OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Hessian Spectrum is Constant Across Minimizers in Regularized Deep Scalar Factorization

Anıl Kamber AKAMBER@UCSD.EDU
Rahul Parhi RAHUL@UCSD.EDU
Department of Electrical and Computer Engineering
University of California, San Diego

Abstract

We characterize the full Hessian spectrum of ℓ^2-regularized deep scalar factorization problems across all minimizers. We prove that the spectrum is constant across all minimizers and, in particular, that the maximum eigenvalue depends on the depth, the (shared) magnitude of optimal layers, and the regularization parameter. The limit of the regularization parameter to zero recovers the unregularized case for flat minima. To the best of our knowledge, our results offer the first complete characterization of Hessian spectrum across minimizers in deep-factorization-type problems.

1. Introduction

A key aspect in understanding the performance of neural networks is based on characterizing the role of the *geometry* of the loss landscape of the training objective. Indeed, a long-standing theory / heuristic used to explain why neural networks generalize well is that gradient-based training algorithms more easily find minima that are "flat" and that flat minima generalize well [5, 6]. This topic has had a recent resurgence and there are a growing number of theoretical results in various settings that aim to understand the role of flatness of the loss landscape [3, 4].

A key part of these analyses is quantifying the flatness of a point in the loss landscape. This is typically done by considering the maximum eigenvalue of the Hessian evaluated at that point. Despite the significance of this quantity, a complete theoretical understanding of the role of flatness still remains unclear. This is in part due to the fact that closed-form expressions for the maximum eigenvalue of the Hessian are unknown in many scenarios. While this is the case, there are some notable exceptions. These include the breakthrough work of Mulayoff and Michaeli [8] that derived a closed-form expression for the maximum eigenvalue of the Hessian for the *flattest minima* for deep linear neural networks trained with squared-error loss and no regularization. More recently, Singh and Hofmann [9] characterized the full Hessian spectrum for univariate, shallow neural networks with linear and rectified linear unit (ReLU) activation functions.

While these exceptions are a good start, obtaining closed-form expressions, especially for the full Hessian spectrum, remains an open problem. Furthermore, to the best of our knowledge, there are no characterizations with explicit regularization. Motivated by this observation, in this paper we provide a complete characterization, i.e., a closed-form expression, of the full Hessian spectrum across all minimizers for the problem of *deep scalar factorization* with explicit regularization.

2. Problem Setup

We consider the objective function for regularized deep scalar factorization

$$\mathcal{L}(\mathbf{w}) := (m - w_L w_{L-1} \cdots w_1)^2 + \frac{\lambda}{L} \sum_{i=1}^{L} w_i^2, \tag{1}$$

where $m \in \mathbb{R}$ denotes the scalar of interest, $L \geq 2$ is the depth, $w_i \in \mathbb{R}$ is the i^{th} factor (layer), $\lambda > 0$ is the regularization parameter, and $\mathbf{w} = (w_1, \ldots, w_L) \in \mathbb{R}^L$ denotes the full set of parameters. This objective is analogous to that of deep linear neural networks with scalar layers.

2.1. General Deep Matrix Factorization

Before we state and prove our complete characterization of the Hessian spectrum (see Theorem 3), we first examine the general deep matrix factorization problem to elucidate some of its properties. Consider the following objective function

$$\mathcal{L}(\mathbf{w}) := \underbrace{\|\mathbf{M} - \mathbf{W}_L \mathbf{W}_{L-1} \cdots \mathbf{W}_1\|_F^2}_{=: D(\mathbf{w})} + \underbrace{\frac{\lambda}{L} \sum_{i=1}^{L} \|\mathbf{W}_i\|_F^2}_{=: R(\mathbf{w})}, \tag{2}$$

