Connecting Parameter Magnitudes and Hessian Eigenspaces at Scale using Sketched Methods

Andres Fernandez, Frank Schneider, Maren Mahreseci, Philipp Hennig

{a.fernandez, f.schneider, maren.mahsereci, philipp.hennig}@uni-tuebingen.de

April 11, 2025

TRANSACTIONS tmlr on ML RESEARCH

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Faculty of Science
Department of Computer Science
Methods of Machine Learning

Early crystallization in deep learning

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Setup:

- **Dataset** of pairs $(\boldsymbol{x}_i, \boldsymbol{y}_i)$
- **Neural Network** $\hat{\boldsymbol{y}}_i = f(\boldsymbol{x}_i, \boldsymbol{\theta})$ with **parameters** $\boldsymbol{\theta} \in \mathbb{R}^D$
- **Loss** $\mathcal{L}(\boldsymbol{\theta}) = \sum_i \ell(\boldsymbol{y}_i, f(\boldsymbol{x}_i, \boldsymbol{\theta})) \in \mathbb{R}_{\geq 0}$ with **gradient** $\boldsymbol{g} \in \mathbb{R}^D$ and **Hessian** $\boldsymbol{H} \in \mathbb{R}^{D \times D}$

Early crystallization of **parameters**:

- Parameters can be *pruned* (Blalock et al. 2020)
- Pruning masks $\boldsymbol{\theta} \odot \boldsymbol{m}$ appear *early* in training (Frankle et al. 2019)
- Magnitude pruning masks *don't change much* during training! (You et al. 2020)

Early crystallization of **loss landscape**:

- $\boldsymbol{H}$ is rank-defficient, i.e. $\boldsymbol{H} \approx \boldsymbol{U}_{top} \boldsymbol{\Lambda}_{top} \boldsymbol{U}_{top}^\top$ (e.g. Sagun et al. 2018)
- $\boldsymbol{g}$ resides mostly in $\boldsymbol{U}_{top}$ (Gur-Ari et al. 2019)
- $\text{span}(\boldsymbol{U}_{top})$ *doesn't change much* during training! (Gur-Ari et al. 2019)

Cheap!

Are those connected?

Expensive!

Research questions and contributions

Questions:

- Can this similarity be measured? If so, how?
- What similarity can be considered high? What are the implications?

Contributions:

- Methodology to compare arbitrary *k*-parameter masks to *top-k* Hessian eigenspaces
- Algorithm and code to perform said measurements at scale $\longrightarrow$ Hessian eigendecompositions
- **In DL, connection is orders of magnitude larger than random**
- Potential implications for pruning, optimization, UQ and loss landscape analysis

Comparing parameters with Hessian subspaces

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Top-k parameter pruning is a projection onto $\boldsymbol{I}_{D,k}$:

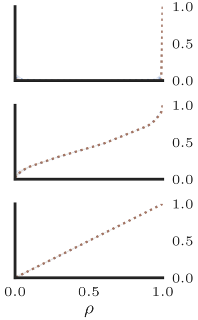
$$\boldsymbol{P}^\top(\boldsymbol{m}_k \odot \boldsymbol{\theta}) = \tilde{\boldsymbol{m}}_k \odot \tilde{\boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{I}_k & 0 \\ 0 & 0 \end{pmatrix} \tilde{\boldsymbol{\theta}} =: \boldsymbol{I}_{D,k} \boldsymbol{I}_{D,k}^\top \tilde{\boldsymbol{\theta}}$$

Also recall the *top-k* eigenbasis $\boldsymbol{U}_{top}$:

$$\boldsymbol{H} = \left(\boldsymbol{U}_{top} \;\vdots\; \boldsymbol{U}_{bulk} \right) \left(\begin{array}{c:c} \boldsymbol{D}_{top} & \\ \hdashline & \boldsymbol{D}_{bulk} \approx 0 \end{array} \right) \left(\begin{array}{c} \boldsymbol{U}_{top}^\top \\ \hdashline \boldsymbol{U}_{bulk}^\top \end{array} \right)$$

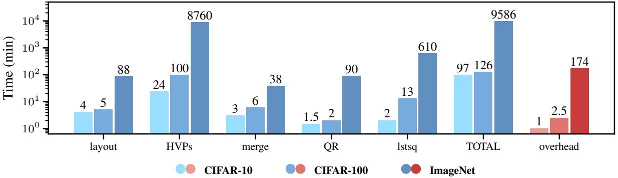
- We have **same-shape, orthogonal matrices** $\boldsymbol{I}_{D,k}$ and $\boldsymbol{U}_{top}$
- **Grassmannian metrics** measure the distance between their **spaces**
- Theoretical and empirical analysis of several Grassmannian metrics
- The **overlap** metric is stable and has a **random baseline value** of $\frac{k}{D}$:

$$\frac{1}{k} \|\boldsymbol{I}_{D,k}^\top \boldsymbol{U}_{top}\|_F^2 \in [0, 1] \quad (\text{higher} \Longleftrightarrow \text{more similar})$$



