrm{vec}\left(\left(\prod_{i=k+1}^{m}\boldsymbol{W}_i\right)\boldsymbol{W}_k\left(\prod_{j=1}^{k-1}\boldsymbol{W}_j\right)\right) = \mathrm{vec}\left(\prod_{i=1}^{m}\boldsymbol{W}_i\right). \tag{S40}$$

Particularly, this means that the value of the term $\boldsymbol{U}_k\boldsymbol{w}_k$ is the same for all k. Hence, $\boldsymbol{U}_k\boldsymbol{w}_k = \boldsymbol{U}_q\boldsymbol{w}_q$ and therefore

$$\frac{\partial}{\partial\boldsymbol{w}_k}\ell(\boldsymbol{w}) = 2\boldsymbol{U}_k^T\left[\left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\boldsymbol{U}_q\boldsymbol{w}_q - \mathrm{vec}\left(\hat{\boldsymbol{\Sigma}}_{yx}\right)\right]. \tag{S41}$$

Now, let us differentiate the vector $\frac{\partial}{\partial\boldsymbol{w}_k}\ell(\boldsymbol{w})$ w.r.t. the scalar $\boldsymbol{w}_{q,l}$, which is the lth element in the vector $\boldsymbol{w}_q$. Notice that $\boldsymbol{U}_k$ and $\boldsymbol{w}_q$ itself are the only terms which depend on $\boldsymbol{w}_q$. Therefore, by the product rule of differentiation using the denominator-layout notation[2]

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{w}_{q,l}\partial\boldsymbol{w}_k}\ell(\boldsymbol{w}) &= 2\frac{\partial}{\partial\boldsymbol{w}_{q,l}}\left(\boldsymbol{U}_k^T\left[\left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\boldsymbol{U}_q\boldsymbol{w}_q - \mathrm{vec}\left(\hat{\boldsymbol{\Sigma}}_{yx}\right)\right]\right)\\ &= 2\left[\left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\boldsymbol{U}_q\boldsymbol{w}_q - \mathrm{vec}\left(\hat{\boldsymbol{\Sigma}}_{yx}\right)\right]^T\left(\frac{\partial}{\partial\boldsymbol{w}_{q,l}}\boldsymbol{U}_k^T\right) + 2\left(\frac{\partial}{\partial\boldsymbol{w}_{q,l}}\boldsymbol{w}_q\right)\boldsymbol{U}_q^T\left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\boldsymbol{U}_k.\end{aligned} \tag{S42}$$

However, at a global minimum

$$\left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\boldsymbol{U}_q\boldsymbol{w}_q = \left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\mathrm{vec}\left(\prod_{i=1}^{m}\boldsymbol{W}_i\right) = \mathrm{vec}\left(\prod_{i=1}^{m}\boldsymbol{W}_i\hat{\boldsymbol{\Sigma}}_x\right) = \mathrm{vec}\left(\hat{\boldsymbol{\Sigma}}_{yx}\hat{\boldsymbol{\Sigma}}_x^{-1}\hat{\boldsymbol{\Sigma}}_x\right) = \mathrm{vec}\left(\hat{\boldsymbol{\Sigma}}_{yx}\right), \tag{S43}$$

where in the first equality we used (S40), in the second equality we used (P1) and in the third equality we used the assumption that $\boldsymbol{w}\in\Omega$. Hence, for all $1\le q\le m$ we have that

$$\left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\boldsymbol{U}_q\boldsymbol{w}_q - \mathrm{vec}\left(\hat{\boldsymbol{\Sigma}}_{yx}\right) = \boldsymbol{0}. \tag{S44}$$

Therefore, (S42) is reduced to

$$\frac{\partial^2}{\partial\boldsymbol{w}_{q,l}\partial\boldsymbol{w}_k}\ell(\boldsymbol{w}) = 2\left(\frac{\partial}{\partial\boldsymbol{w}_{q,l}}\boldsymbol{w}_q\right)\boldsymbol{U}_q^T\left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\boldsymbol{U}_k, \tag{S45}$$

for all $1\le l\le m$. Hence,

$$\frac{\partial^2}{\partial\boldsymbol{w}_q\partial\boldsymbol{w}_k}\ell(\boldsymbol{w}) = 2\left(\frac{\partial}{\partial\boldsymbol{w}_q}\boldsymbol{w}_q\right)\boldsymbol{U}_q^T\left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\boldsymbol{U}_k = 2\boldsymbol{U}_q^T\left(\hat{\boldsymbol{\Sigma}}_x\otimes\boldsymbol{I}\right)\boldsymbol{U}_k = 2\boldsymbol{U}_q^T\left(\hat{\boldsymbol{\Sigma}}_x^{\frac{1}{2}}\otimes\boldsymbol{I}\right)^T\left(\hat{\boldsymbol{\Sigma}}_x^{\frac{1}{2}}\otimes\boldsymbol{I}\right)\boldsymbol{U}_k, \tag{S46}$$

[2]Where the derivative $\frac{\partial\mathbf{A}}{\partial z}$ of a matrix $\mathbf{A}$ w.r.t. a scalar z is laid out according to $\mathbf{A}^T$.

where $\hat{\boldsymbol{\Sigma}}_x^{\frac{1}{2}}$ is the symmetric square root matrix of $\hat{\boldsymbol{\Sigma}}_x$. Let us define the matrices $\{\boldsymbol{\Phi}_k\}_{k=1}^m$ as

$$\boldsymbol{\Phi}_k = \boldsymbol{U}_k^T \left(\hat{\boldsymbol{\Sigma}}_x^{\frac{1}{2}} \otimes \boldsymbol{I}\right)^T = \left[\left(\prod_{j=1}^{k-1} \boldsymbol{W}_j\right) \otimes \left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T\right] \left(\hat{\boldsymbol{\Sigma}}_x^{\frac{1}{2}} \otimes \boldsymbol{I}\right) = \left(\prod_{j=1}^{k-1} \boldsymbol{W}_j \hat{\boldsymbol{\Sigma}}_x^{\frac{1}{2}}\right) \otimes \left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T. \tag{S47}$$