where $\mathbf{M} \in \mathbb{R}^{d_L \times d_0}$ is the matrix of interest, $L \geq 2$ is the depth, $\mathbf{W}_i \in \mathbb{R}^{d_i \times d_{i-1}}$ is the i^{th} factor (layer), $\lambda > 0$ is the regularization parameter, and $\mathbf{w} = \text{vec}(\mathbf{W}_1, \ldots, \mathbf{W}_L)$ denotes the full set of parameters. This form of explicit regularization is well known to be equivalent to Schatten-$(2/L)$ regularization [2, 10]. Define the solution set of the optimization criterion as $\Omega := \arg\min_{\mathbf{w} \in \mathbb{R}^N} \mathcal{L}(\mathbf{w})$, where $N = \sum_{i=1}^{L} d_i d_{i-1}$ is the total number of parameters.

A recent observation by Chen et al. [1, Lemma 3.2] showed that, at each minimizer of $\mathcal{L}$, every layer has exactly the same singular values.

Theorem 1 ([1, Lemma 3.2]) *For any $\mathbf{w}^* \in \Omega$, layers (factors) are balanced, i.e.,*

$$\mathbf{W}_i^* \, \mathbf{W}_i^{*\top} = \mathbf{W}_{i+1}^{*\top} \, \mathbf{W}_{i+1}^* \quad \forall i \in [L-1], \tag{3}$$

which implies that layers possess exactly same singular values. Furthermore, if the singular values are distinct, then their left and right singular vectors align, up to an incurable sign ambiguity.

In particular, this balanced property will play a key role in our analyses in Theorem 2 and Section 2.2. Note that if $\mathbf{M} \neq \mathbf{0}$ then for all $\mathbf{w}^* \in \Omega$ it holds that $\mathcal{L}(\mathbf{w}^*) > 0$. Also observe that the *optimal value* is

$$J^* := \mathcal{L}(\mathbf{w}^*) = D(\mathbf{w}^*) + \lambda R(\mathbf{w}^*) \quad \text{for all } \mathbf{w}^* \in \Omega. \tag{4}$$

What is even more remarkable is that the data-fitting value $D(\cdot)$ and the regularization value $R(\cdot)$ are also constants across all minimizers for regularized deep scalar factorization. This is summarized in the next theorem.

Theorem 2 *Consider the objective function for regularized deep scalar factorization in (1). For any $\mathbf{w}^* \in \Omega$,*

$$D(\mathbf{w}^*) = D^*, \quad R(\mathbf{w}^*) = R^*, \tag{5}$$

where D^* and R^* are the (shared) optimal data-fitting and regularization values, respectively.

Proof Theorem 1 implies that layers (factors) are balanced across all minimizers, i.e., $|w_L^*| = |w_{L-1}^*| = \cdots = |w_1^*|$ for all $\mathbf{w}^* \in \Omega$. Because of this, we can restrict the optimization to a subspace of $\mathbb{R}^L$ where all points are balanced and formulate an equivalent optimization problem to (1) that is

$$\min_{\rho \in \mathbb{R}} (m - \rho)^2 + \lambda |\rho|^{2/L}, \tag{6}$$

where ρ denotes the end-to-end function. Therefore, to prove Theorem 2, it is sufficient to show that the minimizer of (6) is unique. Note that for $L \in \{1, 2\}$, the objective function is strictly convex; therefore, the optimal solution ρ^* is unique. Hence, it is trivial to examine the case where $L \in \{1, 2\}$. Now, suppose $L \geq 3$. To investigate the behavior of the solution set of the new formulation in this case, we first assume that $m = 0$. Then

$$\rho^2 + \lambda |\rho|^{2/L} \geq 0. \tag{7}$$

Since ρ^2 and $|\rho|^{2/L}$ are both nonnegative, the lower bound is achieved if and only if $\rho = 0$. Therefore, ρ^* is unique. If $m \neq 0$, we can investigate the problem in two cases, where $m > 0$ and $m < 0$. Before delving into our analysis, to simplify the notation, let us define $\phi : \mathbb{R} \to \mathbb{R}$ and q such that $\phi(\rho) := (m - \rho)^2 + \lambda |\rho|^q$ and $q := 2/L$, where $L \geq 3$.

