AGRASS: UNDERSTANDING AND SHAPING GRADIENT SPECTRA FOR EEG LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

EEG signals exhibit substantial complexity, inter-subject heterogeneity, and are susceptible to noise, making it challenging to learn meaningful task-relevant representations. While the mainstream methods use neural networks trained through gradient-based optimization, the information carried by different gradient directions during EEG learning remains insufficiently understood. To this end, we investigate the spectral structure of task gradients in EEG models and propose a new optimizer tailored to EEG learning. Our analysis reveals pronounced spectral imbalance of gradients during training and finds a strong association between the leading spectral head and subject-related variation, as well as a low estimated gradient signal-to-noise ratio in the spectral tail. Theoretical analyses with a latent subject-variation model explain such gradient structure of the spectral head. Motivated by this analysis, we propose AGraSS (**A**daptive **Gra**dient **S**pectral **S**caling), an optimizer that combines automatic spectral-head detection with soft spectral shrinkage. AGraSS adaptively rescales the spectral head with additional attenuation on the pre-detected subject-related directions, moderating dominant update directions. Experiments on six EEG datasets with both expert models and pretrained EEG foundation models demonstrate the superiority of AGraSS, outperforming mainstream optimizers such as AdamW and Muon with improved performance. These findings highlight EEG-aware optimizer design as an important direction for improving EEG learning.

1 INTRODUCTION

Electroencephalography (EEG) provides a measure of brain activity and supports a broad range of applications, including brain–computer interfaces, neurological assessment, and cognitive-state decoding. Learning useful representations from EEG, however, remains challenging. Task-dependent neural activity is recorded alongside substantial inter-subject variability, differences in acquisition conditions, and noise. This heterogeneity, together with the low signal-to-noise ratio of EEG recordings, complicates the extraction of stable, task-relevant representations (Jiang et al., 2024; El Ouahidi et al., 2026; Wang et al., 2025).

Recent progress try to address these challenges through EEG-specific architectures and large-scale pretraining-based task-specific finetuning. Expert models incorporate spatial, temporal, and other inductive biases (Lawhern et al., 2018; Song et al., 2022a; Ding et al., 2024), while EEG foundation models leverage large-scale pretraining for task-specific finetuning (Jiang et al., 2024; El Ouahidi et al., 2026; Wang et al., 2025; Zhou et al., 2026; Yang et al., 2023). Despite their architectural differences, these models share a common learning mechanism: their task-specific representations are shaped through gradient-based parameter updates. Gradients therefore mediate how EEG data and training objectives influence the features a model learns. However, the information reflected in different gradient directions and its role in task-specific representations remain insufficiently understood (Xiong et al., 2025). This raises a complementary question: what structure underlies EEG training gradients, and how can understanding this structure inform optimizer design?

Addressing this question requires examining the distribution of gradient energy across update directions, rather than gradient magnitude alone. Many parameters in EEG networks have a matrix structure, with their gradients coupling changes across input and output features. An entrywise distribution or a single gradient norm does not reveal whether update energy is spread broadly or

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.

concentrated in a few coordinated directions (Jordan et al., 2024; Liu et al., 2025). For matrix-valued gradients, singular value decomposition separates these components into orthogonal directions, while the resulting singular-value spectrum describes their relative strengths. This gradient spectrum reveals whether gradient energy is concentrated in a few components or distributed more broadly (Gupta et al., 2018). This gradient spectral perspective therefore provides a way to connect the composition and reliability of EEG learning signals with optimizer design. For EEG, the corresponding question is how this structure reflects the distinctive variability of the underlying signals.

In this work, we investigate the gradient spectral structure in EEG models. Our analyses show that gradient energy is highly concentrated in a small number of leading singular components. And we further find that the gradient spectra heads are strongly associated with subject-related variation. Meanwhile, the spectral tail exhibits heightened noise sensitivity. Motivated by these findings, we propose AGraSS (**A**daptive **Gra**dient **S**pectral **S**caling), an optimizer tailored to EEG learning. Building on residual-preserving spectral-head rescaling, AGraSS determines the head ratio automatically from gradient spectra and introduces additional subject-aware soft shrinkage. A training-data calibration procedure estimates shared attenuation coefficients for gradient spectral head that exhibit consistent subject-related conflict. During optimization, AGraSS rescales the selected head using the residual spectral scale and applies these coefficients for further attenuation. The residual components are preserved during spectral shaping, avoiding tail-specific amplification.

Our contributions are summarized as follows:

- We conduct the gradient spectral analysis of EEG learning and reveal three characteristics of EEG gradients: pronounced spectral concentration, a strong association between leading modes and subject-related variation, and heightened noise sensitivity in the spectral tail. We also provide theoretical analyses to explain such gradient structure of the spectral head.
- We propose a new EEG optimizer AGraSS, combining data-driven head-ratio selection with subject-aware soft shrinkage and efficient low-rank updates, while preserving the residual components during spectral shaping.
- We evaluate AGraSS on six EEG datasets with both task-specific models and pretrained EEG foundation model, demonstrating the benefits of EEG-informed optimization across different models and datasets.