Thus,

$$\frac{\partial^2}{\partial \boldsymbol{w}_q \partial \boldsymbol{w}_k} \ell(\boldsymbol{w}) = 2\boldsymbol{\Phi}_q \boldsymbol{\Phi}_k^T. \tag{S48}$$

Finally, the Hessian matrix of the loss function $\ell(\boldsymbol{w})$ is given by

$$\boldsymbol{H}_{\boldsymbol{w}} = 2\boldsymbol{\Phi}\boldsymbol{\Phi}^T, \tag{S49}$$

where $\boldsymbol{\Phi} = [\boldsymbol{\Phi}_1^T, \boldsymbol{\Phi}_2^T, \ldots, \boldsymbol{\Phi}_m^T]^T$.

V. The Missing Parts of the Proof of Theorem 1

V.1. Proof of Lemma 4

The proof is straightforward. We have

$$\sum_{k=1}^{m} \|\boldsymbol{\Psi}_k\|_{\mathrm{F}}^2 \geq \sum_{k=1}^{m} \|\boldsymbol{\Psi}_k\|_2^2 \geq m \left[\prod_{k=1}^{m} \|\boldsymbol{\Psi}_k\|_2\right]^{\frac{2}{m}} \geq m \left[\left\|\prod_{k=1}^{m} \boldsymbol{\Psi}_k\right\|_2\right]^{\frac{2}{m}}, \tag{S50}$$

where in the first inequality we used the fact that $\|\boldsymbol{\Psi}\|_{\mathrm{F}} \geq \|\boldsymbol{\Psi}\|_2$ for any matrix $\boldsymbol{\Psi} \in \mathbb{R}^{d_1 \times d_2}$. The second inequality is due to the inequality of arithmetic and geometric means. In the final inequality we used the fact that $\|\cdot\|_2$ is a sub-multiplicative matrix norm, meaning $\|\boldsymbol{\Psi}\|_2 \|\boldsymbol{\Phi}\|_2 \geq \|\boldsymbol{\Psi}\boldsymbol{\Phi}\|_2$ for any pair of matrices $\boldsymbol{\Psi} \in \mathbb{R}^{d_1 \times d_2}, \boldsymbol{\Phi} \in \mathbb{R}^{d_2 \times d_3}$.

V.2. Maximal Value of ν

On the one hand, for any $\boldsymbol{B} \in \mathbb{R}^{d_y \times d_x}$ such that $\|\boldsymbol{B}\|_{\mathrm{F}} = 1$,

$$\left\|\left(\boldsymbol{B}\boldsymbol{T}^T\right)^{m-1} \boldsymbol{B}\right\|_2 \leq \|\boldsymbol{B}\|_2^m \|\boldsymbol{T}\|_2^{m-1} \leq (\sigma_{\max}(\boldsymbol{T}))^{m-1}, \tag{S51}$$

where in the second inequality we used $\|\boldsymbol{B}\|_2 \leq \|\boldsymbol{B}\|_{\mathrm{F}} = 1$. On the other hand, this upper bound is achieved by $\boldsymbol{B} = \boldsymbol{u}\boldsymbol{v}^T$, as

$$\begin{aligned} \left\|\left(\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{T}^T\right)^{m-1} \boldsymbol{u}\boldsymbol{v}^T\right\|_2 &= \left\|\boldsymbol{u}\left(\boldsymbol{v}^T\boldsymbol{T}^T\boldsymbol{u}\right)^{m-1} \boldsymbol{v}^T\right\|_2 = \left(\boldsymbol{v}^T\boldsymbol{T}^T\boldsymbol{u}\right)^{m-1} \left\|\boldsymbol{u}\boldsymbol{v}^T\right\|_2 \\ &= (\sigma_{\max}(\boldsymbol{T}))^{m-1} \|\boldsymbol{u}\| \|\boldsymbol{v}\| = (\sigma_{\max}(\boldsymbol{T}))^{m-1}. \end{aligned} \tag{S52}$$

Therefore,

$$\max_{\|\boldsymbol{B}\|_{\mathrm{F}}=1} \nu(\boldsymbol{B}) = 2m \times (\sigma_{\max}(\boldsymbol{T}))^{2(1-\frac{1}{m})}. \tag{S53}$$

V.3. Maximal Eigenvalue at the Canonical Solution (27)

In (20) we have

$$\lambda_{\max}\left(\boldsymbol{H}_{\boldsymbol{w}}\right) = \max_{\|\boldsymbol{b}\|=1} 2\|\boldsymbol{\Phi}\boldsymbol{b}\|^2. \tag{S54}$$

Note that $\|\boldsymbol{\Phi}\boldsymbol{b}\|^2 = \sum_{k=1}^m \|\boldsymbol{\Phi}_k\boldsymbol{b}\|^2$. Using the definition of $\boldsymbol{\Phi}_k$ in (16) we get

$$\|\boldsymbol{\Phi}_k\boldsymbol{b}\|^2 = \left\|\left(\prod_{j=1}^{k-1} \boldsymbol{W}_j \hat{\boldsymbol{\Sigma}}_x^{\frac{1}{2}}\right) \otimes \left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{b}\right\|^2 = \left\|\left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{B} \left(\prod_{j=1}^{k-1} \boldsymbol{W}_j \hat{\boldsymbol{\Sigma}}_x^{\frac{1}{2}}\right)^T\right\|_{\mathrm{F}}^2, \tag{S55}$$

where we used (P1) with $\boldsymbol{b} = \text{vec}\,(\boldsymbol{B})$. Therefore,

$$\lambda_{\max}(\boldsymbol{H_w}) = \max_{\|\boldsymbol{B}\|_{\text{F}}=1} 2\sum_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B} \hat{\boldsymbol{\Sigma}}_x^{\frac{1}{2}} \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_{\text{F}}^2. \tag{S56}$$