We first investigate the case where $m > 0$. For any minimizer ρ^*, we have that

$$\phi(\rho^*) - \phi(-\rho^*) \leq 0. \tag{8}$$

Suppose that $\rho^* < 0$. Then,

$$\phi(\rho^*) - \phi(-\rho^*) = (m - \rho^*)^2 + \lambda |\rho^*|^q - (m + \rho^*)^2 - \lambda |\rho^*|^q = -4m\rho^* > 0, \tag{9}$$

which is a contradiction. We now consider the case $m < 0$. Suppose that $\rho^* > 0$. By symmetry, this also leads to a contradiction. Hence, for any minimizer ρ^*, we have that

$$\operatorname{sign}(\rho^*) = \operatorname{sign}(m). \tag{10}$$

Observe that $|\rho|^q$ is symmetric and for any $m \in \mathbb{R}$ such that $f(x) = (x - m)^2$, we have $f(-x) = (x + m)^2$. Therefore, it is sufficient to prove that $\phi(\rho)$ has a unique minimizer over $\rho > 0$ when $m > 0$. Then,

$$\phi(\rho) = (m - \rho)^2 + \lambda \rho^q, \quad \rho, m > 0, \tag{11}$$

where $0 < q < 1$. Note that

$$\lim_{\rho \to 0} \phi(\rho) = m^2 \quad \text{and} \quad \lim_{\rho \to \infty} \phi(\rho) = \infty, \tag{12}$$

$$\lim_{\rho \to 0} \phi'(\rho) = \infty \quad \text{and} \quad \lim_{\rho \to \infty} \phi'(\rho) = \infty, \tag{13}$$

$$\lim_{\rho \to 0} \phi''(\rho) = -\infty \quad \text{and} \quad \lim_{\rho \to \infty} \phi''(\rho) = 2, \tag{14}$$

$$\phi'''(\rho) = \lambda q(q - 1)(q - 2)\rho^{q-3}. \tag{15}$$

Note that $\phi'(\rho)$ is strictly convex since $\phi'''(\rho)$ is positive for all $\rho > 0$. If $\phi'(\rho) \geq 0$ then $\phi(\rho)$ is monotonically increasing and $\rho^* = 0$. However, this is not the case due to (10). This implies that $\phi(\rho)$ has one local minimum and one local maximum for $\rho > 0$. Finally, since $\rho = 0$ is not a minimizer, the local minimum is also global and unique. ■

The implication of Theorem 2 is that both the data-fitting error and the regularization error remain constant across all minimizers in regularized deep scalar factorization.

2.2. The Hessian Spectrum for Deep Scalar Factorization

In the next theorem, we derive the full Hessian spectrum across all minimizers for the regularized deep scalar factorization problem.

Theorem 3 *Consider the deep scalar factorization objective*

$$\mathcal{L}(\mathbf{w}) := (m - w_L w_{L-1} \cdots w_1)^2 + \frac{\lambda}{L} \sum_{i=1}^{L} w_i^2, \tag{16}$$

where $m, w_1, w_2, \cdots, w_L \in \mathbb{R}$ and $\mathbf{w} = (w_1, \ldots, w_L) \in \mathbb{R}^L$. Then, for any $\mathbf{w}^* \in \Omega$, the spectrum of $\nabla^2 \mathcal{L}(\mathbf{w}^*)$ is constant over the solution set Ω. In particular, the eigenvalues are

$$\lambda_1 = \lambda_2 = \cdots = \lambda_{L-1} = \frac{4\lambda}{L}, \quad \lambda_L = 2Lw^{2L-2} + \frac{4\lambda}{L} - 2\lambda, \tag{17}$$

where $w = |w_1^*| = \cdots = |w_L^*|$. In which case, for any $\mathbf{w}^* \in \Omega$,

$$\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{w}^*)) = \max\Big\{2Lw^{2L-2} - 2\lambda, 0\Big\} + \frac{4\lambda}{L}. \tag{18}$$

