UNDERSTANDING THE SUBSPACE STABILIZATION OF THE HESSIAN AND GRADIENT COVARIANCE MATRIX

Anonymous authors
Paper under double-blind review

ABSTRACT

The phenomenon of the top subspace stabilization of the Hessian matrix is an surprising and critical aspect in study of the second-order information of neural network training. Prior work argues that the top subspace of the Hessian stabilizes by measuring the overlap between the top subspaces of the step-wise Hessian, and explains this stabilization with diminishing parameter change in the late phase of training. In this paper, we define a new instability metric for the subspace evolution, and use it to detect subspace stabilization that is independent of the magnitude of parameter change. In the meantime, we observe that the gradient covariance matrix has a similar property of its top subspace to the Hessian. By using a between-class and within-class decomposition of the gradient covariance matrix, we identify an explicit form that gives a near-perfect approximation of the top-$(C-1)$ subspace of the Hessian and the gradient covariance matrix. In the gradient flow set-up, we show that the slow evolution of the idenfied approximation is due to the separation between the outlier and the bulk eigenvalues of the Hessian matrix, thus providing an explanation to the phenomenon of the top subspace stabilization of the Hessian matrix.

1 INTRODUCTION

The loss landscape of neural network training is believed to be highly nonconvex (Li et al., 2018; Swirszcz et al., 2017; Goldblum et al., 2020), posing great challenge to the optimization process that is central to modern deep learning. Yet, empirical observations do not align well with this belief: training neural networks starting from random initialization often achieves consistent loss decrease (Choromanska et al., 2015; Keskar et al., 2017), especially in large models (Kaplan et al., 2020; Hoffmann et al., 2022). Diving deeper into the convergence behavior, Schaipp et al. (2025) demonstrates that the behavior of learning rate schedules resembles the theory developed from convex non-smooth optimization. This near-convex behavior is further justified by the follow-up work from the perspective of convex dominance (Bu et al., 2026). The various empirical success implies that neural network training may not be as nonconvex as believed it to be.

From the perspective of the Hessian matrix, the near-convex behavior is justified by the low-rankness of the Hessian (Sagun et al., 2017; Gur-Ari et al., 2018; Papyan, 2019), where the Hessian's eigenvalue spectrum is shown to have a few positive outliers with the remaining having a small magnitude. However, the eigenvalue spectrum does not explain all the behaviors in the optimization of deep learning loss. To understand the acceleration of momentum-based optimizer for training neural networks, Liao & Kyrillidis (2024) proposed the subspace partial-convexity, which requires the top subspace to remain nearly non-changing. Wen et al. (2025) based its analysis on the condition that the "river" direction in the river-valley structure of the loss landscapes changes slowly to derive theoretical results of the WSD schedule consistent with the empirical phenomenon. With a even simpler setting, Qiu et al. (2025) derives a predictive scaling dynamic based on the quadratic model, and Meterez et al. (2026) show that training neural networks based on its local quadratic approximation preserves the performance for up to 10% of the training period. Together, these results suggest that the stabilization of at least the top subspace of the Hessian, as demonstrated by Gur-Ari et al. (2018), is crucial to the understanding of neural network training.

Nevertheless, despite the empirical observation in Gur-Ari et al. (2018), it is not clear whether the behavior is consistent in different training setups, and why the top subspace stabilization of the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Hessian happens. In particular, Gur-Ari et al. (2018) measures the subspace overlap between consecutive checkpoint as the cosine angle between the subspaces. Despite being a valid metric of subspace change, this quantity is sensitive to the magnitude of the parameter change between checkpoints: if optimizers does not introduce a large enough parameter update due to reasons such as small learning rate, then the top Hessian subspaces between two iterates are inherently small. This reasoning is adopted by Jaiswal et al. (2025) to theoretically justify the stabilization of the top Hessian subspace when the training loss converges. However, it remains an open question whether the Hessian top subspace stabilization happens independent of the magnitude of the parameter update, and, if so, what is the reason behind it.

To answer the question of *whether the top Hessian subspace really stabilizes*, we introduce a metric that scales the subspace change with the inverse of the parameter update magnitude to more accurately reflect the Lipschitzness of the subspace movement. Using the metric, we demonstrate empirically that the standard training settings in Song et al. (2025); Cohen et al. (2021) indeed exhibit top Hessian subspace stabilization independent of the update magnitude. However, we also show that, by modifying the architecture of the neural network, our metric identifies the ablation of the stabilization that is not observable using the step-wise subspace overlap metric.

To answer the question of *why the top Hessian subspace stabilizes*, we take a detour to first focus on the mini-batch gradient covariance matrix, which is shown to highly correlate with the Hessian matrix (Zhu et al., 2019). We observe that the gradient covariance matrix exhibit low-rank property similar to the Hessian, whose top subspace also stabilizes beyond the initial stage of training. In addition, we observe a surprising alignment between the top subspace of the Hessian and the gradient covariance matrix, despite that the two matrices are different. Focusing on the multi-class classification task, we adopt a similar approach to the with-class and between-class decomposition of the Hessian subspace in Papyan (2020) to show that the top subspace of both the Hessian and the gradient covariance matrix can be described by the span of the centered per-class gradient. With this explicit identifier, we adopt matrix ODE tools to show that, under gradient flow training, the subspace change is small when the largest eigenvalue of the Hessian and per-class Hessian in the *bulk* subspace is small. Together, our discovery makes the following contribution:

- We use the overlap and instability metric to detect settings where the top Hessian subspace stabilizes independent of the parameter update magnitude.
- We show that the mini-batch gradient covariance possesses similar low-rankness and top subspace stabilization property. In addition, we show an alignment between the Hessian's and gradient covariance matrix's top subspace.
- We give an explicit identifier for the top subspace of the Hessian and the gradient covariance matrix. Using this identifier, we give a theoretical account of the top subspace stabilization phenomenon that depend on the separation between the top and bulk eigenvalues.

Notations. We use small regular letters (e.g. a) to denote scalars, small bold-face letters (e.g. $\boldsymbol{a}$) to denote vectors, and capital bold-face letters (e.g. $\boldsymbol{A}$) to denote matrices. For a matrix $\boldsymbol{M}$, we use $\sigma_i(\boldsymbol{M})$ to denote its ith singular value, $\sigma_{\min}(\boldsymbol{M})$ to denote its minimum singular value, and $\|\boldsymbol{M}\|_2, \|\boldsymbol{M}\|_F$ to denote its operator and Frobenius norm, respectively. For a square matrix $\boldsymbol{M}$, we use $\lambda_i(\boldsymbol{M})$ to denote its ith eigenvalue and $\lambda_{\min}(\boldsymbol{M})$ to denote its smallest eigenvalue. For a symmetric matrix $\boldsymbol{M}$, we use $\mathrm{span}(\boldsymbol{M})$ to denote the row/column space of $\boldsymbol{M}$.

2 Does Hessian Top Subspace Really Stabilize?

We are interested in training a neural network $f(\boldsymbol{\theta}, \boldsymbol{x})$ with parameter $\boldsymbol{\theta} \in \mathbb{R}^p$ operating on inputs $\boldsymbol{x} \in \mathbb{R}^d$. Let the training data be $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Our focus is on the C-class classification task with $y_i \in [C]$ denote the labels, and the neural network $f(\boldsymbol{\theta}, \boldsymbol{x})$ trained over the cross-entropy loss

$$\mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(f(\boldsymbol{\theta}, \boldsymbol{x}_i), y_i\right); \quad \ell(\hat{\boldsymbol{y}}, c) = -\log \hat{\boldsymbol{y}}_c$$

Here $f(\boldsymbol{\theta}, \boldsymbol{x})$ outputs logits. For the convenience of the study, we use $\boldsymbol{p}(\boldsymbol{\theta}, \boldsymbol{x}) = \mathtt{softmax}(f(\boldsymbol{\theta}, \boldsymbol{x}))$ to denote the softmax probability. Denote the neural network gradient and Hessian as $\boldsymbol{g}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ and $\boldsymbol{H} = \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$ respectively. Let the eigen-decomposition of

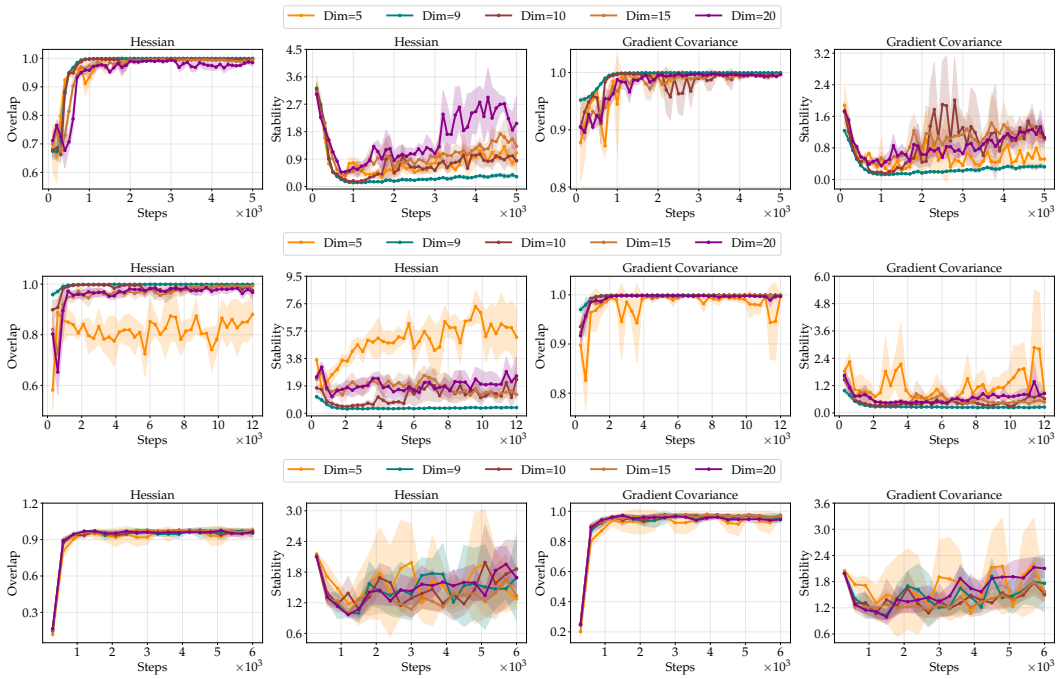


Figure 1: Top-K subspace stabilization of the Hessian and the gradient covariance matrix, with $K \in \{5, 9, 10, 15, 20\}$ shown in different colors. Column 1 and 3 shows the overlap metric, and column 2 and 4 shows the stability metric. **First row:** MLP on MNIST. **Second row:** 2-Layer CNN on CIFAR-10.**Third row:** 3-Layer CNN on CIFAR-10.