2 EEG Learning Analysis

Generic optimizers either process gradient coordinates independently, as in AdamW, or apply nearly uniform transformations across matrix singular modes, as in orthogonalized optimizers. Neither design explicitly accounts for the spectral structure induced by EEG data. We therefore investigate three questions: **(i)** what spectral structure EEG task gradients exhibit, **(ii)** whether small singular modes constitute reliable update directions, and **(iii)** how subject-related variation interacts with the task-gradient spectrum.

2.1 Gradient Spectral Analysis

Let $\mathcal{D}_{\mathrm{tr}} = \{(x_i, y_i, s_i)\}_{i=1}^{N}$ denote the training set, where x_i is an EEG segment, y_i is its task label, and s_i is the subject identity. Consider a weight matrix $W \in \mathbb{R}^{m\times n}$, the gradient matrix of task loss $\mathcal{L}_T$ is $G = \nabla_W \mathcal{L}_T$. The compact singular value decomposition can get $G = U\Sigma V^\top = \sum_{k=1}^{r} \sigma_k u_k v_k^\top$, where $r = \min(m, n)$,$\{\sigma_i\}_{i=1}^{r}$ are the singular values, $\sigma_1 \geq \cdots \geq \sigma_r \geq 0$, $\{u_i\}$ are the left singular vectors, $\{v_i\}$ are the right singular vectors, and we denote $E_k = u_k v_k^\top$.

We first compare early and late-training gradient spectra to assess whether gradient energy remains concentrated in a small number of leading singular modes. Figure 1 compares normalized task-gradient spectra at $t = 0$ and $t = t^\star$.Across compact EEG networks and pretrained EEG foundation models, the spectra are strongly non-uniform: a small number of leading singular modes account for a disproportionate fraction of the gradient energy. More importantly, optimization does not eliminate this concentration. Although the overall gradient scale changes substantially, the separation between a dominant head and a smooth residual spectrum remains visible near convergence.

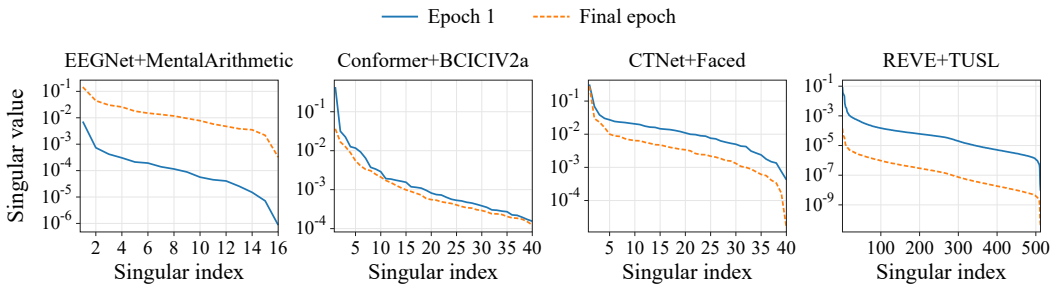


Figure 1: Singular-value spectra of different models with different dataset shows the consistent spectral imbalance. The spectral head is 1-2 orders of magnitude higher than the smooth tail.

2.2 Subject-related variation in the spectral head

The spectral imbalance observed above shows that a few leading singular components account for a disproportionate share of gradient energy. Yet their magnitudes alone do not reveal which sources of variation they capture. Given the substantial inter-subject variability of EEG, a natural question is whether the spectral head is strongly associated with subject-related variation that may confound the task-relevant representations.

To examine where subject-related variation is reflected in the task-gradient spectrum, we introduce an auxiliary subject-identity classifier on the encoder representations. The classifier is fitted using a subject classification loss $\mathcal{L}_S$, with the encoder and task classifier held fixed. Then, we define

$$G_{\text{inv}} = -\nabla_W \mathcal{L}_S, \tag{1}$$

as a gradient-based surrogate for subject invariance, where W denotes the weight matrix of EEG backbone. To characterize how subject-related gradient structure is distributed across the task-gradient spectrum, we project G_{inv} onto each task singular component:

$$b_k = \frac{u_k^\top G_{\text{inv}} v_k}{\|G_{\text{inv}}\|_F + \epsilon}. \tag{2}$$

As shown in Figure 2, the largest absolute projections are concentrated in a small number of leading task-gradient components, while projections onto later components are comparatively small and mostly close to zero. This pattern suggests that the spectral head is associated not only with large update magnitudes, but also with subject-related variation. Importantly, the leading projections can be either positive or negative: positive values indicate alignment with the local subject-invariance surrogate, whereas negative values indicate conflict.

2.3 Tail Singular Modes Are Directionally Unreliable

To further assess the reliability of gradient spectral components, we examine their gradient signal-to-noise ratios (SNRs) in a fixed task-gradient singular basis.