Proof Differentiating $\mathcal{L}(\mathbf{w})$ with respect to w_j, $j \in [L]$, yields

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial w_j} = -2\left(m - \prod_{i=1}^{L} w_i\right) \prod_{\substack{i=1 \\ i \neq j}}^{L} w_i + \frac{2\lambda}{L} w_j. \tag{19}$$

Next, for any $\mathbf{w}^* \in \Omega$, we have that $\nabla \mathcal{L}(\mathbf{w}^*) = 0$. Hence, for any $j \in [L]$, we have that

$$-2\left(m - \prod_{i=1}^{L} w_i^*\right) \prod_{\substack{i=1 \\ i \neq j}}^{L} w_i^* + \frac{2\lambda}{L} w_j^* = 0. \tag{20}$$

Let $\rho(\mathbf{w}) := \prod_{i=1}^{L} w_i$. By Theorem 2, we have that $\rho(\mathbf{w})$ is constant for any $\mathbf{w} \in \Omega$. Write $\rho^* := \rho(\mathbf{w}^*)$ with $\mathbf{w}^* \in \Omega$. Then,

$$2(m - \rho^*) \frac{\rho^*}{w_j^*} = \frac{2\lambda}{L} w_j^*. \tag{21}$$

For any $\mathbf{w} \in \mathbb{R}^L$, we have that

$$\frac{\partial^2 \mathcal{L}(\mathbf{w})}{\partial w_j \partial w_k} = \begin{cases} 2\left(\dfrac{\rho(\mathbf{w})}{w_j}\right)^2 + \dfrac{2\lambda}{L}, & j = k, \\ 2\dfrac{\rho(\mathbf{w})^2}{w_j w_k} - 2(m - \rho(\mathbf{w}))\dfrac{\rho(\mathbf{w})}{w_j w_k}, & j \neq k. \end{cases} \tag{22}$$

Now, we fix $\mathbf{w}^* \in \Omega$ and define

$$\mathbf{s}^* := (\operatorname{sign}(w_1^*), \ldots, \operatorname{sign}(w_L^*)) \in \mathbb{R}^L. \tag{23}$$

By Theorem 1, we know that $|w_1^*| = \cdots = |w_L^*| =: w$, where we note that $w \in \mathbb{R}_{\geq 0}$ is the same across all minimizers by Theorem 2. Then,

$$\left[\nabla^2 \mathcal{L}(\mathbf{w}^*)\right]_{j,k} = \begin{cases} \dfrac{2\rho^{*2}}{w^2} + \dfrac{2\lambda}{L}, & j = k, \\ \dfrac{C}{w^2 s_j s_k}, & j \neq k, \end{cases} \tag{24}$$

where $C = 4\rho^{*2} - 2m\rho^*$. Since $\nabla^2 \mathcal{L}(\mathbf{w}^*)$ is a real symmetric matrix, it admits an orthonormal basis of eigenvectors $\{\mathbf{v}_i\}_{i=1}^L$ with corresponding eigenvalues $\{\lambda_i\}_{i=1}^L$. In particular,

$$\nabla^2 \mathcal{L}(\mathbf{w}^*)\, \mathbf{v}_i \;=\; \lambda_i\, \mathbf{v}_i, \tag{25}$$

and

$$\left[\nabla^2 \mathcal{L}(\mathbf{w}^*)\, \mathbf{v}_i\right]_j = \lambda_i v_{ij} = \left(\frac{2\rho^{*2}}{w^2} + \frac{2\lambda}{L}\right) v_{ij} + \sum_{\substack{k=1 \\ k \neq j}}^{L} \frac{C}{w^2 s_j s_k} v_{ik}. \tag{26}$$