Substituting the canonical solution (27) in this optimization problem, we obtain

$$\sum_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B} \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_{\text{F}}^2 = \sum_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{S}_i^{\frac{1}{m}} \Big)^T \boldsymbol{U}^T \boldsymbol{B} \boldsymbol{V} \Big(\prod_{j=1}^{k-1} \boldsymbol{S}_j^{\frac{1}{m}} \Big)^T \Big\|_{\text{F}}^2, \tag{S57}$$

where in the first and the last terms of the series ($k = 1$ and $k = m$) we used the fact that $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary matrices, so that

$$\Big\| \Big(\prod_{i=2}^{m} \boldsymbol{S}_i^{\frac{1}{m}} \Big)^T \boldsymbol{U}^T \boldsymbol{B} \Big\|_{\text{F}}^2 = \Big\| \Big(\prod_{i=2}^{m} \boldsymbol{S}_i^{\frac{1}{m}} \Big)^T \boldsymbol{U}^T \boldsymbol{B} \boldsymbol{V} \Big\|_{\text{F}}^2, \qquad \Big\| \boldsymbol{B} \boldsymbol{V} \Big(\prod_{j=1}^{m-1} \boldsymbol{S}_j^{\frac{1}{m}} \Big)^T \Big\|_{\text{F}}^2 = \Big\| \boldsymbol{U}^T \boldsymbol{B} \boldsymbol{V} \Big(\prod_{j=1}^{m-1} \boldsymbol{S}_j^{\frac{1}{m}} \Big)^T \Big\|_{\text{F}}^2. \tag{S58}$$

Note that $\prod_{i=k+1}^{m} \boldsymbol{S}_i^{\frac{1}{m}}$ is a diagonal $d_y \times d_k$ matrix, whose qth diagonal entry is $(\sigma_q(\boldsymbol{T}))^{(m-k)/m}$ (where $\sigma_q(\boldsymbol{T})$ is the qth largest singular value of $\boldsymbol{T}$). Similarly, $\prod_{j=1}^{k-1} \boldsymbol{S}_j^{\frac{1}{m}}$ is a diagonal $d_{k-1} \times d_x$ matrix, whose qth diagonal entry is $(\sigma_q(\boldsymbol{T}))^{(k-1)/m}$. Therefore, we can write

$$\sum_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{S}_i^{\frac{1}{m}} \Big)^T \boldsymbol{U}^T \boldsymbol{B} \boldsymbol{V} \Big(\prod_{j=1}^{k-1} \boldsymbol{S}_j^{\frac{1}{m}} \Big)^T \Big\|_{\text{F}}^2 = \sum_{k=1}^{m} \Big\| \Big(\boldsymbol{S}^{\frac{m-k}{m}} \Big)^T \boldsymbol{U}^T \boldsymbol{B} \boldsymbol{V} \Big(\boldsymbol{S}^{\frac{k-1}{m}} \Big)^T \Big\|_{\text{F}}^2, \tag{S59}$$

where $\boldsymbol{S}^{\alpha}$ denotes a $d_y \times d_x$ diagonal matrix whose qth diagonal entry is $(\sigma_q(\boldsymbol{T}))^{\alpha}$. Here, we used the fact that the Frobenius norm is unaffected by zero entries, and thus removed/added zero rows/columns.

Next, we preform the change of variables $\tilde{\boldsymbol{B}} = \boldsymbol{U}^T \boldsymbol{B} \boldsymbol{V} \in \mathbb{R}^{d_y \times d_x}$ to obtain the following optimization problem

$$\max_{\tilde{\boldsymbol{B}} \in \mathbb{R}^{d \times d}} 2\sum_{k=1}^{m} \Big\| \Big(\boldsymbol{S}^{\frac{m-k}{m}} \Big)^T \tilde{\boldsymbol{B}} \Big(\boldsymbol{S}^{\frac{k-1}{m}} \Big)^T \Big\|_{\text{F}}^2 \qquad \text{s.t.} \qquad \|\tilde{\boldsymbol{B}}\|_{\text{F}}^2 = 1. \tag{S60}$$

Writing the objective in terms of the elements of $\tilde{\boldsymbol{B}}$, which we denote by $\{\tilde{b}_{i,j}\}$, gives

$$2\sum_{k=1}^{m} \Big\| \Big(\boldsymbol{S}^{\frac{m-k}{m}} \Big)^T \tilde{\boldsymbol{B}} \Big(\boldsymbol{S}^{\frac{k-1}{m}} \Big)^T \Big\|_{\text{F}}^2 = 2\sum_{k=1}^{m} \sum_{i=1}^{d} \sum_{j=1}^{d} \Big[(\sigma_j(\boldsymbol{T}))^{\frac{m-k}{m}} (\sigma_i(\boldsymbol{T}))^{\frac{k-1}{m}} \tilde{b}_{i,j} \Big]^2, \tag{S61}$$

where $d = \min\{d_x, d_y\}$ is the number of singular values of $\boldsymbol{T}$. By changing the order of the summation, we get

$$\max_{\tilde{b}_{1,1},\ldots,\tilde{b}_{d,d} \in \mathbb{R}} 2\sum_{i,j=1}^{d} \tilde{b}_{i,j}^2 \sum_{k=1}^{m} \Big[(\sigma_j(\boldsymbol{T}))^{\frac{m-k}{m}} (\sigma_i(\boldsymbol{T}))^{\frac{k-1}{m}} \Big]^2 \qquad \text{s.t.} \qquad \sum_{i,j=1}^{d} \tilde{b}_{i,j}^2 = 1. \tag{S62}$$