We construct a reference gradient $\bar{G}_T$ by averaging task gradients over the analysis samples and compute its singular value decomposition, $\bar{G}_T = \sum_k \bar{u}_k \bar{\sigma}_k \bar{v}_k^T$. Let $G_{T,j}$ be the task gradient computed from the j-th fixed-size, montage-compatible microbatch. We project it onto the singular basis

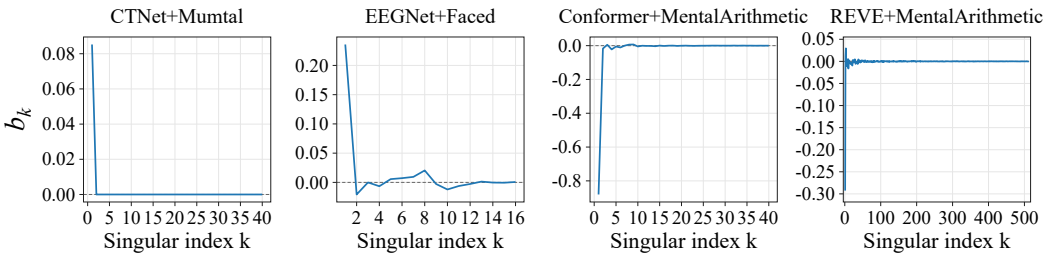


Figure 2: Subject-related gradient projections across the task-gradient spectrum. The largest absolute projections occur at leading indices, whereas gradient spectra tail exhibit smaller values.

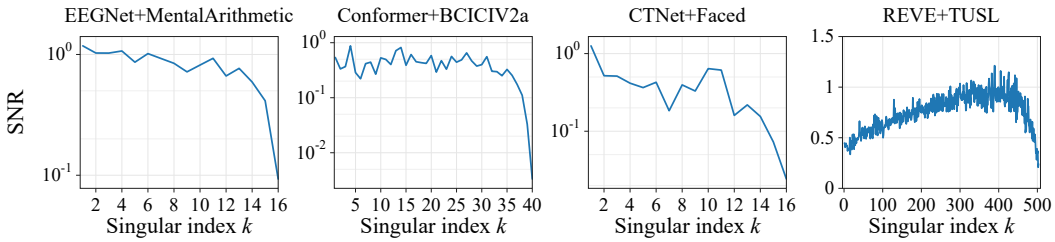


Figure 3: SNR(k) decreases with k, indicating more noise-dominated small-singular directions.

of the reference gradient $g_{j,k} = \langle G_j, E_k \rangle_F = \bar{u}_k^\top G_j \bar{v}_k$. We define $\mathrm{SNR}_k = \frac{\bar{\sigma}_k}{\mathrm{std}_j[g_{j,k}]}$.This scale-normalized measure quantifies the strength of each gradient component relative to its stochastic fluctuations, with lower values indicating greater sensitivity to microbatch sampling noise. Figure 3 shows that the smallest-singular-value modes exhibit low SNR across all four settings, with sampling fluctuations exceeding their reference signal magnitudes. relative variance. Small-singular-value modes are substantially more likely to exhibit stochastic variation comparable to, or larger than, their reference signal.

3 Theoretical Explanation

Our analysis finds that subject-gradient projections are concentrated in the leading task-gradient modes. We now investigate when this pattern can arise from the structure of EEG inputs.

Model and assumptions. We consider an additive input model,

$$x = \underbrace{\mu_y}_{\text{class-dependent component}} + \underbrace{Qa_s}_{\text{subject-dependent variation}} + \underbrace{\varepsilon}_{\text{other variation and noise}}, \qquad Q^\top Q = I_q. \tag{3}$$

where the EEG input $x \in \mathbb{R}^d$, $Q \in \mathbb{R}^{d\times q}$ is composed of q orthonormal, nonredundant bases for the subject-variation subspace with rank q. We assume $a_s \sim \mathcal{N}\big(0, \mathrm{diag}(\boldsymbol{\sigma}_{\mathrm{sub}}^2)\big)$, where $\boldsymbol{\sigma}_{\mathrm{sub}} = (\sigma_{\mathrm{sub}}^1, \ldots, \sigma_{\mathrm{sub}}^q)^\top$ contains the standard deviations along the subject-variation directions, and the square is applied elementwise. We denote $\rho = \min_{1\le j\le q} \sigma_{\mathrm{sub}}^j$.

To simplify the condition, we study a randomly initialized one-layer non-linear neural network $h_W(x) = \mathrm{ReLU}(Wx)$ with a linear task classifier $Ah_W(x)$ and a linear subject classifier $Ch_W(x)$. The encoder W, task classifier A, and initial subject classifier C are Gaussian initialized. Both classifiers use squared losses with one-hot labels. For T-step fitted subject classifier C_T, $T \ge 0$, with fixed W and A, there is

Theorem 1 (Concentration under strong individual variation)**.** *Under the latent additive input generation model and assumptions above, there are*

$$|b_k| = \Theta(1) \quad (1 \le k \le q), \qquad \max_{k>q} |b_k| = O(\rho^{-1}). \tag{4}$$

Consequently, for $1 \le k \le q$, $|b_k|$ exceeds every remaining-mode projection magnitude for sufficiently large ρ.

The theorem shows that under the stated conditions, when the weakest subject direction has sufficiently large variance, the normalized subject-gradient projections onto the task-gradient tail become small, while the head projections remain nonzero. The theorem provides a sufficient mechanism for the observed head concentration. And the latent additive input model offers a stylized description of class-dependent and subject-dependent variation in EEG. More details and proof can be found in the appendix.

4 Adaptive Gradient Spectral Scaling

Our analyses suggest an EEG optimizer design principle: it should control excessive spectral concentration, attenuate leading modes in proportion to their conflict with a subject-invariance sur-

rogate, and further avoid indiscriminately amplifying the variance-sensitive tail. This motivates AGraSS to control the strengths of gradient components through spectral scaling.