$\boldsymbol{H}(\boldsymbol{\theta})$ be $\boldsymbol{H} = \boldsymbol{U}(\boldsymbol{\theta})\boldsymbol{\Lambda}(\boldsymbol{\theta})\boldsymbol{U}(\boldsymbol{\theta})^\top$ with $\boldsymbol{\Lambda}(\boldsymbol{\theta})_{i,i} \geq \boldsymbol{\Lambda}(\boldsymbol{\theta})_{i+1,i+1}$, i.e., the eigenvalues are sorted in descending order. We define the top-K subspace of $\boldsymbol{H}(\boldsymbol{\theta})$ as $\boldsymbol{U}_K(\boldsymbol{\theta}) \in \mathbb{R}^{p\times K}$ containing the first K columns of $\boldsymbol{U}(\boldsymbol{\theta})$ for some $K \leq p$. Given two subspaces $\boldsymbol{U}_K$ ad $\boldsymbol{U}'_K$, Gur-Ari et al. (2018) defines the overlap metric as

$$\text{Overlap}(\boldsymbol{U}_K, \boldsymbol{U}'_K) = \frac{\text{Tr}\left(\boldsymbol{U}_K\boldsymbol{U}_K^\top\boldsymbol{U}'_K\boldsymbol{U}'^\top_K\right)}{\sqrt{\text{Tr}\left(\boldsymbol{U}_K\boldsymbol{U}_K^\top\right)\text{Tr}\left(\boldsymbol{U}'_K\boldsymbol{U}'^\top_K\right)}} = \frac{1}{K}\left\|\boldsymbol{U}_K^\top\boldsymbol{U}'_K\right\|_F^2 \tag{1}$$

However, we note that this metric is sensitive to the choice of the learning rate. Let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ be two consecutive checkpoints during model training. Assuming a universal upper bound on the third-order derivative in the region of the training $\left\|\nabla^3_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\right\|_2 \leq L$, we can obtain that

$$\begin{aligned}\text{Overlap}(\boldsymbol{U}_K(\boldsymbol{\theta}), \boldsymbol{U}_K(\boldsymbol{\theta}')) &= 1 - \frac{1}{2K}\left\|\boldsymbol{U}_K(\boldsymbol{\theta})\boldsymbol{U}_K(\boldsymbol{\theta})^\top - \boldsymbol{U}_K(\boldsymbol{\theta}')\boldsymbol{U}_K(\boldsymbol{\theta}')^\top\right\|_F^2 \\ &\geq 1 - \frac{\gamma}{2K}\left\|\boldsymbol{H}(\boldsymbol{\theta}) - \boldsymbol{H}(\boldsymbol{\theta}')\right\|_F^2 \geq 1 - \frac{\gamma L^2}{2K}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2^2\end{aligned} \tag{2}$$

where the first inequality is due to Davis-Kahan $\sin\Theta$ Theorem , and γ is a quantity that depends on the eigengap of $\boldsymbol{H}(\boldsymbol{\theta})$. Since for two consecutive checkpoints we have $\left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2 \propto \eta$, the overlap metric in (2) can be made arbitrarily close to 1 by choosing a small enough learning rate. To mitigate the confounding factor of the small learning rate, we define the following instability metric that quantifies the sensitivity to perturbation between two subspaces

$$\begin{aligned}\text{Instability}(\boldsymbol{U}_K(\boldsymbol{\theta}), \boldsymbol{U}_K(\boldsymbol{\theta}')) &:= \left(1 - \frac{1}{K}\left\|\boldsymbol{U}_K(\boldsymbol{\theta})^\top\boldsymbol{U}_K(\boldsymbol{\theta}')\right\|_F^2\right)^{\frac{1}{2}} / \left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2 \\ &= \frac{\left\|\boldsymbol{U}_K(\boldsymbol{\theta})\boldsymbol{U}_K(\boldsymbol{\theta})^\top - \boldsymbol{U}_K(\boldsymbol{\theta}')\boldsymbol{U}_K(\boldsymbol{\theta}')^\top\right\|_F}{\sqrt{2K}\left\|\boldsymbol{\theta} - \boldsymbol{\theta}'\right\|_2}\end{aligned} \tag{3}$$

In particular, our instability metric can be interpreted as the Lipschitzness of the projection operator $\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{U}_K(\boldsymbol{\theta})\boldsymbol{U}_K(\boldsymbol{\theta})^\top$ scaled by a factor of $\frac{1}{\sqrt{2K}}$. We note that although the instability metric is not confounded by the small learning rate, it can be made arbitrarily small by

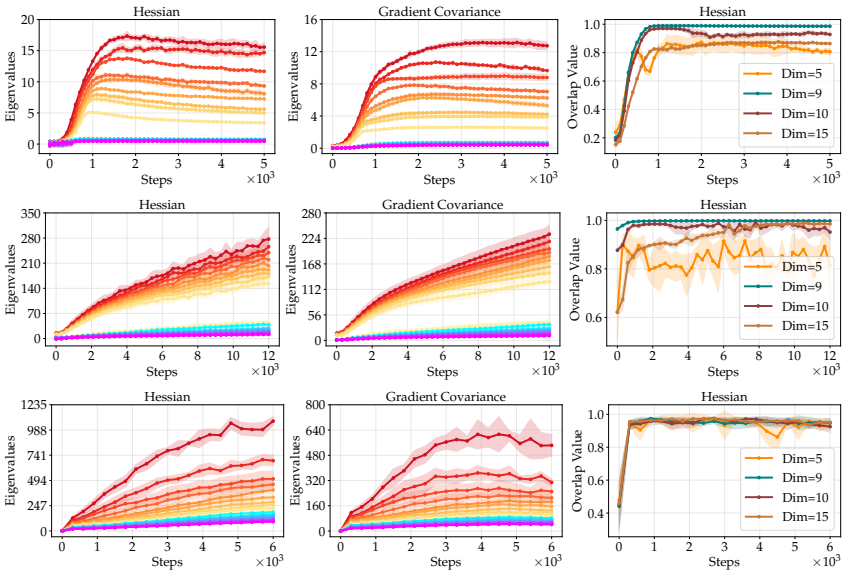


Figure 2: Structural similarity between the Hessian and the gradient covariance matrix. The three columns plots the Hessian eigenvalues, gradient covariance matrix's eigenvalues, and the overlaop between the Hessian and the gradient covariance matrix's top-K subspaces. **First row:** MLP on MNIST. **Second row:** 2-Layer CNN on CIFAR-10.**Third row:** 3-Layer CNN on CIFAR-10.

a parameter change that grows in magnitude, since $\left(1-\frac{1}{K}\left\|\boldsymbol{U}_K\left(\boldsymbol{\theta}\right)^\top \boldsymbol{U}_K\left(\boldsymbol{\theta}'\right)\right\|_F^2\right)^{\frac{1}{2}} \leq 1$ and thus $\text{Stability}\left(\boldsymbol{U}_K\left(\boldsymbol{\theta}\right), \boldsymbol{U}_K\left(\boldsymbol{\theta}'\right)\right) \leq \left\|\boldsymbol{\theta}-\boldsymbol{\theta}'\right\|_2^{-1}$. In our experiment, we rule out this possibility by plotting both $\text{Overlap}\left(\boldsymbol{U}_K\left(\boldsymbol{\theta}\right), \boldsymbol{U}_K\left(\boldsymbol{\theta}'\right)\right)$ and $\text{Stability}\left(\boldsymbol{U}_K\left(\boldsymbol{\theta}\right), \boldsymbol{U}_K\left(\boldsymbol{\theta}'\right)\right)$: if we observe an $\text{Overlap}\left(\boldsymbol{U}_K\left(\boldsymbol{\theta}\right), \boldsymbol{U}_K\left(\boldsymbol{\theta}'\right)\right)$ close to 1 and a small enough $\text{Stability}\left(\boldsymbol{U}_K\left(\boldsymbol{\theta}\right), \boldsymbol{U}_K\left(\boldsymbol{\theta}'\right)\right)$, then we can safely say that the top-K subspace is truly stable.

With the overlap and instability metric in (2) and (3), we test the subspace stabilization in three settings: i). a two-layer MLP trained on MNIST dataset; ii). a two-layer CNN trained on CIFAR-10 dataset; and iii). a three-layer CNN with batch normalization trained on CIFAR-10 dataset.[1] Focusing on the first two columns of Figure 1 that shows the stabilization property of the Hessian, we observe that in the MLP and 2-layer CNN case, pretty much every K in the top-K overlap (except for $K = 5$ in the 2-layer CNN settings) achieves a near-perfect overlap. However, only the $K = C-1 = 9$ curve shows a consistently low stability metric. This demonstrates the insufficiently of using the overlap metric alone, which can be complemented by the instability metric we propose. Combining the overlap and instability metric, we observe that in the only the $K = C-1 = 9$ subspace (in teal) demonstrates true stability. This is possibly due to that reducing the dimensionality below $C-1$ fails to account for the rotation inside the top subspace, while increasing the dimensionality above $C-1$ introduces unstable directions. In the third column, our instability metric also discovers a setting where the top subspace fails to reach stabilization, even though the step-wise overlap between consecutive subspace remains close to 1. We will investigate this setting further in later sections.

3 Structure and Stabilization of Gradient Covariance Matrix

Before diving into the reasons of the stabilization of top Hessian subspace, we take a detour to explore the stabilization of a related matrix, the gradient covariance matrix. The relationship between the Hessian and the gradient covariance matrix is noted in Zhu et al. (2019), where it is observed that for classification task, the gradient covariance matrix can be written as the in terms of the empirical Fisher information matrix (FIM), which resembles the Gauss-Newton matrix in the expansion of the Hessian. In particular, the Hessian $\boldsymbol{H}\left(\boldsymbol{\theta}\right)$ can be written in the follow decomposition

[1]The first two settings follow the standard setting in Song et al. (2025); Cohen et al. (2021). Details of the experiments are provided in Appendix A.

$$\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{N}(\boldsymbol{\theta}) + \boldsymbol{S}(\boldsymbol{\theta}); \quad \boldsymbol{S}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}\frac{\partial \ell\left(f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right), y_i\right)}{\partial f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)_c}\nabla^2_{\boldsymbol{\theta}} f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)_c$$
$$\boldsymbol{N}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n}\sum_{c_1,c_2=1}^{C}\frac{\partial^2 \ell\left(f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right), y_i\right)}{\partial f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)_{c_1} f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)_{c_2}}\nabla_{\boldsymbol{\theta}} f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)_{c_1}\nabla_{\boldsymbol{\theta}} f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)^\top_{c_2} \tag{4}$$