This is a simple linear optimization problem over the unit simplex, whose optimal value is attained at one of the vertices,

$$\max_{i,j \in \{1,\ldots,d\}} 2\sum_{k=1}^{m} \Big[(\sigma_j(\boldsymbol{T}))^{\frac{m-k}{m}} (\sigma_i(\boldsymbol{T}))^{\frac{k-1}{m}} \Big]^2. \tag{S63}$$

The maximal value is attained for $i = j = 1$, thus the value of (21) for the canonical solution is

$$2\sum_{k=1}^{m} \Big[(\sigma_1(\boldsymbol{T}))^{\frac{m-k}{m}} (\sigma_1(\boldsymbol{T}))^{\frac{k-1}{m}} \Big]^2 = 2m \times (\sigma_{\max}(\boldsymbol{T}))^{2(1-\frac{1}{m})}. \tag{S64}$$

This result shows that the canonical solution (27) is indeed a minimizer of the maximal eigenvalue of the Hessian matrix.

V.4. Proof of the Top Eigenvector of $\boldsymbol{H_w}$

On the one hand, according to Section 5, for any flattest minimum point $\boldsymbol{w} \in \Omega_0$, the largest eigenvalue satisfies

$$\lambda_{\max}(\boldsymbol{H_w}) = 2m \times (\sigma_{\max}(\boldsymbol{T}))^{2(1-\frac{1}{m})}. \tag{S65}$$

On the other hand, the maximal eigenvalue of the Hessian matrix is the solution to the optimization problem (S56), in which $\boldsymbol{b} = \text{vec}\,(\boldsymbol{B})$ is the eigenvector of $\hat{\boldsymbol{H}}_{\boldsymbol{w}}$ (see (S54)). Substituting $\boldsymbol{B}^* = \boldsymbol{u}\boldsymbol{v}^T$ (*i.e.* $\boldsymbol{b}^* = \boldsymbol{v} \otimes \boldsymbol{u}$) in the objective function, we get

$$\begin{aligned} 2\sum_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B}^* \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_{\text{F}}^2 &\geq 2m \times \left\| \left[\boldsymbol{B}^* \Big(\prod_{i=1}^{m} \boldsymbol{W}_i \Big)^T \right]^{m-1} \boldsymbol{B}^* \right\|_2^{\frac{2}{m}} \\ &= 2m \times \left\| \left(\boldsymbol{B}^* \boldsymbol{T}^T \right)^{m-1} \boldsymbol{B}^* \right\|_2^{\frac{2}{m}} \\ &= 2m \times (\sigma_{\max}(\boldsymbol{T}))^{2(1-\frac{1}{m})}, \end{aligned} \tag{S66}$$

where in the second inequality we used Lemma 4 and explicitly unrolled the product, as in (24), and in the last step we used (S52). This proves that $\boldsymbol{b}^* = \boldsymbol{v} \otimes \boldsymbol{u}$ is an eigenvector of $\hat{\boldsymbol{H}}_{\boldsymbol{w}}$ corresponding to the maximal eigenvalue. Now, since $\hat{\boldsymbol{H}}_{\boldsymbol{w}} = 2\boldsymbol{\Phi}^T\boldsymbol{\Phi}$ and $\boldsymbol{H_w} = 2\boldsymbol{\Phi}\boldsymbol{\Phi}^T$, we have that $\boldsymbol{\Phi}\boldsymbol{b}^* = \boldsymbol{\Phi}(\boldsymbol{v} \otimes \boldsymbol{u})$ is the eigenvector of $\boldsymbol{H_w}$ corresponding to its maximal eigenvalue.

VI. Proof of Theorem 2

Let us start the proof by presenting two lemmas.

Lemma S1. *Let $\hat{\boldsymbol{\Sigma}}_x = \boldsymbol{I}$. If $\boldsymbol{w} \in \Omega_0$ then for all $k \in \{1, 2, \ldots, m\}$*

$$\Big\| \boldsymbol{u}^T \prod_{i=k+1}^{m} \boldsymbol{W}_i \Big\| \Big\| \prod_{j=1}^{k-1} \boldsymbol{W}_j \boldsymbol{v} \Big\| = (\sigma_{\max}(\boldsymbol{T}))^{1-\frac{1}{m}}. \tag{S67}$$

Proof. First, observe that for $\boldsymbol{B}^* = \boldsymbol{u}\boldsymbol{v}^T$, the left-hand side of (S67) can be written as

$$\Big\| \boldsymbol{u}^T \prod_{i=k+1}^{m} \boldsymbol{W}_i \Big\| \Big\| \prod_{j=1}^{k-1} \boldsymbol{W}_j \, \boldsymbol{v} \Big\| = \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B}^* \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_2. \tag{S68}$$

Now, from Theorem 1 and Eq. (21) we can conclude that

$$\lambda_{\max}(\boldsymbol{H_w}) = 2\sum_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B}^* \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_2^2 = 2m \times (\sigma_{\max}(\boldsymbol{T}))^{2(1-\frac{1}{m})}. \tag{S69}$$

Note that since $\boldsymbol{B}^*$ is a rank-1 matrix, the entire expression within the norm is rank-1, which is the reason we could replace the Frobenius norm appearing in (21) by the operator norm (the two norms coincide for rank-1 matrices). Furthermore, by the inequality of arithmetic and geometric means

$$\begin{aligned} 2\sum_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B}^* \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_2^2 &\geq 2m \left[\prod_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B}^* \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_2^2 \right]^{\frac{1}{m}} \\ &\geq 2m \left[\Big\| \prod_{k=1}^{m} \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B}^* \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_2^2 \right]^{\frac{1}{m}} \\ &= 2m \times (\sigma_{\max}(\boldsymbol{T}))^{2(1-\frac{1}{m})}, \end{aligned} \tag{S70}$$

where the second inequality is due to the sub-multiplicativity property of the operator norm, and in the last step we unrolled the product, as in (24), and used (S52). From (S69) and (S70) we obtain that the inequality of arithmetic and geometric means in (S70) is achieved with equality. This happens if and only if all summands in the series are equal. Thus, we conclude that for all $k \in \{1, 2, \ldots, m\}$,

$$\Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B}^* \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_2 = (\sigma_{\max}(\boldsymbol{T}))^{1-\frac{1}{m}}, \tag{S71}$$

and together with (S68), this implies that

$$\Big\| \boldsymbol{u}^T \prod_{i=k+1}^{m} \boldsymbol{W}_i \Big\| \Big\| \prod_{j=1}^{k-1} \boldsymbol{W}_j \, \boldsymbol{v} \Big\| = (\sigma_{\max}(\boldsymbol{T}))^{1-\frac{1}{m}}. \tag{S72}$$