4.1 KEY INSIGHT

To translate this principle into an update rule, AGraSS operates in the singular basis of the weight gradient, allowing component strengths to be adjusted while preserving their associated directions.

For the weight gradient G, the original singular value decomposition

$$G = U\Sigma V^\top = \sum_{k=1}^{r} \sigma_k u_k v_k^\top. \tag{5}$$

This decomposition separates each singular component $u_k v_k^\top$ from its strength σ_k. AGraSS reshapes these strengths to construct

$$\widehat{G} = U\widetilde{\Sigma}V^\top = \sum_{k=1}^{r} \widehat{\sigma}_k u_k v_k^\top. \tag{6}$$

Specifically, we rescale an automatically identified spectral head to reduce its disproportionate dominance, and apply additional soft shrinkage at leading indices exhibiting consistent subject-related conflict during calibration. The residual tail is preserved during spectral shaping, avoiding tail-specific amplification of noise-sensitive components. Specifically, the shaped singular coefficients are

$$\widehat{\sigma}_k = \begin{cases} a_k \sigma_{\mathrm{res}}, & 1 \le k \le K_c, \\ \sigma_{\mathrm{res}}, & K_c < k \le K_m. \\ \sigma_k, & k > K_m. \end{cases} \tag{7}$$

where σ_{res} is the root-mean-square spectral scale of the residual spectral components, a_k is the subject-aware scaling coefficient, K_m and K_c are the gradient spectral head ratio and subject-aware gradient spectral ratio respectively. The calculation details will be introduced in following subsections.

AGraSS consists of two stages: automatic ratio detection and iterative optimization. Section 4.2 determines the spectral-head ratio and the subject-aware scaling ratios used in subsequent updates. Section 4.3 describes the optimization procedure, including power-iteration-based head estimation, spectral shrinkage, and parameter updates, followed by an analysis of computational costs.

4.2 AUTOMATIC RATIO SELECTION BEFORE TRAINING

For computational efficiency and practical feasibility, AGraSS selects its gradient spectral ratios once before formal optimization, rather than repeatedly estimating them throughout training. In this stage, AGraSS automatically determines two sets of ratios from training data: the spectral-head ratio, which controls how many leading modes are rescaled, and the subject-aware scaling ratios, which determine their additional attenuation.

4.2.1 AUTOMATIC GRADIENT SPECTRAL HEAD DETECTION

In this subsection, we introduce how the formal optimizer automatically determines *how many* components belong to the gradient spectral head.

As gradient spectra from different layers differ in both scale and rank, directly averaging their raw singular values would be dominated by large-scale matrices. We therefore conduct the normalized log transform of the singular values. Additionally, since the matrices may have different gradient spectral ranks, each normalized curve is interpolated onto a common relative-index grid $\in [0, 1]$. To obtain a robust spectrum representative of all routed matrices, we compute the pointwise median. The gradient spectral head is expected to correspond to the initial rapidly-decaying portion of the representative spectrum. We therefore locate its endpoint using the elbow with the largest perpendicular deviation from the endpoint chord. For each matrix weight, the selected relative ratio is converted into a matrix-specific integer head rank.

Algorithm 1 Automatic Ratio Selection

Input: Interpolated normalized initial log-spectra $\{y_m\}$; calibrated statistics $\{b_{m,k}, b_{m,k}, \sigma_{m,k}\}$
Output: Head ranks $\{K_m\}$, subject-aware rank K_c, coefficients $\{a_k\}_{k=1}^{K_c}$
$\bar{y}(x) \leftarrow \mathrm{median}_{m \in \mathcal{M}_{\mathrm{valid}}}\, y_m(x)$
$\rho^\star \leftarrow \arg\max_{0<x\le\rho_{\max}} \dfrac{1-x-\bar{y}(x)}{\sqrt{2}}$
$K_m \leftarrow \min(\lfloor \rho_{\max} r_m \rfloor, \mathrm{round}(\rho^\star r_m))$
$p_k \leftarrow |\mathcal{M}|^{-1} \sum_{m\in\mathcal{M}} \mathbb{I}[b_{m,k} < 0]$
$P_k \leftarrow \prod_{j=1}^{k} \big(\mathbb{I}[p_j \ge 0.95]\big)$
$K_c \leftarrow \max\left\{k :\ 1 \le k \le \min_{m\in\mathcal{M}} K_m,\ P_k = 1\right\}$
$h_k \leftarrow -\,\mathrm{median}_{m\in\mathcal{M}}\, b_{m,k}$; $a_k \leftarrow (1+h_k^2)^{-1}$, for $k \le K_c$