For classification task with cross-entropy loss, the matrix $\boldsymbol{N}(\boldsymbol{\theta})$ is the same as the FIM

$$\boldsymbol{F}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{y\sim \boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x})}\left[\nabla_{\boldsymbol{\theta}}\log \boldsymbol{p}_{\boldsymbol{\theta}}\left(y \mid \boldsymbol{x}_i\right)\nabla_{\boldsymbol{\theta}}\log \boldsymbol{p}_{\boldsymbol{\theta}}\left(y \mid \boldsymbol{x}_i\right)^\top\right] \tag{5}$$

In particular, it has been observed in Papyan (2020) that the top subspace of $\boldsymbol{H}(\boldsymbol{\theta})$ attributes to $\boldsymbol{N}(\boldsymbol{\theta})$ (or $\boldsymbol{F}(\boldsymbol{\theta})$. The gradient covariance matrix, on the other hand, is often associated with the *empirical* Fisher information matrix $\tilde{\boldsymbol{F}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n}\hat{\boldsymbol{g}}_i(\boldsymbol{\theta})\hat{\boldsymbol{g}}_i(\boldsymbol{\theta})^\top$[2]. Let $\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\ell\left(f\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right), y_i\right)$ denote the per-sample gradient. Define the gradient covariance matrix

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)^\top = \tilde{\boldsymbol{F}}(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})^\top$$

Here $\tilde{\boldsymbol{F}}(\boldsymbol{\theta})$ can be seen as $\boldsymbol{F}(\boldsymbol{\theta})$ in (5) with deterministic ground-truth labels y. The overlap between $\boldsymbol{F}(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{F}}(\boldsymbol{\theta})$ explains the overlap between $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{H}(\boldsymbol{\theta})$ observed in Zhu et al. (2019). However, Kunstner et al. (2019) argues that the empirical FIM $\tilde{\boldsymbol{F}}(\boldsymbol{\theta})$ does not always capture properties of $\boldsymbol{H}(\boldsymbol{\theta})$, leaving the question of how much information of $\boldsymbol{H}(\boldsymbol{\theta})$ resides in $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ open.

We offer a new finding of the relationship between the Hessian and the gradient covariance matrix from the perspective of their *top subspace*, which has largely been ignored by previous work. From the first two columns of Figure 2, we observe that the gradient covariance matrix has a similar eigenvalue profile to the Hessian matrix, with a clear separation of the top-$(C-1)$ outlier and the bulk in the first two rows. In the third column, we measure the overlap between the top-K subspace of the Hessian and the gradient covariance matrix and the Hessian, and observes a consistent near-perfect overlap for the case of $K = C-1 = 9$. Moreover, in column 3 and 4 in Figure 1, we observe that the top-K subspace of the gradient covariance matrix has a similar stability phenomenon to the top subspace of the Hessian with the most significant stability happens at $K = C-1$. This suggest that the top-K subspace of the gradient covariance matrix and the Hessian shares a similar.

3.1 Identifying the Stable Subspace of the Gradient Covariance

To understand what directs the evolution of the top subspace of the gradient covariance matrix, we take a similar decomposition approach as in Papyan (2020) for the Hessian. In particular, we decompose $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ into class-specific gradient covariance matrices $\boldsymbol{\Sigma}_c(\boldsymbol{\theta})$. Recall that C denotes the number of classes. Let $\mathcal{I}_c \subseteq [n]$ denote the indices of the samples belonging to class c for $c \in [C]$. Let $n_c = |\mathcal{I}_c|$ denote the number of samples belonging to class c. Denote the class-aggregated gradient $\boldsymbol{g}_c(\boldsymbol{\theta})$ as $\boldsymbol{g}_c(\boldsymbol{\theta}) = \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\hat{\boldsymbol{g}}_i(\boldsymbol{\theta})$ so that $\boldsymbol{g}(\boldsymbol{\theta}) = \sum_{c=1}^{C}\frac{n_c}{n}\boldsymbol{g}_c(\boldsymbol{\theta})$. We define $\boldsymbol{\Sigma}_c(\boldsymbol{\theta})$ as

$$\begin{aligned}\boldsymbol{\Sigma}_c(\boldsymbol{\theta}) &= \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)^\top \\ &= \left(\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)\left(\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)^\top + \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right)\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right)^\top\end{aligned} \tag{6}$$

Intuitively, $\boldsymbol{\Sigma}_c(\boldsymbol{\theta})$ is the per-class gradient covariance matrix. Then we have that

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \sum_{c=1}^{C}\frac{n_c}{n}\boldsymbol{\Sigma}_c(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_{\text{top}}(\boldsymbol{\theta}) + \boldsymbol{\Sigma}_{\text{bulk}}(\boldsymbol{\theta})$$
$$\boldsymbol{\Sigma}_{\text{top}}(\boldsymbol{\theta}) := \sum_{c=1}^{C}\frac{n_c}{n}\left(\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)\left(\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)^\top$$
$$\boldsymbol{\Sigma}_{\text{bulk}}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{c=1}^{C}\sum_{i\in\mathcal{I}_c}\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right)\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right)^\top$$

[2]Here we are restricting the definition to the case of classification task with cross-entropy loss.

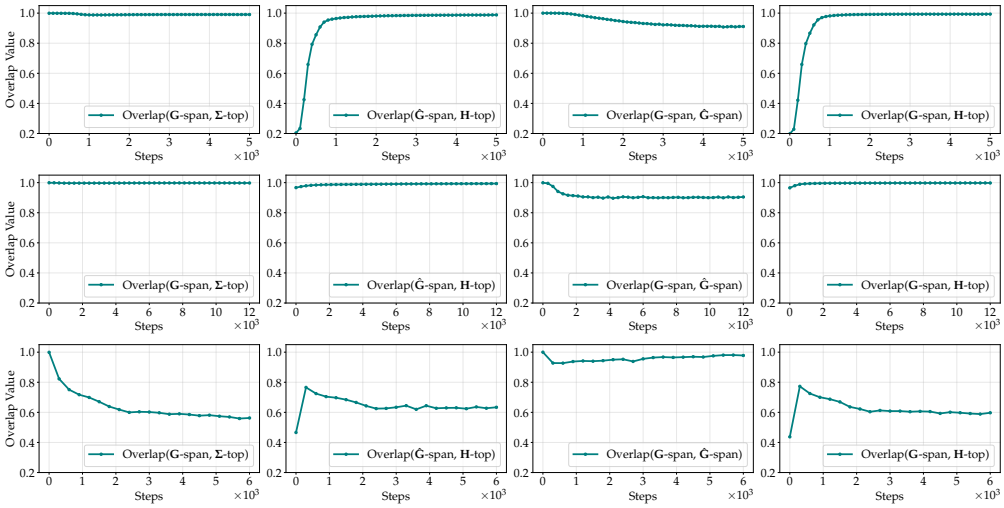


Figure 3: Overlaps between $\boldsymbol{G}$-span, $\hat{\boldsymbol{G}}$-span, and the top-$(C-1)$ subspace of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{H}(\boldsymbol{\theta})$ throughout training. **First row:** MLP on MNIST. **Second row:** 2-Layer CNN on CIFAR-10.**Third row:** 3-Layer CNN on CIFAR-10.

The decomposition of the $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ into $\boldsymbol{\Sigma}_{\text{top}}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_{\text{top}}(\boldsymbol{\theta})$ can be seen as separating the covariance into a between-class covariance $\boldsymbol{\Sigma}_{\text{top}}(\boldsymbol{\theta})$ and an average within class covariance $\boldsymbol{\Sigma}_{\text{bulk}}(\boldsymbol{\theta})$. Our hypothesis is that the between-class covariance $\boldsymbol{\Sigma}_{\text{top}}(\boldsymbol{\theta})$ explains the top-$(C-1)$ subspace of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ in the setting where the top subspace stabilizes.

To test the hypothesis, we notice that the column/row space of $\boldsymbol{\Sigma}_{\text{top}}(\boldsymbol{\theta})$ is spanned by vectors $\{\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\}_{c=1}^{C}$. Since $\sum_{c=1}^{C} \frac{n_c}{n}(\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})) = 0$, the same subspace must also be the span of $\{\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\}_{c=1}^{C-1}$. Let $\boldsymbol{G}(\boldsymbol{\theta}) \in \mathbb{R}^{(C-1)\times p}$ be the matrix such that the cth row is $\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})$. In the first column of Figure 3, we show that in the setting where the top-$(C-1)$ subspace of the Hessian and the gradient covariance matrix stabilizes (MLP and 2-layer CNN), the overlap between row space of $\boldsymbol{G}(\boldsymbol{\theta})$ and the top-$(C-1)$ subspace of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ stays very close to 1 throughout training. When the vectors $\{\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\}_{c=1}^{C-1}$ are linearly independent, $\boldsymbol{\Sigma}_{\text{top}}(\boldsymbol{\theta})$ is exactly rank-$(C-1)$. Therefore, the top-$(C-1)$ subspace of $\boldsymbol{\Sigma}$ is explicitly identified by span$(\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}), \ldots \boldsymbol{g}_C(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}))$, instead of covered by it. Thus, showing the stabilization of the top subspace of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ directly corresponding to upper bounding the change of span$(\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}), \ldots \boldsymbol{g}_C(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}))$.

3.2 Dynamic of Identified Subspace

Next, we analyze how fast the subspace span$(\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}), \ldots \boldsymbol{g}_C(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}))$ changes throughout training. We assume that at least $C-1$ vectors from $\{\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\}_{c=1}^{C}$ are linearly independent, which implies that span$(\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}), \ldots \boldsymbol{g}_C(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}))$ is a $(C-1)$-dimensional space. Indeed, this is easy to achieve in overparameterized model with $p \gg C$. To formally understand the change speed of the subspace, we study the projection matrix onto the row space of $\boldsymbol{g}$, defined by

$$\boldsymbol{P}(\boldsymbol{\theta}) = \boldsymbol{G}(\boldsymbol{\theta})^{\top}\left(\boldsymbol{G}(\boldsymbol{\theta})\boldsymbol{G}(\boldsymbol{\theta})^{\top}\right)^{-1}\boldsymbol{G}(\boldsymbol{\theta}) \in \mathbb{R}^{p\times p} \tag{7}$$

Consider the continuous dynamic of training with gradient flow

$$\dot{\boldsymbol{\theta}}(t) = -\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = -\boldsymbol{g}(\boldsymbol{\theta}(t)) \tag{8}$$

We let $\boldsymbol{g}(t) = \boldsymbol{g}(\boldsymbol{\theta}(t))$ and correspondingly $\boldsymbol{P}(t) = \boldsymbol{P}(\boldsymbol{\theta}(t))$ and $\boldsymbol{G}(t) = \boldsymbol{G}(\boldsymbol{\theta}(t))$. Showing the stability of the subspace span$(\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}), \ldots \boldsymbol{g}_C(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}))$ translates to bounding $\left\|\dot{\boldsymbol{P}}(t)\right\|_F$.