Since $|s_j| = 1$, we can move s_j to the numerator. Therefore,

$$\lambda_i v_{ij} = \left(\frac{2\rho^{*2}}{w^2} + \frac{2\lambda}{L}\right) v_{ij} + \sum_{\substack{k=1 \\ k \neq j}}^{L} \frac{C s_j s_k}{w^2} v_{ik}. \tag{27}$$

Furthermore, we can express the right-hand side as a dot product, i.e.,

$$\lambda_i v_{ij} = \left(\frac{2\rho^{*2}}{w^2} + \frac{2\lambda}{L}\right) v_{ij} + \frac{C s_j}{w^2}\left(\mathbf{s}^{*\top} \mathbf{v}_i - s_j\, v_{ij}\right) \tag{28}$$

$$= \left(\frac{2\rho^{*2}}{w^2} + \frac{2\lambda}{L}\right) v_{ij} + \frac{C s_j}{w^2} \mathbf{s}^{*\top} \mathbf{v}_i - \frac{C}{w^2} v_{ij}. \tag{29}$$

We can rearrange the terms on the right-hand side to find that

$$v_{ij}\left(\lambda_i - \left(\frac{2\rho^{*2}}{w^2} + \frac{2\lambda}{L}\right) + \frac{C}{w^2}\right) = \frac{C}{w^2} s_j \mathbf{s}^{*\top} \mathbf{v}_i. \tag{30}$$

Next, define

$$\widetilde{\lambda}_i := \lambda_i - \left(\frac{2\rho^{*2}}{w^2} + \frac{2\lambda}{L}\right) + \frac{C}{w^2}. \tag{31}$$

Without loss of generality, assume that $C/w^2 \neq 0$. Then, we obtain

$$\mathbf{v}_i \frac{w^2 \widetilde{\lambda}_i}{C} \;=\; \mathbf{s}^*\, \mathbf{s}^{*\top}\, \mathbf{v}_i. \tag{32}$$

This implies that $\mathbf{v}_i$ is an eigenvector of $\mathbf{s}^* \, \mathbf{s}^{*\top}$ with corresponding eigenvalue $(w^2\widetilde{\lambda}_i)/C$ for all $i \in [L]$. Therefore, we have the eigendecomposition

$$\mathbf{s}^* \, \mathbf{s}^{*\top} = \begin{bmatrix} \dfrac{\mathbf{s}^*}{\|\mathbf{s}^*\|_2} & \mathbf{v}_2 & \cdots & \mathbf{v}_L \end{bmatrix} \begin{bmatrix} \|\mathbf{s}^*\|_2^2 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} \dfrac{\mathbf{s}^*}{\|\mathbf{s}^*\|_2} & \mathbf{v}_2 & \cdots & \mathbf{v}_L \end{bmatrix}^\top. \tag{33}$$

Observe that the only nonzero eigenvalue of $\mathbf{s}^* \, \mathbf{s}^{*\top}$ is $\|\mathbf{s}^*\|_2^2 = L$, where note that this equality holds for *any minimizer* $\mathbf{w}^* \in \Omega$. Without loss of generality, we will identify this eigenvalue with $i = 1$, i.e., $(w^2\widetilde{\lambda}_1)/C = L$. Therefore, for any minimizer, we must satisfy the two equalities

$$\frac{w^2\widetilde{\lambda}_i}{C} = 0 \quad \text{and} \quad \frac{w^2\widetilde{\lambda}_1}{C} = L \quad \Leftrightarrow \quad \widetilde{\lambda}_i = 0 \quad \text{and} \quad \widetilde{\lambda}_1 = \frac{C}{w^2}L, \quad i \in \{2, \ldots, L\}. \tag{34}$$

Now, suppose that $i \in \{2, \ldots, L\}$, i.e., $\widetilde{\lambda}_i = 0$, . Then,

$$\lambda_i = \Big(\frac{2\rho^{*2}}{w^2} + \frac{2\lambda}{L}\Big) - \frac{C}{w^2} = \frac{2\rho^{*2}}{w^2} - \frac{4\rho^{*2}}{w^2} + \frac{2m\,\rho^*}{w^2} + \frac{2\lambda}{L} = \frac{2\rho^*(m - \rho^*)}{w^2} + \frac{2\lambda}{L}. \tag{35}$$