■

While Lemma S1 characterizes the norms of the vectors $\boldsymbol{u}^T \prod_{i=k+1}^{m} \boldsymbol{W}_i$ and $\prod_{j=1}^{k-1} \boldsymbol{W}_j \boldsymbol{v}$, the next Lemma characterizes their directions.

Lemma S2. *Let $\hat{\boldsymbol{\Sigma}}_x = \boldsymbol{I}$. If $\boldsymbol{w} \in \Omega_0$ then for all $k \in \{0, 1, 2, \ldots, m\}$*

$$\frac{1}{\Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{u} \Big\|} \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{u} = \frac{1}{\Big\| \prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v} \Big\|} \prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v}. \tag{S73}$$

Proof. From Lemma S1 we have

$$\prod_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{u} \Big\| \Big\| \boldsymbol{v}^T \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\| = (\sigma_{\max}(\boldsymbol{T}))^{m-1}. \tag{S74}$$

Recall that in our convention (see Section 4), for $k = 1, m$ we have

$$\Big\| \boldsymbol{v}^T \Big(\prod_{j=1}^{0} \boldsymbol{W}_j \Big)^T \Big\| = \|\boldsymbol{v}^T\| = 1, \qquad \Big\| \Big(\prod_{i=m+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{u} \Big\| = \|\boldsymbol{u}\| = 1. \tag{S75}$$

Therefore, (S74) can be written as

$$\prod_{k=1}^{m-1} \Big\| \boldsymbol{v}^T \Big(\prod_{j=1}^{k} \boldsymbol{W}_j \Big)^T \Big\| \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{u} \Big\| = (\sigma_{\max}(\boldsymbol{T}))^{m-1}. \tag{S76}$$

On the other hand, by the Cauchy–Schwarz inequality we have

$$\begin{aligned} \prod_{k=1}^{m-1} \Big\| \boldsymbol{v}^T \Big(\prod_{j=1}^{k} \boldsymbol{W}_j \Big)^T \Big\| \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{u} \Big\| &\geq \prod_{k=1}^{m-1} \boldsymbol{v}^T \Big(\prod_{j=1}^{k} \boldsymbol{W}_j \Big)^T \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{u} \\ &= \prod_{k=1}^{m-1} \boldsymbol{v}^T \boldsymbol{T}^T \boldsymbol{u} \\ &= (\sigma_{\max}(\boldsymbol{T}))^{(m-1)}. \end{aligned} \tag{S77}$$

From (S76) we have that the Cauchy–Schwartz inequalities are achieved with equality. Thus, for all $k \in \{0, 1, 2, \ldots, m\}$

$$\frac{1}{\Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{u} \Big\|} \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{u} = \frac{1}{\Big\| \prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v} \Big\|} \prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v}. \tag{S78}$$

■

Now we are ready to prove Theorem 2. From Lemma S2, we have

$$\frac{1}{\left\|\prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v}\right\|} \prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v} = \frac{1}{\left\|\left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{u}\right\|} \left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{u}. \tag{S79}$$

Multiplying by $(\prod_{j=1}^{k} \boldsymbol{W}_j)^T$ from the left, (S79) becomes

$$\frac{1}{\left\|\prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v}\right\|} \left(\prod_{j=1}^{k} \boldsymbol{W}_j\right)^T \prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v} = \frac{1}{\left\|\left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{u}\right\|} \boldsymbol{T}^T \boldsymbol{u}. \tag{S80}$$

Note that $\boldsymbol{T}^T \boldsymbol{u} = \sigma_{\max}(\boldsymbol{T}) \boldsymbol{v}$. Therefore,

$$\left[\left(\prod_{j=1}^{k} \boldsymbol{W}_j\right)^T \prod_{j=1}^{k} \boldsymbol{W}_j\right] \boldsymbol{v} = \frac{\left\|\prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v}\right\|}{\left\|\left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{u}\right\|} \sigma_{\max}(\boldsymbol{T}) \, \boldsymbol{v}. \tag{S81}$$

This shows that $\boldsymbol{v}$ is an eigenvector of $\left(\prod_{j=1}^{k} \boldsymbol{W}_j\right)^T \prod_{j=1}^{k} \boldsymbol{W}_j$, *i.e.* a singular vector of $\prod_{j=1}^{k} \boldsymbol{W}_j$. To compute the corresponding singular value, let us multiply this equation by $\boldsymbol{v}^T$ form the left to get the following result.

$$\left\|\prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v}\right\| \left\|\left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{u}\right\| = \sigma_{\max}(\boldsymbol{T}). \tag{S82}$$

Recall from Lemma S1 that $\|\boldsymbol{u}^T \prod_{i=k+1}^{m} \boldsymbol{W}_i\| = (\sigma_{\max}(\boldsymbol{T}))^{1-1/m} / \|\prod_{j=1}^{k-1} \boldsymbol{W}_j \boldsymbol{v}\|$. Substituting into (S82), we obtain that

$$\left\|\prod_{j=1}^{k} \boldsymbol{W}_j \, \boldsymbol{v}\right\| = (\sigma_{\max}(\boldsymbol{T}))^{\frac{1}{m}} \times \left\|\prod_{j=1}^{k-1} \boldsymbol{W}_j \, \boldsymbol{v}\right\|. \tag{S83}$$

By unwrapping this recursive formula with an initial condition for $k = 0$ of $\left\|\prod_{j=1}^{0} \boldsymbol{W}_j \boldsymbol{v}\right\| = \|\boldsymbol{v}\| = 1$, we get

$$\left\|\prod_{j=1}^{k} \boldsymbol{W}_j \boldsymbol{v}\right\| = \sigma_{\max}(\boldsymbol{T})^{\frac{k}{m}}. \tag{S84}$$

The proof for the left singular vector and its corresponding singular value is the same.