Theorem 1. *Let $\boldsymbol{P}(t) = \boldsymbol{P}(\boldsymbol{\theta}(t))$ be defined in (7). Consider the parameter dynamic induced by gradient flow on the loss $\mathcal{L}(\boldsymbol{\theta})$, as given in (8). Then we have that*

$$\left\|\dot{\boldsymbol{P}}(t)\right\|_F^2 \leq \frac{4}{\sigma_{\min}(\boldsymbol{G}(t))^2}\left((C-1)\left\|(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}(t)\boldsymbol{g}(t)\right\|_2^2 + \sum_{c=1}^{C-1}\left\|(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}_c(t)\boldsymbol{g}(t)\right\|_2^2\right)$$

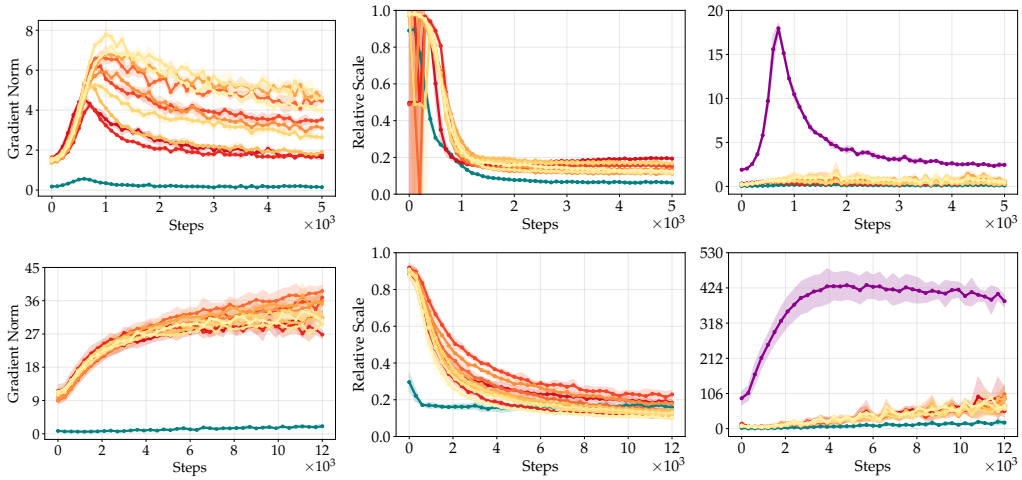


Figure 4: **Left column:** norm comparisons between $\boldsymbol{g}_c(\boldsymbol{\theta})$s (yellow to red for $c \in [C]$) and $\boldsymbol{g}(\boldsymbol{\theta})$ (teal). **Mid column:** $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}_c\|_2 / \|\boldsymbol{H}_c\|_2$ for $c \in [C]$ (yellow to red), and $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}\|_2 / \|\boldsymbol{H}\|_2$ (teal). **Right column:** comparison of $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}_c\boldsymbol{g}\|_2$ (yellow to red) and $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}\boldsymbol{g}\|_2$ (teal) against $\sigma_{\min}(\boldsymbol{G}(t))$ (purple). The first and second row correspond to the setting of the first and second row in Figure 1.

where $\boldsymbol{H}(t) = \boldsymbol{H}(\boldsymbol{\theta}(t))$ and $\boldsymbol{H}_c(t)$ is the Hessian restricted to class-c samples at $\boldsymbol{\theta}(t)$

$$\boldsymbol{H}_c(t) := \frac{1}{n_c}\sum_{i=1}^{n_c}\nabla^2_{\boldsymbol{\theta}}\ell\left(f\left(\boldsymbol{\theta},\boldsymbol{x}_{i,c}\right),y_c\right),\ \textit{s.t.}\ \boldsymbol{H}(t) = \sum_{c=1}^{C}\frac{n_c}{n}\boldsymbol{H}_c(t)$$

Theorem 1 states that the top subspace change scales with the norm of the projected Hessian-gradient product $(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}(t)\boldsymbol{g}(t)$ and $(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}_c(t)\boldsymbol{g}(t)$. As a naive simplification, we have that $\|(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}(t)\boldsymbol{g}(t)\|_2 \propto \|(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}(t)\|_2$ and $\|(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}_c(t)\boldsymbol{g}(t)\|_2 \propto \|(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}_c(t)\|_2$. In the theorem below, we show that $\|(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}_c(t)\|_2$ can we well controlled by $\|(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}(t)(\boldsymbol{I}-\boldsymbol{P}(t))\|_2$.

Theorem 2. *Fix any $t \geq 0$ and let $\boldsymbol{H}(t) = \boldsymbol{H}, \boldsymbol{H}_c(t) = \boldsymbol{H}_c, \boldsymbol{P}(t) = \boldsymbol{P}$. Assume that $\lambda_{\min}(\boldsymbol{H}_c) \geq -\nu$ for all $c \in [C]$. Let $\varepsilon \geq \|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}(\boldsymbol{I}-\boldsymbol{P})\|_2$. Then we have that $\lambda_{\min}(\boldsymbol{H}) \geq -\nu$, and*

$$\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}\|_2 \leq 2\sqrt{(\varepsilon+\nu)(\|\boldsymbol{H}\|_2+\nu)};\quad \|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}_c\|_2 \leq \frac{2n}{n_c}\sqrt{(\varepsilon+\nu)(\|\boldsymbol{H}\|_2+\nu)}$$

Theorem 2 relates the operator norm of $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}$ and $(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}_c$ to ε, the upper bound on the Hessian's operator norm when projected onto $\boldsymbol{I}-\boldsymbol{P}$, which roughly corresponds to the operator norm of the Hessian in the bulk subspace. Recall that $\boldsymbol{P}(t)$ predicts the top-$(C-1)$ subspace of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, which coincides with the top-$(C-1)$ subspace of the Hessian. Therefore, if $\boldsymbol{H}_c(t)$ shares a simlar subspace to $\boldsymbol{H}(t)$, then $(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}(t)$ and $(\boldsymbol{I}-\boldsymbol{P}(t))\boldsymbol{H}_c(t)$ should correspond roughly to the bulk subspace of the Hessian, which is significantly smaller than the top-$(C-1)$ eigenvalues of $\boldsymbol{H}(t)$. In the second column of Figure 4, we plot $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}_c\|_2 / \|\boldsymbol{H}_c\|_2$ for $c \in [C]$ in color spectrum yellow to red, and $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}\|_2 / \|\boldsymbol{H}\|_2$ in teal. We observe that these ratios are consistently small after the initial period of training.

In the meantime, we point out that simply measuring $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}\|_2$ may not result in a tight bound, since $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}\boldsymbol{g}\|_2$ and $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}_c\boldsymbol{g}\|_2$ also involves the alignment between the gradient and the Hessian outside the top subspace. Therefore, in the third column of Figure 4 we also directly measured the norms $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}_c\boldsymbol{g}\|_2$ in the color spectrum yellow to red and $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}\boldsymbol{g}\|_2$ in teal, and compare directly against $\sigma_{\min}(\boldsymbol{G}(t))$ in purple. We observe that $\sigma_{\min}(\boldsymbol{G}(t))$ is much larger in magnitude than $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}_c\boldsymbol{g}\|_2$ and $\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}\boldsymbol{g}\|_2$.

4 Stabilization of the Hessian Top Subspace

From the perspective of the gradient covariance matrix, Section 3 shows that i). the top-$(C-1)$ subspace of the gradient covariance matrix has a near-perfect alignment with the

top-$(C-1)$ subspace of the Hessian; ii). the top-$(C-1)$ subspace of the gradient covariance matrix can be identified by span $(\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}), \ldots \boldsymbol{g}_C(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}))$; and iii). the subspace span $(\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}), \ldots \boldsymbol{g}_C(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}))$ changes slowly when bulk subspace of the Hessian and the per-class Hessian are small. Therefore, it it reasonable to conclude that span $(\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}), \ldots \boldsymbol{g}_C(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}))$ serves also as an identifier for the top-$(C-1)$ subspace of the Hessian, and adopt the reason of its slow change as an explanation of the stabilization of the top-$(C-1)$ subspace of the Hessian. In this section, we provide an explanation of why span $(\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}), \ldots \boldsymbol{g}_C(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}))$ should be an identifier of the top-$(C-1)$ subspace of the Hessian based on the cross-class structure argument in Papyan (2020).

4.1 Connecting to the Subspace Identifier from Papyan (2020)

Papyan (2020)studies the outlier subspace using a classification setup similar to our case. To ease the explanation, we use a slightly different notation from previous sections. In particular, for the following discussion we fix $\boldsymbol{\theta}$ and omit the specification of e.g. $\boldsymbol{g}(\boldsymbol{\theta})$ to endorse the simplified notation of $\boldsymbol{g}$. Let $\boldsymbol{x}_{i,c}$ denote the ith input sample from cth class, and y_c be the indicator of the cth class. Define

$$\tilde{\boldsymbol{g}}_{i,c,c'} := \nabla_{\boldsymbol{\theta}} \ell \left(f\left(\boldsymbol{\theta}, \boldsymbol{x}_{i,c}\right), y_{c'} \right)$$

as the per-sample *cross-class* gradient, i.e., the gradient of $\boldsymbol{x}_{i,c}$ computed bsaed on a potentially incorrect label $y_{c'}$. Then, let $\omega_{i,c,c'} = \frac{p_{i,c,c'}}{n_c C}$ with $p_{i,c,c'} = \boldsymbol{p}\left(\boldsymbol{\theta}, \boldsymbol{x}_{i,c}\right)_{c'}$. Define auxiliary variables

$$\begin{aligned} \tilde{\boldsymbol{g}}_{c,c'} &:= \sum_{i \in \mathcal{I}_c} \pi_{i,c,c'} \tilde{\boldsymbol{g}}_{i,c,c'}; \quad \pi_{i,c,c'} = \frac{\omega_{i,c,c'}}{\omega_{c,c'}}; \quad \omega_{c,c'} = \sum_{i \in \mathcal{I}_c} \omega_{i,c,c'} \\ \tilde{\boldsymbol{g}}_{c} &:= \sum_{c \neq c'} \pi_{c,c'} \boldsymbol{g}_{c,c'}; \quad \pi_{c,c'} = \frac{\omega_{c,c'}}{\omega_{c}}; \quad \omega_{c} = \sum_{c \neq c'} \omega_{c,c'} \end{aligned} \tag{9}$$

Intuitively, $\tilde{\boldsymbol{g}}_{c,c'}$ and $\tilde{\boldsymbol{g}}_c$ can be seen as a gradual weighted aggregation first over samples of each class and over cross-classes. Papyan (2020) shows that the top-C eigenvector subspace of the Hessian attributes to the rank-C matrix $\hat{\boldsymbol{G}} := \sum_{c=1}^{C} \omega_c \tilde{\boldsymbol{g}}_c \tilde{\boldsymbol{g}}_c^\top$. Therefore, we should also expect the row/column span of $\hat{\boldsymbol{G}}$, namely span$(\hat{\boldsymbol{G}})$, to cover the top-$(C-1)$ subspace of the Hessian. In cases where the top-$(C-1)$ subspaces stabilizes (first and second row of Figure 3), it can be verified that $\hat{\boldsymbol{G}}$ covers the top-$(C-1)$ subspace of the Hessian. However, there are two issue with this attribution:

1. The directions $\boldsymbol{g}'_c$ involves complicated definition that is not easy to interpret and analyze.
2. The subspace span$(\hat{\boldsymbol{G}})$ has is a C-dimensional subspace. Although it has been shown that span$(\hat{\boldsymbol{G}})$ covers the subspace spanned by the top-$(C-1)$ eigenvectors of the Hessian, the dimensionality mismatch makes the characterization of the top Hessian subspace difficulty.