Since $2(m - \rho^*)\rho^* = w^2(2\lambda)/L$, we get

$$\lambda_i = \frac{2\lambda}{L} + \frac{2\lambda}{L} = \frac{4\lambda}{L}. \tag{36}$$

When $i = 1$, i.e., $\widetilde{\lambda}_1 = LC/w^2$, we have that

$$\lambda_1 - \left(\frac{2\rho^{*2}}{w^2} + \frac{2\lambda}{L}\right) + \frac{C}{w^2} = L\frac{C}{w^2} \quad \Rightarrow \quad \lambda_1 = 2L(w^{2L-2}) + \frac{4\lambda}{L} - 2\lambda. \tag{37}$$

To complete the proof, we observe that we can write the $\lambda_{\max}$ as follows

$$\lambda_{\max} = \max\{2L(w^{2L-2}) - 2\lambda, 0\} + \frac{4\lambda}{L}. \tag{38}$$

■

Remark 4 *Note that when the regularization parameter $\lambda = 0$, we have that $\lambda_i = 0$ for all $i \in \{2, 3, \cdots, L\}$. Furthermore, observe that $w^L = |m| = \sigma_{\max}(m)$, which implies that*

$$\lambda_{\max} = 2L\,\sigma_{\max}(m)^{2\left(1-\frac{1}{L}\right)}. \tag{39}$$

Therefore, our theorem recovers the result of Mulayoff and Michaeli [8, Theorem 1].

3. Discussion and Conclusion

In this paper, we derived the full Hessian spectrum for minimizers of regularized deep scalar factorization problems. An interesting consequence of our results is that the Hessian spectrum is constant across all minimizers. Without regularization, as shown by Mulayoff and Michaeli [8], the Hessian is rank-deficient at minima by at least the order of $1-(1/L)$; that is, at least $1-(1/L)$ of the singular values of the Hessian are zero at a global minimum. Meanwhile, with regularization, as we have shown in Theorem 3, the Hessian is full-rank at minimizers. This implies that all minimizers are *isolated* in the parameter space. Recently, it has been empirically observed that explicit regularization prevents edge-of-stability phenomenon [7, Figure 3]. The results of this paper may help to understand why explicit regularization prevents edge-of-stability phenomenon, and we leave the details to future work. Furthermore, finding a closed-form expression for the largest eigenvalue of the Hessian at a minimum of the general ℓ^2-regularized deep matrix factorization still remains an open problem.

References

[1] Po Chen, Rujun Jiang, and Peng Wang. A complete loss landscape analysis of regularized deep matrix factorization. *arXiv preprint arXiv:2506.20344*, 2025.

[2] Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. *Advances in Neural Information Processing Systems*, 34:26884–26896, 2021.

[3] Lijun Ding, Dmitriy Drusvyatskiy, Maryam Fazel, and Zaid Harchaoui. Flat minima generalize for low-rank matrix recovery. *Information and Inference: A Journal of the IMA*, 13(2):iaae009, 2024.

[4] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[5] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[6] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

[7] Tongtong Liang, Dan Qiao, Yu-Xiang Wang, and Rahul Parhi. Stable minima of ReLU neural networks suffer from the curse of dimensionality: The neural shattering phenomenon. In *Advances in Neural Information Processing Systems*, 2025.

[8] Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In *International Conference on Machine Learning*, pages 7108–7118. PMLR, 2020.

[9] Sidak Pal Singh and Thomas Hofmann. Closed form of the Hessian spectrum for some neural networks. In *High-dimensional Learning Dynamics*, 2024.

[10] Yifei Wang, Tolga Ergen, and Mert Pilanci. Parallel deep neural networks have zero duality gap. In *The Eleventh International Conference on Learning Representations*, 2023.