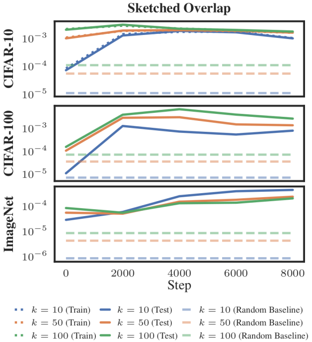
Computing the EIGH via sketched methods

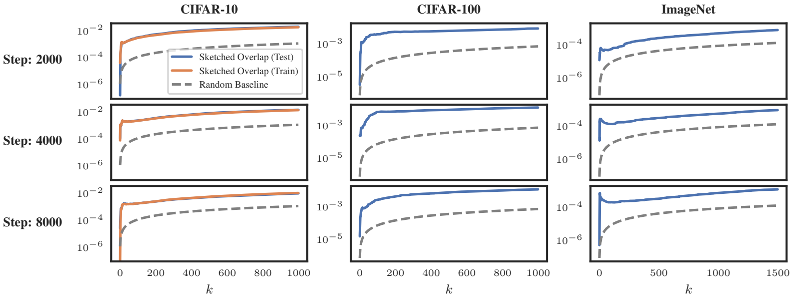
- Computing *overlap* **requires top-k Hessian eigendecomposition** $\frac{1}{k}||\boldsymbol{I}_{D,k}^{\top}\boldsymbol{U}_{top}||_F^2$
- **Intractable:** $\mathcal{O}(D^2)$ memory, $\mathcal{O}(D^3)$ arithmetic (Golub et al. 2013)
- **Expensive measurements:** Each $\boldsymbol{w} = \boldsymbol{H}\boldsymbol{v}$ costs 2 forward+backpropagations (Pearlmutter 1994)
- Sketched methods: $\mathcal{O}(k)$ *parallel* measurements, $\mathcal{O}(Dk)$ memory (Halko et al. 2011)
- **PyTorch library**: `skerch`



`pip install skerch`

High overlap in DL





- **Scalable**: Rank-1500 eigendecompositions on 12M-parameter networks
- **Orders of magnitude higher for all observed splits, steps, rank sizes and problems**
- Parameter inspection cheaply informs about curvature $\rightarrow$ training, pruning, UQ, analysis
- Still, spaces are far from identical ($\frac{k}{D}$ is small), so no direct mapping

Thank you!

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Conclusions:

- Grassmannian metrics to compare arbitrary parameters and Hessian eigenspaces
- Sketched eigendecompositions to measure *overlap* at scale $\rightarrow$ `skerch`
- DL *overlap* orders-of-magnitude larger than baseline (albeit far from identical)
- Connecting expensive Hessian quantities with cheap parameter observations

Future work:

- Scalability: We also explore faster alternatives like perturbation-based and GGN
- Explaining why do we observe high overlap
- Leveraging this effect in downstream applications

References I

Blalock, Davis et al. (2020). "What is the State of Neural Network Pruning?" In: *Proceedings of Machine Learning and Systems (MLSys)*.

Frankle, Jonathan and Michael Carbin (2019). "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rJl-b3RcF7.

Golub, Gene H. and Charles F. Van Loan (2013). *Matrix Computations*. Fourth. The Johns Hopkins University Press.

Gur-Ari, Guy, Daniel A. Roberts, and Ethan Dyer (2019). *Gradient Descent Happens in a Tiny Subspace*. URL: https://openreview.net/forum?id=ByeTHsAqtX.

Halko, N., P. G. Martinsson, and J. A. Tropp (2011). "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions". In.

Pearlmutter, Barak A. (1994). "Fast Exact Multiplication by the Hessian". In.

Sagun, Levent et al. (2018). *Empirical Analysis of the Hessian of Over-Parametrized Neural Networks*. ICLR 2018 Workshop. URL: https://openreview.net/forum?id=rJrTwxbCb.

You, Haoran et al. (2020). "Drawing Early-Bird Tickets: Toward More Efficient Training of Deep Networks". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=BJxsrgStvr.