Connect to our identifier. Nevertheless, span$(\hat{\boldsymbol{G}})$ provides a good starting point to be connected to our identifier span $(\boldsymbol{g}_1(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}), \ldots \boldsymbol{g}_C(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta}))$. To see the connection, we start by simplifying the definition of $\boldsymbol{g}_c$ to an equivalent set of vectors $\boldsymbol{g}'_c$, justified by the theorem below.

Theorem 3. *Let $\hat{\boldsymbol{G}} := \sum_{c=1}^{C} \omega_c \tilde{\boldsymbol{g}}_c \tilde{\boldsymbol{g}}_c^\top$ with ω_c and $\tilde{\boldsymbol{g}}_c$ defined in (9). Let $\boldsymbol{g}'_c = \frac{1}{n_c} \sum_{i \in \mathcal{I}_c} \boldsymbol{p}\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)_c \hat{\boldsymbol{g}}_i\left(\boldsymbol{\theta}\right)$ be the aggregated gradients of samples from class-c, weighted by the softmax probability $\boldsymbol{p}\left(\boldsymbol{\theta}, \boldsymbol{x}_{i,c}\right)_c$ for $c \in [C]$. If $\{\boldsymbol{g}_c\}_{c=1}^{C}$ is an linearly independent list, then we have that span $(\boldsymbol{g}_{class}) =$ span $(\boldsymbol{g}'_1, \ldots, \boldsymbol{g}'_C)$.*

Theorem 3 simplified the row space of $\boldsymbol{g}_{\text{classs}}$ into the span of vectors $\boldsymbol{g}'_c$s, providing better interpretability of the subspace directions. Instead of relying on the notion of cross-class gradients, the definition of $\boldsymbol{g}'_c$ is simply a weighted average of the per-sample gradients in class c.

Comparing the form of $\boldsymbol{g}'_c$ to $\boldsymbol{g}_c$ in our setup, the difference is that our $\boldsymbol{g}_c$ is an unweighted average of the per-sample gradients. To bridge this gap, our second observation is that the weighting probabilities $\boldsymbol{p}\left(\boldsymbol{\theta}, \boldsymbol{x}_{i,c}\right)_c$ in $\boldsymbol{g}'_c$ can be dropped without affecting the direction of $\boldsymbol{g}'_c$ too much.

Theorem 4. *Let $\boldsymbol{U}, \boldsymbol{U}' \in \mathbb{R}^{d \times C}$ denote the matrices such that the columns of $\boldsymbol{U}$ and $\boldsymbol{U}'$ are orthonormal basis of span $(\boldsymbol{g}_1\left(\boldsymbol{\theta}\right), \ldots, \boldsymbol{g}_C\left(\boldsymbol{\theta}\right))$ and span $(\boldsymbol{g}'_1\left(\boldsymbol{\theta}\right), \ldots, \boldsymbol{g}'_C\left(\boldsymbol{\theta}\right))$, respectively. Let*

$\boldsymbol{G}' \in \mathbb{R}^{C\times p}$ *such that the cth row of* $\boldsymbol{G}'$ *is* $\boldsymbol{g}'_c$*. Then we have that*

$$\sigma_{\min}\left(\boldsymbol{U}^\top \boldsymbol{U}'\right)^2 \geq 1 - \frac{\|\boldsymbol{\Sigma}_{\textit{bulk}}(\boldsymbol{\theta})\|_2}{4\sigma_C(\boldsymbol{G}')^2}$$

In particular, Theorem 4 states that the overlap between the subspace spanned by $\{\boldsymbol{g}'_c\}_{c=1}^C$ and $\{\boldsymbol{g}_c\}_{c=1}^C$ is off by a magnitude that scales with $\|\boldsymbol{\Sigma}_{\text{bulk}}(\boldsymbol{\theta})\|_2$. By the form of the covariance matrix, $\|\boldsymbol{\Sigma}(\boldsymbol{\theta})\|_2$ scales as $\|\boldsymbol{g}_c - \boldsymbol{g}\|_2^2$. We observe that the full gradient $\boldsymbol{g}$ is usually much smaller in magnitude than the per-class gradient $\boldsymbol{g}_c$, as verified in the first column of Figure 4. In the meantime, $\boldsymbol{g}_c$'s can be considered as well separated, as justified by the neural collapse perspective Papyan et al. (2020); Zhu et al. (2021). This implies that $\sigma_C(\boldsymbol{G}')^2$ roughly has the same scale as $\|\boldsymbol{\Sigma}(\boldsymbol{\theta})\|_2$. Thus, a well-separated top-$(C-1)$ and the bulk subspace such that $\|\boldsymbol{\Sigma}_{\text{bulk}}(\boldsymbol{\theta})\|_2 \ll \|\boldsymbol{\Sigma}(\boldsymbol{\theta})\|_2$ implies a near-perfect overlap between the span of $\{\boldsymbol{g}'_c\}_{c=1}^C$ and $\{\boldsymbol{g}_c\}_{c=1}^C$.

To recover our identifier $\{\boldsymbol{g}_c - \boldsymbol{g}\}_{c=1}^C$, it remains to explain why subtracting the full gradient from $\boldsymbol{g}_c$ still covers the top-$(C-1)$ subspace of the Hessian. Since $\sum_c \frac{n_c}{n}\boldsymbol{g}_c\boldsymbol{g}_c^\top = \boldsymbol{\Sigma}_{\text{top}} + \boldsymbol{g}\boldsymbol{g}^\top$, centering the class gradients removes exactly one rank-one component, along the full gradient $\boldsymbol{g}$ and of magnitude $\|\boldsymbol{g}\|_2^2$. The left column of Figure 4 shows $\|\boldsymbol{g}\|_2$ to be much smaller than $\|\boldsymbol{g}_c\|_2$ throughout training, so this component is the weakest direction of the uncentered between-class moment rather than one of its top-$(C-1)$ outliers. Centering therefore discards precisely the surplus direction in the C-dimensional $\text{span}(\hat{\boldsymbol{G}})$ while preserving the outlier subspace it covers.

4.2 Accounting for the Reason When Top Hessian Subspace Fails to Stabilize

In the third row of Figure 1, we observed that the top subspace of the Hessian and the gradient covariance does not stabilize when we train a 3-layer CNN with batch normalization. The same setting also gives the experimental result in the third row of Figure 2 and Figure 3. Combining the results let us infer some reasons where the top subspace fails to stabilize in this setting.

Recall from our previous reasoning that when the top subspace stabilizes $i)$. the top subspace can be near-perfectly identified by $\text{span}(\boldsymbol{g}_1 - \boldsymbol{g}, \ldots, \boldsymbol{g}_C - \boldsymbol{g})$, and $ii)$. $\text{span}(\boldsymbol{g}_1 - \boldsymbol{g}, \ldots, \boldsymbol{g}_C - \boldsymbol{g})$ moves slowly because of the separate between the eigenvalues of the top and bulk subspace of the Hessian. From Figure 2 and Figure 3, both conditions seems to break: in the right-most figure of the third rown in Figure 2, we see that $\text{span}(\boldsymbol{g}_1 - \boldsymbol{g}, \ldots, \boldsymbol{g}_C - \boldsymbol{g})$ only achieves a 0.6 overlap with the top-$(C-1)$ subspace of the Hessian, leaving a significant part of the Hessian's top-$(C-1)$ subspace unexplained ans susceptible to chaotic forces during training. In the third row Figure 2, we observe that the clear separation between the top and bulk subspace in both the Hessian and the gradient covariance disappears, which lead to a faster change of $\text{span}(\boldsymbol{g}_1 - \boldsymbol{g}, \ldots, \boldsymbol{g}_C - \boldsymbol{g})$, and thus a faster change of the component that $\text{span}(\boldsymbol{g}_1 - \boldsymbol{g}, \ldots, \boldsymbol{g}_C - \boldsymbol{g})$ overlaps with the top-$(C-1)$ subspace of the Hessian. While both phenomenon seems to be a predictor of when the top subspace stabilization fails, we leave the precise mechanistic understanding of this failure as a future work.

5 Conclusion

In this paper we study the phenomenon that the top subspace of the Hessian stabilizes shortly after the training starts. We designed a new instability metric that removes the confound of small parameter update magnitude in the overlap metric, and used it to to confirm the occurrence of the stabilization in some setting of neural network training. We also identified a setting where stabilization fails despite a near-perfect step-wise overlap. To explain the stabilization, we first investigated the relationship between the top subspace of the gradient covariance matrix and the Hessian, and constructed an explicit identifier for this subspace. Under gradient flow, we upper bound the change of this subspace by the operator norm of the Hessian restricted to the bulk, showing that the separation between outlier and bulk eigenvalues is sufficient to explain the early stabilization we observe.

Our work leaves several open questions: $i)$. a theoretical justification of why the span of $\{\boldsymbol{g}_c - \boldsymbol{g}\}_{c=1}^C$ identifies the top subspace of the Hessian; $ii)$. a precise characterization of the stabilization failure in some setup; and $iii)$. an understanding of the growth of the top-$(C-1)$ eigenvalues and why they are separated from the bulk. We leave them as future directions.

AI USE STATEMENT

In this work, we used generative AI tools for helping develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, implement methods. We have not used generative AI tools for generating synthetic data sets, formulating mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, cleaning and reformating dataset, supporting qualitative and thematic data analysis, interpreting results. Assisting with translation are not applicable to this work. Additionally, we used generative AI tools for drafting parts of a research paper, summarizing or analyzing existing literature, brainstorming, sourcing/searching for information, and proposing a title or keywords for a research paper. We have reviewed all AI-assisted work. Mathematical models proposed by AI is verified by author for mathematical correctness, and supported by experimental results. Critical gradients for proving mathematical claims are verified by author before incorporating into the proof. Implementations are checked and tested by author for correctness. Drafted sections, summarized literature, and suggested keywords are all proofread and edited by the author before put into the paper. Brainstorming and sourced information does not directly influence the outcome of the work.

