Sharpness of Minima in Deep Matrix Factorization[†]

Anıl Kamber

Joint work with Prof. Rahul Parhi

UCSD

[†] Anıl Kamber and Rahul Parhi (2026). "Sharpness of Minima in Deep Matrix Factorization". In: *The Fortieth Annual Conference on Neural Information Processing Systems*.

Motivation

- Generally, in overparameterized settings, the loss function has infinitely many global minima that achieve zero training error (interpolation regime). Yet, these models still perform well.
- It has been shown that *sharp minima* can be dynamically unstable for gradient-based methods; therefore, they avoid sharp minima and converge to *flat minima*.[1]
- However, the link between flatness and generalization remains unclear.[2]
- A central quantity to characterize the loss-landscape geometry is the maximum eigenvalue of the Hessian of the loss. Its precise role has been obfuscated because no exact expressions for this sharpness measure are known in general settings.

[1] Lei Wu, Chao Ma, et al. (2018). "How SGD selects the global minima in over-parameterized learning: A dynamical stability perspective". In: *Advances in Neural Information Processing Systems* 31.

[2] Laurent Dinh et al. (2017). "Sharp minima can generalize for deep nets". In: *International Conference on Machine Learning*. PMLR; Yiding Jiang et al. (2020). "Fantastic Generalization Measures and Where to Find Them". In: *International Conference on Learning Representations*.

Problem Setup

We consider the following optimization problem

$$\min_{\mathbf{w}\in\mathbb{R}^N} \mathcal{L}(\mathbf{w}) := \|\mathbf{M} - \mathbf{W}_L\mathbf{W}_{L-1}\cdots\mathbf{W}_1\|_F^2,$$

where $\mathbf{w} = \mathrm{vec}([\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_L])$ denotes the collection of all parameters. We define the set of global minima of $\mathcal{L}(\mathbf{w})$ as

$$\Omega := \arg\min_{\mathbf{w}\in\mathbb{R}^N} \mathcal{L}(\mathbf{w}) = \left\{\mathbf{w} \in \mathbb{R}^N : \prod_{i=1}^{L} \mathbf{W}_i = \mathbf{M}\right\}$$

and define the set of flat minima as

$$\Omega_F := \arg\min_{\mathbf{w}^*\in\Omega} \lambda_{\max}\left(\nabla^2\mathcal{L}(\mathbf{w}^*)\right).$$

To guarantee the feasibility of factorization for all points in $\mathbb{R}^{d_L\times d_0}$, we require

$$d_i \geq \min\{d_0, d_L\} \quad \forall i \in [L-1].$$

Main Results

Theorem (3.1)

If $\mathbf{w}^* \in \Omega$, *then*

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}\left(\mathbf{w}^*\right)\right) = 2\sigma_{\max}\left(\sum_{i=1}^{L} \mathbf{B}_i^\top \mathbf{B}_i \otimes \mathbf{A}_i \mathbf{A}_i^\top\right),$$

where $\mathbf{A}_k = \prod_{i=k+1}^{L} \mathbf{W}_i^*$ *and* $\mathbf{B}_k = \prod_{i=1}^{k-1} \mathbf{W}_i^*$.

Main Results cont.

Corollary (3.2)

Consider the following objective function

$$\mathcal{L}(\mathbf{w}) := (m - \mathbf{w}_L^\top \mathbf{W}_{L-1} \cdots \mathbf{W}_2 \mathbf{w}_1)^2,$$

where $m \in \mathbb{R}$, $d_0 = d_L = 1$. *Then, for all* $\mathbf{w}^* \in \Omega$,

$$\lambda_{\max}\left(\nabla^2 \mathcal{L}\left(\mathbf{w}^*\right)\right) = 2\sum_{i=1}^{L} \sigma_{\max}\left(\prod_{j=i+1}^{L} \mathbf{W}_j^*\right)^2 \sigma_{\max}\left(\prod_{j=1}^{i-1} \mathbf{W}_j^*\right)^2.$$

Corollary (3.3)