Algorithm 2 AGraSS Optimization Step

Input: Weights W_{t-1}, momentum-based update M_t, previous right factor $V_{t-1,K}$, head rank K, coefficients $\{a_k\}$, learning rate η
Output: Updated weights W_t and right factor $V_{t,K}$
$(U_{t,K}, \Sigma_{t,K}, V_{t,K}) \leftarrow \mathrm{PowerIteration}(M_t, K, V_{t-1,K})$
$R_{\mathrm{res},t} \leftarrow M_t - U_{t,K}\Sigma_{t,K}V_{t,K}^\top$; $\sigma_{\mathrm{res},t} \leftarrow \sqrt{\frac{\|R_{\mathrm{res},t}\|_F^2}{\max(1, r-K)}}$
$\widehat{\sigma}_{t,k} \leftarrow a_k \sigma_{\mathrm{res},t}$, for $k = 1, \ldots, K$
$\widehat{M}_t \leftarrow R_{\mathrm{res},t} + U_{t,K}\,\mathrm{diag}(\widehat{\sigma}_t) V_{t,K}^\top$
$\eta' \leftarrow \frac{0.2\eta}{\mathrm{RMS}(\widehat{M}_t)+\varepsilon}$
$W_t \leftarrow W_{t-1} - \eta' \widehat{M}_t$

4.2.2 Subject-Aware Gradient Spectral Number Detection

Building on the preceding analysis, we estimate subject-related gradient spectral projections $b_{m,k}$. We seek to prevent optimization from being dominated by updates that favor subject discrimination. Under our sign convention, such updates correspond to negative projections. We therefore identify leading spectral indices with consistently negative projections across matrices $\mathcal{M}$ and apply additional attenuation to these components. Thus we compute $p_k = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} \mathbb{I}[b_{m,k} < 0]$. Then we restrict additional attenuation to a contiguous leading prefix: $P_k = \prod_{j=1}^{k} \left(\mathbb{I}[p_j \ge 0.95]\right)$. The consensus threshold requires negative projections in at least 95% of matrices, preventing a small subset from determining the shared attenuation. We restrict additional attenuation to a contiguous leading prefix within the detected heads. Let $K_c = \max\{k :\ 1 \le k \le \min_{m\in\mathcal{M}} K_m,\ P_k = 1\}$ denote the length of this prefix. Selection therefore stops at the first index that fails cross-matrix sign consensus. For each selected index, we define the conflict magnitude as $h_k = -\,\mathrm{median}_{m\in\mathcal{M}}\, b_{m,k}, 1 \le k \le K_c$.

We determine the additional scaling coefficients by balancing deviation from the baseline coefficients against a conflict-weighted penalty:

$$\min_{\boldsymbol{a}\in[0,1]^{K_c}} \frac{1}{2}\sum_{k=1}^{K_c} \left[(a_k - 1)^2 + h_k^2 a_k^2\right], \tag{8}$$

The first term tries to preserve the baseline spectral scaling, whereas the second penalizes the squared retained conflict score $(h_k a_k)^2$. The objective is separable and strictly convex, yielding $a_k = \frac{1}{1+h_k^2}$. Consequently, larger conflict magnitudes induce stronger attenuation.

4.3 Optimization Procedure

We describe a single matrix update and omit the matrix index. Let M denote the momentum-based update, K its selected head rank, and r its minimum dimension.

The pronounced spectral concentration of EEG gradients motivates targeting the leading components while retaining the remaining update. We estimate the head factors (U_K, Σ_K, V_K) via power iteration, where Σ_K is diagonal. Then we can get the residual tail as

$$R_{\text{res}} = M - U_K \Sigma_K V_K^\top. \tag{9}$$

To moderate head dominance, we anchor the spectral head to a spectrum scale from the residual:

$$\sigma_{\text{res}} = \sqrt{\frac{\|R_{\text{res}}\|_F^2}{\max(1, r-K)}}. \tag{10}$$

The association between gradient spectral head and subject-related variation does not imply that every head component should receive additional attenuation. We therefore combine head rescaling with the calibrated coefficients a_k calculated in last subsection into Eq. 7 to get the final result. Following recent work on matrix-based optimization (Liu et al., 2025), we rescale the resulting update to a target RMS magnitude of 0.2 before applying the learning rate.

For $W \in \mathbb{R}^{m\times n}$, AGraSS has a per-step time complexity of $\mathcal{O}(TmnK)$ with T power iterations. It stores $mn + nK$ optimizer-state entries per matrix, together with $\mathcal{O}(K_c)$ coefficients shared across matrices; for $K \ll \min(m, n)$, this requires less persistent state memory than AdamW's $2mn$ entries per matrix. The one-time ratio-selection stage uses only three calibration epochs, corresponding to 6–10% of the 30–50 epochs used for formal training.

Table 1: Comparison results on optimizers in expert models. Bold and underline denote the best and second-best optimizer per column.