REPRODUCIBILITY STATEMENT

Our paper involves contribution in both theoretical and experimental results. For the experimental results, information about the experiment setup is provided in Appendix A where we we discussed dataset, model architecture, and hyperparameter details. We also provide the code as supplementary material. For the theoretical results:

- Proof of Theorem 1 is in Appendix B
- Proof of Theorem 2 is in Appendix C
- Proof of Theorem 3 is in Appendix D
- Proof of Theorem 4 is in Appendix E
- All lemmas used in the proof of the theorems are provided in Appendix F

REFERENCES

Zhiqi Bu, Shiyun Xu, and Jialin Mao. Convex dominance in deep learning i: A scaling law of loss and learning rate, 2026. URL https://arxiv.org/abs/2602.07145.

Anna Choromanska, MIkael Henaff, Michael Mathieu, Gerard Ben Arous, and Yann LeCun. The Loss Surfaces of Multilayer Networks. In Guy Lebanon and S. V. N. Vishwanathan (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 192–204, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/choromanska15.html.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Micah Goldblum, Jonas Geiping, Avi Schwarzschild, Michael Moeller, and Tom Goldstein. Truth or backpropaganda? an empirical investigation of deep learning theory. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HyxyIgHFvr.

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace, 2018. URL https://arxiv.org/abs/1812.04754.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre.

Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Ajay Kumar Jaiswal, Yifan Wang, Lu Yin, Shiwei Liu, Runjin Chen, Jiawei Zhao, Ananth Grama, Yuandong Tian, and Zhangyang Wang. From low rank gradient subspace stabilization to low-rank weights: Observations, theories, and applications. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=fsf7LhbYdf.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

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. URL https://openreview.net/forum?id=H1oyRlYgg.

Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/46a558d97954d0692411c861cf78ef79-Paper.pdf.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.

Fangshuo Liao and Anastasios Kyrillidis. Provable accelerated convergence of nesterov's momentum for deep relu neural networks. In Claire Vernade and Daniel Hsu (eds.), *Proceedings of The 35th International Conference on Algorithmic Learning Theory*, volume 237 of *Proceedings of Machine Learning Research*, pp. 732–784. PMLR, 25–28 Feb 2024. URL https://proceedings.mlr.press/v237/liao24a.html.

Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, Sham Kakade, and Alex Damian. A defense of the quadratic model, 2026. URL https://arxiv.org/abs/2607.21716.

Vardan Papyan. Measurements of three-level hierarchical structure in the outliers in the spectrum of deepnet hessians, 2019. URL https://arxiv.org/abs/1901.08244.

Vardan Papyan. Traces of class/cross-class structure pervade deep learning spectra, 2020. URL https://arxiv.org/abs/2008.11865.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. ISSN 1091-6490. doi: 10.1073/pnas.2015509117. URL http://dx.doi.org/10.1073/pnas.2015509117.

Shikai Qiu, Lechao Xiao, Andrew Gordon Wilson, Jeffrey Pennington, and Atish Agarwala. Scaling collapse reveals universal dynamics in compute-optimally trained neural networks. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Fvq9ogLnLN.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond, 2017. URL https://arxiv.org/abs/1611.07476.

Fabian Schaipp, Alexander Hägele, Adrien Taylor, Umut Simsekli, and Francis Bach. The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=b836TGkRSw.

Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does SGD really happen in tiny subspaces? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=v6iLQBoIJw.

Grzegorz Swirszcz, Wojciech Marian Czarnecki, and Razvan Pascanu. Local minima in training of neural networks, 2017. URL https://arxiv.org/abs/1611.06310.

Kaiyue Wen, Zhiyuan Li, Jason S. Wang, David Leo Wright Hall, Percy Liang, and Tengyu Ma. Understanding warmup-stable-decay learning rates: A river valley loss landscape view. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=m51BgoqvbP.

Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from sharp minima and regularization effects. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7654–7663. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/zhu19e.html.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29820–29834. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf.

A Experimental Details

A.1 Common experimental protocol

Each experiment trains a randomly initialized classifier with plain SGD on a fixed-size subset of a standard image classification benchmark.

- **Optimizer**: vanilla SGD, no momentum (`momentum=0.0`), no weight decay, no learning-rate schedule, constant learning rate throughout.
- **Loss**: cross-entropy on the full 10-way label, both for training and for every Hessian/gradient statistic below.
- **Trials**: each setting runs 4 independent repetitions. Each trial gets a freshly initialized model and a fresh optimizer; training data (and, for CIFAR, the specific random subset of the dataset) is the same across trials within a notebook.
- **Results**: all results are aggregated over 4 trials, and the mean is plotted with shaded region denoting the mean$\pm$std.

A.2 Datasets and preprocessing

MNIST Experiment (MLP):

- **Resolution**: native 28×28, 1 channel
- **Preprocessing**: raw pixel values in $[0, 1]$, no standardization
- **Training subset**: random 10,000-example subset of the 60,000-example train split
- **Test set**: full standard 10,000-example test split (no subsetting)

CIFAR experiments (2- and 3-layer CNN)

- **Resolution**: native 32×32, 3 channels
- **Preprocessing**: standardized per input dimension (channel $\times$ row $\times$ col) using the *training subset's own* mean/std (computed after subsetting), applied to both train and test splits

- **Training subset**: random 5,000-example subset of the 50,000-example train split
- **Test set**: full standard 10,000-example test split (no subsetting)

A.3 Model architectures

- **MLP** (MNIST): fully connected, biases enabled. $784 \to 256 \to 128 \to 10$, ReLU activations after the first two linear layers, no activation on the output layer. No normalization layers, no dropout.
- **2-Layer CNN** (CIFAR-10): two (`Conv2d(3x3, pad=1)` $\to$ ReLU $\to 2 \times 2$ max-pool) blocks at constant width 32 ($32 \to 32$ channels), *no BatchNorm*, followed by a flatten and a single linear classifier head ($32 \times 8 \times 8 = 2048 \to 10$). Matches the "edge-of-stability" repository's `convnet` with `widths=[32, 32]`.
- **3-Layer CNN** (CIFAR-10): three (`Conv2d(3x3, pad=1)` $\to$ ReLU $\to$ `BatchNorm2d` $\to 2 \times 2$ max-pool) blocks at constant width 32, followed by a flatten and a single linear classifier head ($32 \times 4 \times 4 = 512 \to 10$).

A.4 Per-experiment hyperparameters

	MNIST	CIFAR 2-layer	CIFAR 3-layer
Architecture	MLP	2-Layer CNN	3-Layer CNN
Train subset size	10,000	5,000	5,000
Minibatch size	100	50	50
Learning rate	0.01	0.001	0.001
Total SGD steps	5,000	12,000	6,000
Steps / epoch	100	100	100
Epochs	50	120	60
Ckpt Freq	100 (every epoch)	300 (every 3 epochs)	300 (every 3 epochs)
Ckpt number	51 (initial + 50)	41 (initial + 40)	21 (initial + 20)

Compute. All experiments are run on Google Colab L4 and A100 GPU.

B Proof of Theorem 1

Proof. For a matrix function $\boldsymbol{A}(t)$, its inverse's time derivative is given by

$$\mathbf{0} = \frac{d}{dt}\left(\boldsymbol{A}(t)\boldsymbol{A}^{-1}(t)\right) = \dot{\boldsymbol{A}}(t)\boldsymbol{A}^{-1}(t) + \boldsymbol{A}(t)\dot{\boldsymbol{A}}^{-1}(t)$$

Therefore

$$\dot{\boldsymbol{A}}^{-1}(t) = -\boldsymbol{A}^{-1}(t)\dot{\boldsymbol{A}}(t)\boldsymbol{A}^{-1}(t)$$

Notice that $\frac{d}{dt}\left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right) = \dot{\boldsymbol{G}}(t)\boldsymbol{G}(t)^\top + \boldsymbol{G}(t)\dot{\boldsymbol{G}}(t)^\top$. Using this formula, we have that

$$\begin{aligned}
\dot{\boldsymbol{P}}(t) &= \dot{\boldsymbol{G}}(t)^\top \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \boldsymbol{G}(t) + \boldsymbol{G}(t)^\top \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \dot{\boldsymbol{G}}(t) \\
&\quad - \boldsymbol{G}(t)^\top \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \left(\dot{\boldsymbol{G}}(t)\boldsymbol{G}(t)^\top + \boldsymbol{G}(t)\dot{\boldsymbol{G}}(t)^\top\right) \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \boldsymbol{G}(t) \\
&= \dot{\boldsymbol{G}}(t)^\top \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \boldsymbol{G}(t) + \boldsymbol{G}(t)^\top \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \dot{\boldsymbol{G}}(t) \\
&\quad - \boldsymbol{G}(t)^\top \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \dot{\boldsymbol{G}}(t)\boldsymbol{P}(t) - \boldsymbol{P}(t)\dot{\boldsymbol{G}}(t)^\top \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \boldsymbol{G}(t) \\
&= \boldsymbol{G}(t)^\top \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \dot{\boldsymbol{G}}(t) \left(\boldsymbol{I} - \boldsymbol{P}(t)\right) \\
&\quad + \left(\boldsymbol{I} - \boldsymbol{P}(t)\right) \dot{\boldsymbol{G}}(t)^\top \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \boldsymbol{G}(t) \\
&= \boldsymbol{X}(t) + \boldsymbol{X}(t)^\top
\end{aligned}$$

where in the last line we defined

$$\boldsymbol{X}(t) = \left(\boldsymbol{I} - \boldsymbol{P}(t)\right) \dot{\boldsymbol{G}}(t)^\top \left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1} \boldsymbol{G}(t)$$

This gives that

$$\begin{aligned}\left\|\dot{\boldsymbol{P}}(t)\right\|_F &\leq 2\left\|\boldsymbol{X}(t)\right\|_F \\ &\leq 2\left\|\left(\boldsymbol{I}-\boldsymbol{P}(t)\right)\dot{\boldsymbol{G}}(t)\right\|_F\left\|\left(\boldsymbol{G}(t)\boldsymbol{G}(t)^\top\right)^{-1}\boldsymbol{G}(t)\right\|_2 \\ &\leq \frac{2}{\sigma_{\min}\left(\boldsymbol{G}(t)\right)}\left\|\left(\boldsymbol{I}-\boldsymbol{P}(t)\right)\dot{\boldsymbol{G}}(t)^\top\right\|_F\end{aligned}$$

To understand $\dot{\boldsymbol{g}}(t)$, we plug in the gradient flow dynamic to obtain that

$$\frac{d}{dt}\boldsymbol{g}_c(t) = \boldsymbol{H}_c(t)\dot{\boldsymbol{\theta}}(t) = -\boldsymbol{H}_c(t)\boldsymbol{g}(t);\ \frac{d}{dt}\hat{\boldsymbol{g}}(t) = \boldsymbol{H}(t)\dot{\boldsymbol{\theta}}(t) = -\boldsymbol{H}(t)\boldsymbol{g}(t)$$