Next, we prove the bound on the intermediate gain. By Theorem 1 and Eq. (21),

$$\max_{\|\boldsymbol{B}\|_{\mathrm{F}}=1} \sum_{l=1}^{m} \left\|\left(\prod_{i=l+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{B} \left(\prod_{j=1}^{l-1} \boldsymbol{W}_j\right)^T\right\|_{\mathrm{F}}^2 = m \times (\sigma_{\max}(\boldsymbol{T}))^{2(1-\frac{1}{m})}. \tag{S85}$$

Now, for any k, we have that

$$\max_{\|\boldsymbol{B}\|_{\mathrm{F}}=1} \sum_{l=1}^{m} \left\|\left(\prod_{i=l+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{B} \left(\prod_{j=1}^{l-1} \boldsymbol{W}_j\right)^T\right\|_{\mathrm{F}}^2 \geq \max_{\|\boldsymbol{B}\|_{\mathrm{F}}=1} \left\|\left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{B} \left(\prod_{j=1}^{k-1} \boldsymbol{W}_j\right)^T\right\|_{\mathrm{F}}^2. \tag{S86}$$

Furthermore, note that

$$\begin{aligned}
\max_{\|\boldsymbol{B}\|_{\mathrm{F}}=1} \left\|\left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{B} \left(\prod_{j=1}^{k-1} \boldsymbol{W}_j\right)^T\right\|_{\mathrm{F}} &= \max_{\|b\|=1} \left\|\left(\prod_{j=1}^{k-1} \boldsymbol{W}_j\right) \otimes \left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T \boldsymbol{b}\right\| \\
&= \sigma_{\max}\left(\left(\prod_{j=1}^{k-1} \boldsymbol{W}_j\right) \otimes \left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right)^T\right) \\
&= \sigma_{\max}\left(\prod_{i=1}^{k-1} \boldsymbol{W}_i\right) \times \sigma_{\max}\left(\prod_{i=k+1}^{m} \boldsymbol{W}_i\right).
\end{aligned} \tag{S87}$$

Therefore, this implies that

$$\sigma_{\max}\left(\prod_{i=1}^{k-1}\boldsymbol{W}_i\right)\times\sigma_{\max}\left(\prod_{i=k+1}^{m}\boldsymbol{W}_i\right)\leq\sqrt{m}\times(\sigma_{\max}(\boldsymbol{T}))^{(1-\frac{1}{m})}, \tag{S88}$$

or, equivalently, that

$$\sigma_{\max}\left(\prod_{i=1}^{k-1}\boldsymbol{W}_i\right)\leq\frac{\sqrt{m}\times(\sigma_{\max}(\boldsymbol{T}))^{(1-\frac{1}{m})}}{\sigma_{\max}\big(\prod_{i=k+1}^{m}\boldsymbol{W}_i\big)}. \tag{S89}$$

By the first part of Theorem 2 we have

$$\sigma_{\max}\left(\prod_{i=k+1}^{m}\boldsymbol{W}_i\right)\geq\left(\prod_{i=k+1}^{m}\boldsymbol{W}_i\boldsymbol{u}\right)=(\sigma_{\max}(\boldsymbol{T}))^{(1-\frac{k}{m})}. \tag{S90}$$

Hence, we can further bound (S89) from above as

$$\sigma_{\max}\left(\prod_{i=1}^{k-1}\boldsymbol{W}_i\right)\leq\frac{\sqrt{m}\times(\sigma_{\max}(\boldsymbol{T}))^{(1-\frac{1}{m})}}{\sigma_{\max}\big(\prod_{i=k+1}^{m}\boldsymbol{W}_i\big)}\leq\sqrt{m}\times(\sigma_{\max}(\boldsymbol{T}))^{\frac{k-1}{m}}. \tag{S91}$$

The proof of the other direction is similar.

VII. Proof of Theorem 3

By Lemma S2

$$\frac{1}{\left\|\left(\prod_{i=k+1}^{m}\boldsymbol{W}_i\right)^T\boldsymbol{u}\right\|}\left(\prod_{i=k+1}^{m}\boldsymbol{W}_i\right)^T\boldsymbol{u}=\frac{1}{\left\|\prod_{j=1}^{k}\boldsymbol{W}_j\boldsymbol{v}\right\|}\prod_{j=1}^{k}\boldsymbol{W}_j\boldsymbol{v}. \tag{S92}$$

Multiplying both sides by $\boldsymbol{W}_k^T$ from the left, we obtain

$$\frac{1}{\left\|\left(\prod_{i=k+1}^{m}\boldsymbol{W}_i\right)^T\boldsymbol{u}\right\|}\left(\prod_{i=k}^{m}\boldsymbol{W}_i\right)^T\boldsymbol{u}=\frac{1}{\left\|\prod_{j=1}^{k}\boldsymbol{W}_j\boldsymbol{v}\right\|}\left(\boldsymbol{W}_k^T\boldsymbol{W}_k\right)\prod_{j=1}^{k-1}\boldsymbol{W}_j\boldsymbol{v}. \tag{S93}$$