Model	Optimizer	BCIC-IV-2a (4-class)			Mental Arithmetic (2-class)		
		ACC-B	Kappa	F1-W	ACC-B	AUROC	AUC-PR
EEGNet	AdamW	0.579 ± 0.018	0.438 ± 0.023	0.570 ± 0.019	0.559 ± 0.063	**0.765 ± 0.043**	0.556 ± 0.077
	SGD	0.492 ± 0.032	0.322 ± 0.043	0.461 ± 0.038	0.550 ± 0.047	0.713 ± 0.083	0.514 ± 0.108
	Muon	0.570 ± 0.019	0.426 ± 0.025	0.558 ± 0.018	0.565 ± 0.070	0.757 ± 0.085	0.555 ± 0.139
	AGraSS (Ours)	**0.588 ± 0.021**	**0.450 ± 0.027**	**0.577 ± 0.024**	**0.569 ± 0.073**	0.754 ± 0.085	**0.573 ± 0.150**
Conformer	AdamW	0.405 ± 0.090	0.206 ± 0.120	0.338 ± 0.134	0.584 ± 0.088	0.756 ± 0.021	0.603 ± 0.018
	SGD	0.245 ± 0.005	0.007 ± 0.007	0.133 ± 0.028	0.500 ± 0.000	0.615 ± 0.092	0.339 ± 0.073
	Muon	0.500 ± 0.050	0.333 ± 0.066	0.458 ± 0.007	0.508 ± 0.009	0.780 ± 0.021	0.611 ± 0.074
	AGraSS (Ours)	**0.502 ± 0.015**	**0.336 ± 0.020**	**0.477 ± 0.017**	**0.708 ± 0.038**	**0.799 ± 0.016**	**0.644 ± 0.057**
Shallow convnet	AdamW	0.535 ± 0.025	0.380 ± 0.034	0.531 ± 0.031	0.560 ± 0.063	0.777 ± 0.056	0.595 ± 0.101
	SGD	0.452 ± 0.032	0.269 ± 0.043	0.444 ± 0.035	0.579 ± 0.047	0.758 ± 0.077	0.510 ± 0.130
	Muon	0.541 ± 0.020	0.388 ± 0.026	0.542 ± 0.022	0.561 ± 0.046	0.790 ± 0.080	0.589 ± 0.135
	AGraSS (Ours)	**0.549 ± 0.027**	**0.398 ± 0.036**	**0.542 ± 0.043**	**0.585 ± 0.078**	**0.794 ± 0.101**	**0.610 ± 0.153**
CTNet	AdamW	0.573 ± 0.045	0.430 ± 0.060	0.563 ± 0.053	0.519 ± 0.020	0.663 ± 0.015	0.389 ± 0.046
	SGD	0.576 ± 0.023	0.434 ± 0.030	0.570 ± 0.030	0.516 ± 0.017	0.656 ± 0.024	0.377 ± 0.034
	Muon	0.552 ± 0.030	0.403 ± 0.040	0.570 ± 0.036	0.531 ± 0.022	0.663 ± 0.015	0.389 ± 0.047
	AGraSS (Ours)	**0.578 ± 0.022**	**0.437 ± 0.045**	**0.572 ± 0.028**	**0.535 ± 0.019**	**0.667 ± 0.017**	**0.391 ± 0.042**

Model	Optimizer	FACED (9-class)			Mumtaz2016 (2-class)		
		ACC-B	Kappa	F1-W	ACC-B	AUROC	AUC-PR
EEGNet	AdamW	0.310 ± 0.015	0.220 ± 0.017	0.286 ± 0.019	0.867 ± 0.032	0.934 ± 0.021	0.939 ± 0.025
	SGD	0.111 ± 0.000	0.000 ± 0.000	0.030 ± 0.007	0.864 ± 0.057	0.941 ± 0.021	0.949 ± 0.019
	Muon	0.296 ± 0.021	0.207 ± 0.023	0.279 ± 0.021	0.868 ± 0.030	0.945 ± 0.013	0.951 ± 0.015
	AGraSS (Ours)	**0.316 ± 0.013**	**0.226 ± 0.015**	**0.290 ± 0.019**	**0.916 ± 0.024**	**0.961 ± 0.012**	**0.966 ± 0.012**
Conformer	AdamW	0.556 ± 0.009	0.496 ± 0.009	0.557 ± 0.007	0.901 ± 0.016	0.970 ± 0.003	0.972 ± 0.003
	SGD	0.225 ± 0.025	0.128 ± 0.027	0.196 ± 0.043	**0.913 ± 0.004**	0.968 ± 0.006	0.974 ± 0.004
	Muon	0.560 ± 0.006	0.501 ± 0.007	**0.560 ± 0.006**	0.865 ± 0.048	0.965 ± 0.009	0.969 ± 0.007
	AGraSS (Ours)	**0.561 ± 0.010**	**0.502 ± 0.011**	0.559 ± 0.011	0.902 ± 0.027	**0.971 ± 0.006**	**0.975 ± 0.006**
Shallow convnet	AdamW	0.394 ± 0.015	0.318 ± 0.016	0.399 ± 0.012	0.911 ± 0.006	0.973 ± 0.002	0.977 ± 0.003
	SGD	0.155 ± 0.053	0.050 ± 0.060	0.119 ± 0.081	0.914 ± 0.010	0.969 ± 0.006	0.975 ± 0.005
	Muon	0.413 ± 0.017	0.337 ± 0.019	0.414 ± 0.016	0.908 ± 0.009	0.970 ± 0.004	0.974 ± 0.004
	Ours	**0.416 ± 0.011**	**0.341 ± 0.012**	**0.418 ± 0.011**	**0.915 ± 0.010**	**0.975 ± 0.004**	**0.979 ± 0.003**
CTNet	AdamW	0.232 ± 0.023	0.136 ± 0.026	0.228 ± 0.026	0.907 ± 0.028	0.966 ± 0.008	0.973 ± 0.007
	SGD	0.164 ± 0.022	0.060 ± 0.025	0.108 ± 0.037	0.908 ± 0.015	0.964 ± 0.009	0.970 ± 0.007
	Muon	0.214 ± 0.012	0.116 ± 0.026	0.211 ± 0.025	0.887 ± 0.027	0.950 ± 0.002	0.961 ± 0.002
	AGraSS (Ours)	**0.276 ± 0.021**	**0.183 ± 0.022**	**0.259 ± 0.015**	**0.915 ± 0.016**	**0.967 ± 0.007**	**0.974 ± 0.005**