Therefore,

$$\begin{aligned}\left\|\left(\boldsymbol{I}-\boldsymbol{P}(t)\right)\dot{\boldsymbol{g}}(t)^\top\right\|_F^2 &= \sum_{c=1}^{C-1}\left\|\left(\boldsymbol{I}-\boldsymbol{P}(t)\right)\left(\boldsymbol{H}(t)-\boldsymbol{H}_c(t)\right)\boldsymbol{g}(t)\right\|_2^2 \\ &\leq (C-1)\left\|\left(\boldsymbol{I}-\boldsymbol{P}(t)\right)\boldsymbol{H}(t)\hat{\boldsymbol{g}}(t)\right\|_2^2 + \sum_{c=1}^{C-1}\left\|\left(\boldsymbol{I}-\boldsymbol{P}(t)\right)\boldsymbol{H}_c(t)\hat{\boldsymbol{g}}(t)\right\|_2^2\end{aligned}$$

Pugging into the bound of $\left\|\dot{\boldsymbol{P}}(t)\right\|_F$ gives

$$\left\|\dot{\boldsymbol{P}}(t)\right\|_F^2 \leq \frac{4}{\sigma_{\min}\left(\boldsymbol{g}(t)\right)^2}\left((C-1)\left\|\left(\boldsymbol{I}-\boldsymbol{P}(t)\right)\boldsymbol{H}(t)\hat{\boldsymbol{g}}(t)\right\|_2^2 + \sum_{c=1}^{C-1}\left\|\left(\boldsymbol{I}-\boldsymbol{P}(t)\right)\boldsymbol{H}_c(t)\hat{\boldsymbol{g}}(t)\right\|_2^2\right)$$

□

C Proof of Theorem 2

Proof. By definition of the minimum eigenvalue, we have that

$$\begin{aligned}\lambda_{\min}\left(\boldsymbol{H}\right) &= \min_{\boldsymbol{v}:\|\boldsymbol{v}\|=1}\boldsymbol{v}^\top\boldsymbol{H}\boldsymbol{v} \\ &= \min_{\boldsymbol{v}:\|\boldsymbol{v}\|}\sum_{c=1}^{C}\frac{n_c}{n}\boldsymbol{v}^\top\boldsymbol{H}_c\boldsymbol{v} \\ &\geq \sum_{c=1}^{C}\frac{n_c}{n}\min_{\boldsymbol{v}:\|\boldsymbol{v}\|}\boldsymbol{v}^\top\boldsymbol{H}_c\boldsymbol{v} \\ &= -\sum_{c=1}^{C}\frac{n_c}{n}\cdot\nu \\ &= -\nu\end{aligned}$$

Therefore, we have that

$$\begin{aligned}\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{H}\right\|_2 &\leq \left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{H}\left(\boldsymbol{I}-\boldsymbol{P}\right)\right\|_2 + \left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{H}\boldsymbol{P}\right\| \\ &\leq \left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\left(\boldsymbol{H}+\nu\boldsymbol{I}\right)\boldsymbol{P}\right\|_2 + \nu + \varepsilon\end{aligned}$$

Notice that $\boldsymbol{H}+\nu\boldsymbol{I}$ is positive semi-definite. Therefore, we can envoke Lemma 1 to obtain that

$$\begin{aligned}\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\left(\boldsymbol{H}+\nu\boldsymbol{I}\right)\boldsymbol{P}\right\|_2 &\leq \left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\left(\boldsymbol{H}+\nu\boldsymbol{I}\right)\left(\boldsymbol{I}-\boldsymbol{P}\right)\right\|_2^{\frac{1}{2}}\left\|\boldsymbol{P}\left(\boldsymbol{H}+\nu\boldsymbol{I}\right)\boldsymbol{P}\right\|_2^{\frac{1}{2}} \\ &\leq \sqrt{\left(\varepsilon+\nu\right)\left(\left\|\boldsymbol{H}\right\|_2+\nu\right)}\end{aligned}$$

This gives that

$$\left\|\left(\boldsymbol{I}-\boldsymbol{P}\right)\boldsymbol{H}\right\|_2 \leq \sqrt{\left(\varepsilon+\nu\right)\left(\left\|\boldsymbol{H}\right\|_2+\nu\right)} + \varepsilon + \nu \leq 2\sqrt{\left(\varepsilon+\nu\right)\left(\left\|\boldsymbol{H}\right\|_2+\nu\right)}$$

since $\varepsilon \leq \|\boldsymbol{H}\|_2$. Now, we focus on $\boldsymbol{H}_c$. Using a similar technique, we obtain that

$$\begin{aligned}\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}_c\|_2 &= \|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{H}_c+\nu\boldsymbol{I})\| + \nu \\ &\leq \|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{H}_c+\nu\boldsymbol{I})(\boldsymbol{I}-\boldsymbol{P})\|_2 + \|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{H}_c+\nu\boldsymbol{I})\boldsymbol{P}\|_2 + \nu\end{aligned}$$

For the first term, we notice that

$$\sum_{c'=1}^{C}\frac{n_c'}{n}(\boldsymbol{H}_c'+\nu\boldsymbol{I}) = \boldsymbol{H}+\nu\boldsymbol{I} \;\Rightarrow\; (\boldsymbol{H}+\nu\boldsymbol{I}) - \frac{n_c}{n}(\boldsymbol{H}_c+\nu\boldsymbol{I}) = \sum_{c'\neq c}\frac{n_c'}{n}(\boldsymbol{H}_c'+\nu\boldsymbol{I}) \succeq 0$$

since each $\boldsymbol{H}_c'+\nu\boldsymbol{I}$ must be positive semi-definite. Therefore, $\|\boldsymbol{H}_c+\nu\boldsymbol{I}\|_2 \leq \frac{n}{n_c}\|\boldsymbol{H}+\nu\boldsymbol{I}\|_2$, and

$$\frac{n_c}{n}\|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{H}_c+\nu\boldsymbol{I})(\boldsymbol{I}-\boldsymbol{P})\|_2 \leq \|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{H}+\nu\boldsymbol{I})(\boldsymbol{I}-\boldsymbol{P})\|_2 \leq \nu+\varepsilon$$

Thus, we have that

$$\|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{H}_c+\nu\boldsymbol{I})(\boldsymbol{I}-\boldsymbol{P})\|_2 \leq \frac{n}{n_c}(\nu+\varepsilon)$$

Moreover, for $\|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{H}_c+\nu\boldsymbol{I})\boldsymbol{P}\|_2$, by Lemma 1

$$\begin{aligned}\|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{H}_c+\nu\boldsymbol{I})\boldsymbol{P}\|_2 &\leq \|(\boldsymbol{I}-\boldsymbol{P})(\boldsymbol{H}_c+\nu\boldsymbol{I})(\boldsymbol{I}-\boldsymbol{P})\|_2^{\frac{1}{2}}\,\|\boldsymbol{P}(\boldsymbol{H}_c+\nu\boldsymbol{I})\boldsymbol{P}\|_2^{\frac{1}{2}} \\ &\leq \frac{n}{n_c}\sqrt{(\nu+\varepsilon)(\|\boldsymbol{H}\|_2+\nu)}\end{aligned}$$

Combining, we have that

$$\|(\boldsymbol{I}-\boldsymbol{P})\boldsymbol{H}_c\|_2 \leq \frac{n}{n_c}(\nu+\varepsilon) + \frac{n}{n_c}\sqrt{(\nu+\varepsilon)(\|\boldsymbol{H}\|_2+\nu)} \leq \frac{2n}{n_c}\sqrt{(\nu+\varepsilon)(\|\boldsymbol{H}\|_2+\nu)}$$

□

D Proof of Theorem 3

Proof. We write

$$\tilde{\boldsymbol{g}}_c = \sum_{c\neq c'}\sum_{i\in\mathcal{I}_c}\pi_{c,c'}\pi_{i,c,c'}\tilde{\boldsymbol{g}}_{i,c,c'} = \frac{1}{\omega_c}\sum_{c\neq c'}\sum_{i\in\mathcal{I}_c}\omega_{i,c,c'}\tilde{\boldsymbol{g}}_{i,c,c'} = \frac{1}{\omega_c}\sum_{i\in\mathcal{I}_c}\sum_{c\neq c'}\omega_{i,c,c'}\tilde{\boldsymbol{g}}_{i,c,c'}$$

Define the Jacobian $\mathcal{J}_{i,c} = \frac{\partial f(\boldsymbol{\theta},\boldsymbol{x}_{i,c})}{\boldsymbol{\theta}} \in \mathbb{R}^{C\times p}$. Then we have that $\tilde{\boldsymbol{g}}_{i,c,c'} = \mathcal{J}_{i,c}^{\top}(\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_{i,c}) - \boldsymbol{e}_{c'})$. Therefore,

$$\begin{aligned}\sum_{c'=1}^{C}\omega_{i,c,c'}\boldsymbol{g}_{i,c,c'} &= \frac{1}{n_c C}\mathcal{J}_{i,c}\sum_{c'=1}^{C}p_{i,c,c'}(\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_{i,c}) - \boldsymbol{e}_{c'}) \\ &= \frac{1}{n_c C}\mathcal{J}_{i,c}\left(\sum_{c'=1}^{C}p_{i,c,c'}\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_{i,c}) - \sum_{c'=1}^{C}p_{i,c,c'}\boldsymbol{e}_{c'}\right) \\ &= \frac{1}{n_c C}\mathcal{J}_{i,c}(\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_{i,c}) - \boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_{i,c})) \\ &= 0\end{aligned}$$

Therefore, we have that

$$\tilde{\boldsymbol{g}}_c = -\frac{1}{\omega_c}\sum_{i\in\mathcal{I}_c}\omega_{i,c,c}\tilde{\boldsymbol{g}}_{i,c,c} = -\frac{1}{\omega_c n_c C}\sum_{i\in\mathcal{I}_c}p_{i,c,c}\tilde{\boldsymbol{g}}_{i,c,c} = -\frac{1}{\omega_c n_c C}\sum_{i\in\mathcal{I}_c}\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_{i,c})_c\,\hat{\boldsymbol{g}}_i(\boldsymbol{\theta})$$

Redefine $\boldsymbol{g}_c' = \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_{i,c})_c\,\hat{\boldsymbol{g}}_i(\boldsymbol{\theta})$. Since $\boldsymbol{g}_c$'s are linearly independent, we have that

$$\text{span}(\boldsymbol{g}_{\text{class}}) = \text{span}(\tilde{\boldsymbol{g}}_1,\ldots,\tilde{\boldsymbol{g}}_C) = \text{span}(\boldsymbol{g}_1',\ldots,\boldsymbol{g}_C')$$