Using Lemma S2 again for $k-1$ we have

$$\left(\prod_{i=k}^{m}\boldsymbol{W}_i\right)^T\boldsymbol{u}=\frac{\left\|\left(\prod_{i=k}^{m}\boldsymbol{W}_i\right)^T\boldsymbol{u}\right\|}{\left\|\prod_{j=1}^{k-1}\boldsymbol{W}_j\boldsymbol{v}\right\|}\prod_{j=1}^{k-1}\boldsymbol{W}_j\boldsymbol{v}. \tag{S94}$$

Plugging this equation in (S93) we get

$$\frac{\left\|\left(\prod_{i=k}^{m}\boldsymbol{W}_i\right)^T\boldsymbol{u}\right\|}{\left\|\prod_{j=1}^{k-1}\boldsymbol{W}_j\boldsymbol{v}\right\|\left\|\left(\prod_{i=k+1}^{m}\boldsymbol{W}_i\right)^T\boldsymbol{u}\right\|}\prod_{j=1}^{k-1}\boldsymbol{W}_j\boldsymbol{v}=\frac{1}{\left\|\prod_{j=1}^{k}\boldsymbol{W}_j\boldsymbol{v}\right\|}\left(\boldsymbol{W}_k^T\boldsymbol{W}_k\right)\prod_{j=1}^{k-1}\boldsymbol{W}_j\boldsymbol{v}. \tag{S95}$$

From Theorem 2, we know that $\|(\prod_{i=k}^{m}\boldsymbol{W}_i)^T\boldsymbol{u}\|=(\sigma_{\max}(\boldsymbol{T}))^{(m-k+1)/m}$, $\|\prod_{j=1}^{k}\boldsymbol{W}_j\boldsymbol{v}\|=(\sigma_{\max}(\boldsymbol{T}))^{k/m}$, and $\|\prod_{j=1}^{k-1}\boldsymbol{W}_j\boldsymbol{v}\|\|(\prod_{i=k+1}^{m}\boldsymbol{W}_i)^T\boldsymbol{u}\|=(\sigma_{\max}(\boldsymbol{T}))^{1-1/m}$. Therefore, (S95) can be reduced to

$$(\sigma_{\max}(\boldsymbol{T}))^{\frac{2}{m}}\times\boldsymbol{r}_k=\big(\boldsymbol{W}_k^T\boldsymbol{W}_k\big)\boldsymbol{r}_k. \tag{S96}$$

Hence, $\boldsymbol{r}_k$ is an eigenvector of $\boldsymbol{W}_k^T\boldsymbol{W}_k$ with a corresponding eigenvalue of $(\sigma_{\max}(\boldsymbol{T}))^{2/m}$. Namely, $\boldsymbol{r}_k/\|\boldsymbol{r}_k\|$ is a singular vector of $\boldsymbol{W}_k$ with a corresponding singular value of $(\sigma_{\max}(\boldsymbol{T}))^{1/m}$. Using Lemma S2 we have

$$\frac{1}{\|\boldsymbol{r}_k\|}\boldsymbol{W}_k\boldsymbol{r}_k=(\sigma_{\max}(\boldsymbol{T}))^{\frac{1}{m}}\times\frac{1}{\|\boldsymbol{q}_k\|}\boldsymbol{q}_k. \tag{S97}$$

From this equation we deduce that $\bar{\boldsymbol{r}}_k$ and $\bar{\boldsymbol{q}}_k$ are pair of singular vectors of $\boldsymbol{W}_k$, with a singular value of $(\sigma_{\max}(\boldsymbol{T}))^{1/m}$. Note that the equality $\bar{\boldsymbol{r}}_{k+1}=\bar{\boldsymbol{q}}_k$ is in fact the result of Lemma S2.

VIII. Proof of Theorem 4

On the one hand, according to Theorem 1

$$\min_{\tilde{\boldsymbol{w}}\in\Omega} \lambda_{\max}(\boldsymbol{H}_{\tilde{\boldsymbol{w}}}) = 2m \times \sigma_{\max}\big(\boldsymbol{T}\big)^{2(1-\frac{1}{m})}. \tag{S98}$$

On the other hand, given an arbitrary minimum point $\boldsymbol{w} \in \Omega$

$$\begin{aligned}\min_{\tilde{\boldsymbol{w}}\in\Omega} \lambda_{\max}(\boldsymbol{H}_{\tilde{\boldsymbol{w}}}) &\leq \lambda_{\max}(\boldsymbol{H}_{\boldsymbol{w}}) \\ &= \max_{\boldsymbol{B}\in\mathbb{R}^{d_y\times d_x}} 2\sum_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B} \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_{\mathrm{F}}^2 \qquad \text{s.t.} \qquad \|\boldsymbol{B}\|_{\mathrm{F}} = 1 \qquad \text{(S99)} \\ &\leq \max_{\boldsymbol{B}_1,\ldots,\boldsymbol{B}_m\in\mathbb{R}^{d_y\times d_x}} 2\sum_{k=1}^{m} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B}_k \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \|\boldsymbol{B}_1\|_{\mathrm{F}} = \cdots = \|\boldsymbol{B}_m\|_{\mathrm{F}} = 1.\end{aligned}$$

Here we obtain a separable optimization problem. Let us examine one term from the series

$$\max_{\boldsymbol{B}_k\in\mathbb{R}^{d_y\times d_x}} \Big\| \Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)^T \boldsymbol{B}_k \Big(\prod_{j=1}^{k-1} \boldsymbol{W}_j \Big)^T \Big\|_{\mathrm{F}}^2 \qquad \text{s.t.} \qquad \|\boldsymbol{B}_k\|_{\mathrm{F}} = 1. \tag{S100}$$