5 EXPERIMENT

5.1 EXPERIMENTAL SETUP

Dataset. We evaluate the performance of AGraSS on six public EEG datasets spanning motor imagery (BCIC-IV-2a (Brunner et al., 2008)), dementia classification (ADFTD (Miltiadous et al., 2023)), slowing-event classification (TUSL (von Weltin et al., 2017)), emotion recognition (FACED (Chen et al., 2023)), mental workload assessment (MentalArithmetic (Zyma et al., 2019)), and depression detection (Mumtaz (Mumtaz et al., 2017)).

Metrics. For multiclass benchmarks, performance is measured using balanced accuracy (ACC-B), weighted F1 (F1-W), and Cohen's κ. Binary benchmarks are evaluated using ACC-B, the area under the ROC curve (AUROC), and the area under the precision–recall curve (AUC-PR). Results are reported as mean $\pm$ standard deviation over five random seeds.

Baseline. We evaluate the efficacy of AGraSS on both 4 expert models and 3 EEG foundation models. The expert models include EEGNet (Lawhern et al., 2018), Conformer (Song et al., 2022b), ShallowConvNet (Schirrmeister et al., 2017) and CTNet(Zhao et al., 2024). The EEG foundation models include REVE(El Ouahidi et al., 2026), LaBraM (Jiang et al., 2024) and CBraMod (Wang et al., 2025). For each model, we compare AGraSS with AdamW(Loshchilov & Hutter, 2017), SGD(Bottou, 2010), and Muon(Jordan et al., 2024). The final model for each task is selected based on its validation set performance, with results reported on the held-out test set.

5.2 MAIN RESULTS

Table 1 shows the performance of different optimizers on various EEG expert models. Our AGraSS shows the advantage across different architectures and achieves superior performance over other optimizers. As shown in Table 2, the benefits also extend to the downtask performance of pretrained EEG foundation models. These results demonstrate broad benefits across both task-specific architectures and pretrained models, supporting EEG-aware optimizer design as a complementary avenue for improving EEG learning.

Table 2: Comparison results on optimizers in pretrained foundation models. Bold and underline denote the best and second-best optimizer results for each model and metric, respectively.

Model	Optimizer	ADFTD (3-class)			TUSL (3-class)		
		ACC-B	Kappa	F1-W	ACC-B	Kappa	F1-W
REVE	AdamW	0.502 ± 0.004	0.292 ± 0.021	0.507 ± 0.019	0.716 ± 0.066	0.502 ± 0.082	0.655 ± 0.048
	SGD	0.419 ± 0.041	0.037 ± 0.067	0.305 ± 0.050	0.417 ± 0.051	0.101 ± 0.064	0.311 ± 0.088
	Muon	0.507 ± 0.068	0.306 ± 0.086	0.520 ± 0.062	0.738 ± 0.030	0.540 ± 0.052	$\mathbf{0.674 \pm 0.053}$
	AGraSS (Ours)	$\mathbf{0.576 \pm 0.036}$	$\mathbf{0.403 \pm 0.036}$	$\mathbf{0.596 \pm 0.029}$	$\mathbf{0.749 \pm 0.044}$	$\mathbf{0.550 \pm 0.091}$	0.673 ± 0.008
LaBraM	AdamW	0.331 ± 0.027	0.009 ± 0.049	0.329 ± 0.044	0.670 ± 0.004	0.451 ± 0.016	0.624 ± 0.017
	SGD	0.332 ± 0.008	0.001 ± 0.013	0.214 ± 0.054	0.369 ± 0.090	0.044 ± 0.142	0.305 ± 0.070
	Muon	0.306 ± 0.062	0.001 ± 0.106	0.276 ± 0.085	0.576 ± 0.066	0.434 ± 0.096	0.293 ± 0.064
	AGraSS (Ours)	$\mathbf{0.339 \pm 0.022}$	$\mathbf{0.029 \pm 0.044}$	$\mathbf{0.341 \pm 0.037}$	$\mathbf{0.740 \pm 0.033}$	$\mathbf{0.556 \pm 0.050}$	$\mathbf{0.700 \pm 0.034}$
CBraMod	AdamW	0.449 ± 0.056	0.194 ± 0.085	0.473 ± 0.055	0.688 ± 0.133	0.487 ± 0.182	0.629 ± 0.144
	SGD	0.350 ± 0.024	0.032 ± 0.047	0.266 ± 0.058	0.338 ± 0.022	0.012 ± 0.027	0.188 ± 0.102
	Muon	0.448 ± 0.056	0.187 ± 0.084	0.454 ± 0.074	0.701 ± 0.035	0.514 ± 0.054	0.671 ± 0.035
	AGraSS (Ours)	$\mathbf{0.469 \pm 0.009}$	$\mathbf{0.221 \pm 0.013}$	$\mathbf{0.486 \pm 0.015}$	$\mathbf{0.702 \pm 0.022}$	$\mathbf{0.515 \pm 0.031}$	$\mathbf{0.674 \pm 0.018}$