□

E PROOF OF THEOREM 4

Proof. Let $\boldsymbol{q} \in \text{span}\left(\boldsymbol{g}_1\left(\boldsymbol{\theta}\right), \ldots \boldsymbol{g}_C\left(\boldsymbol{\theta}\right)\right)^\top$ be a unit vector. Then we must have that $\boldsymbol{q} \in \text{col}\left(\boldsymbol{\Sigma}_{\text{top}}\left(\boldsymbol{\theta}\right)\right)^\perp$. By definition, we have that

$$\left(\boldsymbol{q}^\top \boldsymbol{g}_c'\left(\boldsymbol{\theta}\right)\right)^2 = \left(\boldsymbol{q}^\top \left(\boldsymbol{g}_c'\left(\boldsymbol{\theta}\right) - n_c \bar{p}_c \boldsymbol{g}_c\left(\boldsymbol{\theta}\right)\right)\right)^2 \leq \frac{1}{4}\boldsymbol{q}^\top \boldsymbol{\Sigma}_c\left(\boldsymbol{\theta}\right)\boldsymbol{q}$$

where the last inequality follows Lemma 2. Let $\boldsymbol{g} \in \mathbb{R}^{d\times C}$ be the matrix such that the cth column is $\sqrt{\frac{n_c}{n}}\boldsymbol{g}_c'\left(\boldsymbol{\theta}\right)$. Then we have that

$$\left\|\boldsymbol{g}^\top \boldsymbol{q}\right\|_2^2 = \sum_{c=1}^{C} \frac{n_c}{n}\left(\boldsymbol{q}^\top \boldsymbol{g}_c'\left(\boldsymbol{\theta}\right)\right)^2 \leq \frac{1}{4}\boldsymbol{q}^\top \sum_{c=1}^{C}\frac{n_c}{n}\boldsymbol{\Sigma}_c\left(\boldsymbol{\theta}\right)\boldsymbol{q} = \frac{1}{4}\boldsymbol{q}^\top \boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right)\boldsymbol{q}$$

Notice that $\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top$ is the projection operator onto $\text{span}\left(\boldsymbol{g}_1\left(\boldsymbol{\theta}\right), \ldots \boldsymbol{g}_C\left(\boldsymbol{\theta}\right)\right)^\perp$. Then we have that

$$\left\|\left(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top\right)\boldsymbol{g}\right\|_2^2 \leq \sup_{\boldsymbol{q}\in \text{span}\left(\boldsymbol{g}_1\left(\boldsymbol{\theta}\right), \ldots \boldsymbol{g}_C\left(\boldsymbol{\theta}\right)\right)^\top} \left\|\boldsymbol{g}^\top \boldsymbol{q}\right\|_2^2 \leq \frac{1}{4}\sup_{\boldsymbol{q}\in \text{col}\left(\boldsymbol{\Sigma}_{\text{top}}\left(\boldsymbol{\theta}\right)\right)^\perp} \boldsymbol{q}^\top \boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right)\boldsymbol{q} \leq \frac{1}{4}\left\|\boldsymbol{\Sigma}_{\text{bulk}}\left(\boldsymbol{\theta}\right)\right\|_2$$

Therefore

$$\left\|\left(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top\right)\boldsymbol{u}'\right\|_2^2 \leq \frac{1}{\sigma_{\min}\left(\boldsymbol{g}\right)^2}\left\|\left(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^\top\right)\boldsymbol{g}\right\|_2^2 = \frac{\left\|\boldsymbol{\Sigma}_{\text{bulk}}\left(\boldsymbol{\theta}\right)\right\|_2}{4\sigma_{\min}\left(\boldsymbol{g}\right)^2}$$

This implies that

$$\sigma_{\min}\left(\boldsymbol{u}^\top \boldsymbol{u}'\right)^2 \geq 1 - \frac{\left\|\boldsymbol{\Sigma}_{\text{bulk}}\left(\boldsymbol{\theta}\right)\right\|_2}{4\sigma_{\min}\left(\boldsymbol{g}\right)^2}$$

□

F AUXILIARY LEMMAS

Lemma 1. *Let $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ be a symmetric positive semi-definite matrix. Then for all matrices $\boldsymbol{q} \in \mathbb{R}^{d\times m_1}$ and $\boldsymbol{P} \in \mathbb{R}^{d\times m_2}$ we have that*

$$\left\|\boldsymbol{q}^\top \boldsymbol{M}\boldsymbol{P}\right\|_2^2 \leq \left\|\boldsymbol{q}^\top \boldsymbol{M}\boldsymbol{q}\right\|_2 \left\|\boldsymbol{P}^\top \boldsymbol{M}\boldsymbol{P}\right\|_2$$

Proof. For any vector $\boldsymbol{A}, \boldsymbol{b} \in \mathbb{R}^d$ such that $\boldsymbol{A}^\top \boldsymbol{M}\boldsymbol{b} \geq 0$ we have that

$$\left(\boldsymbol{A}^\top \boldsymbol{M}\boldsymbol{b}\right)^2 = \left\langle \boldsymbol{M}^{\frac{1}{2}}\boldsymbol{A}, \boldsymbol{M}^{\frac{1}{2}}\boldsymbol{b}\right\rangle^2 \leq \left\|\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{A}\right\|_2^2 \left\|\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{b}\right\|_2^2 = \boldsymbol{A}^\top \boldsymbol{M}\boldsymbol{A}\cdot \boldsymbol{b}^\top \boldsymbol{M}\boldsymbol{b}$$

Now, let $\boldsymbol{A} = \boldsymbol{q}\boldsymbol{u}$ and $\boldsymbol{b} = \boldsymbol{P}\boldsymbol{v}$ with $\left\|\boldsymbol{u}\right\|_2 = \left\|\boldsymbol{v}\right\|_2 = 1$. Then we have that

$$\left(\boldsymbol{u}^\top \boldsymbol{q}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{v}\right)^2 \leq \left(\boldsymbol{u}^\top \boldsymbol{q}^\top \boldsymbol{M}\boldsymbol{q}\boldsymbol{u}\right)\left(\boldsymbol{v}^\top \boldsymbol{P}^\top \boldsymbol{M}\boldsymbol{P}\boldsymbol{v}\right) \leq \left\|\boldsymbol{q}^\top \boldsymbol{M}\boldsymbol{q}\right\|_2 \left\|\boldsymbol{P}^\top \boldsymbol{M}\boldsymbol{P}\right\|_2$$

Since the inequality above holds for all $\boldsymbol{u}, \boldsymbol{v}$ with unit norm, we can conclude that

$$\left\|\boldsymbol{q}^\top \boldsymbol{M}\boldsymbol{P}\right\|_2^2 \leq \left\|\boldsymbol{q}^\top \boldsymbol{M}\boldsymbol{q}\right\|_2 \left\|\boldsymbol{P}^\top \boldsymbol{M}\boldsymbol{P}\right\|_2$$

□

Lemma 2. *Let $\boldsymbol{g}_c = \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\hat{\boldsymbol{g}}_i$ and let $\boldsymbol{g}_c' = \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\boldsymbol{p}\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)_c\hat{\boldsymbol{g}}_i$ for a fixed $\boldsymbol{\theta}$. Then for any vector $\boldsymbol{q}$ we have that*

$$\left(\boldsymbol{q}^\top\left(\boldsymbol{g}_c' - \bar{p}_c\boldsymbol{g}_c\left(\boldsymbol{\theta}\right)\right)\right)^2 \leq \frac{1}{4}\boldsymbol{q}^\top \boldsymbol{\Sigma}_c\left(\boldsymbol{\theta}\right)\boldsymbol{q}$$

where $\bar{p}_c = \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\boldsymbol{p}\left(\boldsymbol{\theta}, \boldsymbol{x}_i\right)_c$ and $\boldsymbol{\Sigma}_c\left(\boldsymbol{\theta}\right)$ is defined in (6).

Proof. Recall the definition

$$\boldsymbol{g}_c(\boldsymbol{\theta}) = \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}); \quad \boldsymbol{g}'_c(\boldsymbol{\theta}) = \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_i)_c\,\hat{\boldsymbol{g}}_i(\boldsymbol{\theta})$$

Fix $\boldsymbol{\theta}$. Let $\bar{p}_c := \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\boldsymbol{P}(\boldsymbol{\theta},\boldsymbol{x}_i)_c$. Then we have that

$$\begin{aligned}\boldsymbol{g}'_c(\boldsymbol{\theta}) &= \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_i)_c\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right) + \bar{p}_c\boldsymbol{g}_c(\boldsymbol{\theta}) \\ &= \frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\left(\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_i)_c - \bar{p}_c\right)\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right) + \bar{p}_c\boldsymbol{g}_c(\boldsymbol{\theta})\end{aligned}$$

For any vector $\boldsymbol{q}\in\mathbb{R}^p$ we have that

$$\begin{aligned}&\left(\boldsymbol{q}^\top\left(\boldsymbol{g}'_c - \bar{p}_c\boldsymbol{g}_c(\boldsymbol{\theta})\right)\right)^2 \\ &\quad= \left(\frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\left(\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_i)_c - \bar{p}_c\right)\boldsymbol{q}^\top\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right)\right)^2 \\ &\quad\le \left(\frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\left(\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_i)_c - \bar{p}_c\right)^2\right)\boldsymbol{q}^\top\left(\frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right)\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right)^\top\right)\boldsymbol{q}\end{aligned}$$

Notice that

$$\frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\left(\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_i)_c - \bar{p}_c\right)^2 = \mathrm{Var}\left(\boldsymbol{p}(\boldsymbol{\theta},\boldsymbol{x}_i)_c\right) \le \bar{p}_c(1-\bar{p}_c) \le \frac{1}{4}$$

In the meantime

$$\frac{1}{n_c}\sum_{i\in\mathcal{I}_c}\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right)\left(\hat{\boldsymbol{g}}_i(\boldsymbol{\theta}) - \boldsymbol{g}_c(\boldsymbol{\theta})\right)^\top = \boldsymbol{\Sigma}_c(\boldsymbol{\theta}) - \left(\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)\left(\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)^\top$$

Therefore

$$\left(\boldsymbol{q}^\top\left(\boldsymbol{g}'_c - \bar{p}_c\boldsymbol{g}_c(\boldsymbol{\theta})\right)\right)^2 \le \frac{1}{4}\left(\boldsymbol{q}^\top\boldsymbol{\Sigma}_c(\boldsymbol{\theta})\,\boldsymbol{q} - \left(\boldsymbol{q}^\top\left(\boldsymbol{g}_c(\boldsymbol{\theta}) - \boldsymbol{g}(\boldsymbol{\theta})\right)\right)^2\right) \le \frac{1}{4}\boldsymbol{q}^\top\boldsymbol{\Sigma}_c(\boldsymbol{\theta})\,\boldsymbol{q}$$

□