In (S87), we saw that the value of (S100) is

$$\Big(\sigma_{\max}\Big(\prod_{i=k+1}^{m} \boldsymbol{W}_i \Big)\Big)^2 \times \Big(\sigma_{\max}\Big(\prod_{i=1}^{k-1} \boldsymbol{W}_i \Big)\Big)^2 \leq \prod_{i\neq k} (\sigma_{\max}(\boldsymbol{W}_i))^2, \tag{S101}$$

where we used the sub-multiplicity property of the operator norm (the top singular value). If $\sigma_{\max}(\boldsymbol{W}_k) = (\sigma_{\max}(\boldsymbol{T}))^{1/m}$ for all k, then

$$\prod_{i\neq k} (\sigma_{\max}(\boldsymbol{W}_i))^2 = \prod_{i\neq k} (\sigma_{\max}(\boldsymbol{T}))^{\frac{2}{m}} = (\sigma_{\max}(\boldsymbol{T}))^{2(1-\frac{1}{m})}. \tag{S102}$$

Thus,

$$\lambda_{\max}(\boldsymbol{H}_{\boldsymbol{w}}) = 2m \times (\sigma_{\max}(\boldsymbol{T}))^{2(1-\frac{1}{m})}. \tag{S103}$$

Therefore $\boldsymbol{w}$ is a flattest minimum.

IX. More Details About Our Experiments

IX.1. Linear Networks

To sample arbitrary global minima, we started with the canonical solution (27), and multiplied the weight matrices by random matrices from the left and right, such that the the left matrix of one layer cancels out the right matrix of the next (thus keeping the end-to-end function unmodified). Specifically, let $\{\boldsymbol{A}_i\}_{i=1}^{m-1}$ be Gaussian random matrices with i.i.d. entries, distributed $\mathcal{N}(0,1)$. Then the weights for arbitrary solutions were generated as

$$\boldsymbol{W}_m = \boldsymbol{U}\boldsymbol{S}_m^{\frac{1}{m}}\boldsymbol{A}_{m-1}, \qquad \boldsymbol{W}_i = \boldsymbol{A}_i^{-1}\boldsymbol{S}_i^{\frac{1}{m}}\boldsymbol{A}_{i-1}, \qquad \boldsymbol{W}_1 = \boldsymbol{A}_1^{-1}\boldsymbol{S}_1^{\frac{1}{m}}\boldsymbol{V}^T. \tag{S104}$$

To obtain flattest minima, we minimized $\lambda_{\max}(\boldsymbol{H}_{\boldsymbol{w}})$ w.r.t. the weights, by taking random steps over the manifold of global minima Ω, and greedily progressing towards a flattest solution. In detail, we randomly generated a set of matrices $\{\boldsymbol{A}_i^0\}_{i=1}^m$ with i.i.d. normally distributed entries. We then set the initial weights of the network to be $\boldsymbol{W}_i^0 = \boldsymbol{A}_i^0$, for all $i \neq j$, and

$$\boldsymbol{W}_j^0 = \Big(\prod_{i=j+1}^{m} \boldsymbol{A}_i \Big)^{-1} \boldsymbol{T} \Big(\prod_{i=1}^{j-1} \boldsymbol{A}_i \Big)^{-1}, \tag{S105}$$

where j was a random integer chosen uniformly over $\{1,\ldots,m\}$. Next, we iteratively took small random steps over the manifold of global minima according to the following update rule.

$$\begin{aligned}\boldsymbol{W}_m^{t+1} &= \boldsymbol{W}_m^t\big(\boldsymbol{I}+\varepsilon_t\boldsymbol{A}_{m-1}^t\big),\\ \boldsymbol{W}_i^{t+1} &= \big(\boldsymbol{I}+\varepsilon_t\boldsymbol{A}_i^t\big)^{-1}\boldsymbol{W}_i^t\big(\boldsymbol{I}+\varepsilon_t\boldsymbol{A}_{i-1}^t\big),\\ \boldsymbol{W}_1^{t+1} &= \big(\boldsymbol{I}+\varepsilon_t\boldsymbol{A}_1^t\big)^{-1}\boldsymbol{W}_1^t,\end{aligned} \tag{S106}$$

where ε_t is the step size at the tth iteration, and $\{\boldsymbol{A}_i^t\}_{i=1}^m$ are again random matrices with i.i.d. normally distributed entries. We continued to the next iteration only if the spectral norm of the Hessian decreased. Otherwise, we generated an additional set of direction matrices $\{\boldsymbol{A}_i^t\}_{i=1}^m$ until we got a decrement. We stopped this process when the objective achieved its minimal value of $2m\times(\sigma_{\max}(\boldsymbol{T}))^{2(1-1/m)}$, up to a minor error.

IX.2. Nonlinear Networks

The table below summarizes the parameters and the results for the methods we used in the nonlinear setting for Fig. 6.

	Method 1	Method 2
Optimization Algorithm	SGD	Adam
Learning rate	$1/2$	3×10^{-4}
Other parameters	momentum = 0	$\beta_1=0.8,\ \beta_2=0.99$
Batch size	100	100
Train loss	$2.69\times10^{-2}\pm2.53\times10^{-5}$	$2.74\times10^{-2}\pm4.03\times10^{-5}$
Validation loss	$2.70\times10^{-2}\pm1.68\times10^{-4}$	$2.80\times10^{-2}\pm1.94\times10^{-4}$
$\lambda_{\max}$	$1.76\pm9.43\times10^{-3}$	12.9 ± 2.2

Table S1. Summary of the two methods we used to train the network.

References

Arora, S., Cohen, N., and Hazan, E. E. On the optimization of deep networks: Implicit acceleration by overparameterization. In *35th International Conference on Machine Learning, ICML 2018*, pp. 372–389. International Machine Learning Society (IMLS), 2018.

Nar, K. and Sastry, S. Step size matters in deep learning. In *Advances in Neural Information Processing Systems*, pp. 3436–3444, 2018.

Wu, L., Ma, C., and Weinan, E. How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective. In *Advances in Neural Information Processing Systems*, pp. 8279–8288, 2018.