Consider the following depth-2 matrix factorization objective

$$\mathcal{L}(\mathbf{L}, \mathbf{R}) = \left\|\mathbf{M} - \mathbf{L}\mathbf{R}^\top\right\|_F^2,$$

where $\mathbf{L} \in \mathbb{R}^{d_L \times k}$, $\mathbf{R} \in \mathbb{R}^{d_0 \times k}$ *such that* $k \geq \min\{d_0, d_L\}$. *If* $(\mathbf{L}^*, \mathbf{R}^*) \in \Omega$, *then*

$$\lambda_{\max}(\nabla^2 \mathcal{L}(\mathbf{L}^*, \mathbf{R}^*)) = 2(\sigma_{\max}(\mathbf{L}^*)^2 + \sigma_{\max}(\mathbf{R}^*)^2).$$

Flatness

These results lead to the following remarkable observations:

- A minimizer of the deep overparameterized scalar factorization loss is flat **if and only if**, across all layers, the product of the spectral norms of the left and right intermediate factors equals a constant determined by the depth and the magnitude of the target (Corollary 4.1):

$$\sigma_{\max}(\mathbf{A}_k)\sigma_{\max}(\mathbf{B}_k) = |m|^{1-\frac{1}{L}} \quad \forall k \in [L].$$

- Flat minima are spectral-norm balanced in depth-2 matrix factorization and not necessarily balanced[3] (Corollary 4.2):

$$\sigma_{\max}(\mathbf{L}^*) = \sigma_{\max}(\mathbf{R}^*) = \sqrt{\sigma_{\max}(\mathbf{M})}.$$

- A minimizer of the deep matrix factorization loss is flat **if**, across all layers, the product of spectral norms of the left and right intermediate factors equals a constant determined by the depth and the spectral norm of the target matrix (Corollary 4.4). We further show, via an explicit counterexample, that the converse does not hold in general (Remark 4.5).

[3] A minimizer $\mathbf{w}^* \in \Omega$ is balanced if $(\mathbf{W}_{k+1}^*)^\top \mathbf{W}_{k+1}^* = \mathbf{W}_k^* (\mathbf{W}_k^*)^\top$ for all $k \in [L-1]$.

Are Flat Minima Frobenius-norm Balanced?

- When flatness is measured by the *scaled trace*[4], it has been shown that flat minima are Frobenius-norm balanced in depth-2 matrix factorization.

Lemma

Consider the deep matrix factorization problem with squared error loss. Then for any $\mathbf{w}^* \in \Omega$, *the scaled trace of the Hessian matrix is given by*

$$\mathrm{str}(\nabla^2 \mathcal{L}(\mathbf{w}^*)) = 2\|\prod_{i=2}^{L} \mathbf{W}_i^*\|_F^2 + 2\left(\sum_{k=2}^{L-1} \|\prod_{i=k+1}^{L} \mathbf{W}_i^*\|_F^2 \|\prod_{i=1}^{k-1} \mathbf{W}_i^*\|_F^2\right) + 2\|\prod_{i=1}^{L-1} \mathbf{W}_i^*\|_F^2.$$

[4] Lijun Ding et al. (2024). "Flat minima generalize for low-rank matrix recovery". In: *Information and Inference: A Journal of the IMA* 13.2.

Summary

- Sharpness measures are delusive!
- There is still no consensus in the literature regarding the definition of flatness. For instance, while several works define flat minima as global minima that minimize the maximum eigenvalue of the Hessian of the loss[5], others define them as global minima that minimize the trace of the Hessian.[6] Even a scaled version of the Hessian trace is designed to define flat minima in depth-2 matrix factorization problems, as we mentioned before. As we have shown, they can lead to contradictory interpretations of the loss landscape.
- Thus, we need a robust sharpness measure that consolidates the interpretations from existing ones!

[5] Pierre Marion and Lénaïc Chizat (2024). "Deep linear networks for regression are implicitly regularized towards flat minima". In: *Advances in Neural Information Processing Systems* 37, pp. 76848–76900.

[6] Khashayar Gatmiry et al. (2023). "What is the inductive bias of flatness regularization? A study of deep matrix factorization models". In: *Advances in Neural Information Processing Systems* 36, pp. 28040–28052.