Model	Optimizer	BCIC-IV-2a (4-class)			Mental Arithmetic (2-class)		
		ACC-B	Kappa	F1-W	ACC-B	AUROC	AUC-PR
REVE	AdamW	0.464 ± 0.022	0.285 ± 0.029	0.424 ± 0.026	0.741 ± 0.001	0.835 ± 0.017	0.644 ± 0.037
	SGD	0.264 ± 0.013	0.019 ± 0.018	0.239 ± 0.039	0.514 ± 0.031	0.479 ± 0.059	0.244 ± 0.026
	Muon	0.500 ± 0.023	0.334 ± 0.031	0.474 ± 0.031	0.755 ± 0.049	0.889 ± 0.028	0.788 ± 0.042
	AGraSS (Ours)	$\mathbf{0.517 \pm 0.014}$	$\mathbf{0.356 \pm 0.019}$	$\mathbf{0.500 \pm 0.017}$	$\mathbf{0.788 \pm 0.004}$	$\mathbf{0.901 \pm 0.013}$	$\mathbf{0.790 \pm 0.041}$
LaBraM	AdamW	0.312 ± 0.015	0.083 ± 0.020	0.246 ± 0.029	$\mathbf{0.614 \pm 0.052}$	0.689 ± 0.011	0.344 ± 0.011
	SGD	0.251 ± 0.005	0.001 ± 0.007	0.121 ± 0.022	0.494 ± 0.011	0.490 ± 0.105	0.251 ± 0.046
	Muon	0.288 ± 0.018	0.050 ± 0.024	0.211 ± 0.034	0.515 ± 0.041	0.668 ± 0.085	0.328 ± 0.033
	AGraSS (Ours)	$\mathbf{0.315 \pm 0.011}$	$\mathbf{0.087 \pm 0.014}$	$\mathbf{0.285 \pm 0.025}$	0.595 ± 0.041	$\mathbf{0.691 \pm 0.049}$	$\mathbf{0.368 \pm 0.011}$
CBraMod	AdamW	0.320 ± 0.020	0.094 ± 0.027	0.240 ± 0.035	0.682 ± 0.038	0.746 ± 0.023	0.478 ± 0.024
	SGD	0.253 ± 0.013	0.005 ± 0.018	0.116 ± 0.022	0.500 ± 0.000	0.409 ± 0.021	0.208 ± 0.009
	Muon	0.338 ± 0.022	0.118 ± 0.029	0.265 ± 0.036	0.665 ± 0.042	0.740 ± 0.051	0.486 ± 0.056
	AGraSS (Ours)	$\mathbf{0.344 \pm 0.012}$	$\mathbf{0.125 \pm 0.016}$	$\mathbf{0.264 \pm 0.028}$	$\mathbf{0.709 \pm 0.032}$	$\mathbf{0.782 \pm 0.015}$	$\mathbf{0.508 \pm 0.034}$

Table 3: Ablation study of AGraSS components using REVE on Mental Arithmetic and EEGNet on BCIC-IV-2a. The best result for each metric is highlighted in bold.

Head shrinkage	Subject-aware shrinkage	REVE+Mental Arithmetic			EEGNet+BCIC-IV-2a		
		ACC-B	AUROC	AUC-PR	ACC-B	Kappa	F1-W
None	✗	0.741 ± 0.001	0.835 ± 0.017	0.644 ± 0.037	0.579 ± 0.018	0.438 ± 0.023	0.570 ± 0.019
Fixed 10%	✗	0.695 ± 0.070	0.850 ± 0.034	0.728 ± 0.056	0.582 ± 0.018	0.443 ± 0.023	0.571 ± 0.018
Adaptive	✗	0.770 ± 0.050	0.873 ± 0.032	0.739 ± 0.066	0.584 ± 0.022	0.445 ± 0.030	0.575 ± 0.022
Adaptive	✓	$\mathbf{0.788 \pm 0.004}$	$\mathbf{0.901 \pm 0.013}$	$\mathbf{0.790 \pm 0.041}$	$\mathbf{0.588 \pm 0.021}$	$\mathbf{0.450 \pm 0.027}$	$\mathbf{0.577 \pm 0.024}$

5.3 Ablation study

Table 3 examines the contributions of automatic head selection and subject-aware shrinkage in AGraSS. Adaptive head selection outperforms both no head shrinkage and fixed-ratio shrinkage across reported metrics in both settings. Notably, fixing the head ratio may decrease the performance, as a single prescribed ratio may not suit different model–dataset settings. These results support adapting the spectral head extent to the observed gradient spectrum. With automatic head selection held fixed, subject-aware shrinkage provides additional gains, supporting targeted attenuation of leading components identified as exhibiting subject-related conflict.

6 Related works

Optimization for EEG. Representative EEG models, including EEGNet (Lawhern et al., 2018), LaBraM (Jiang et al., 2024), EEGPT (Wang et al., 2024), CSBrain(Zhou et al., 2026), and CBraMod (Wang et al., 2025), use general-purpose Adam-family optimizers while advancing architectures and pretrained representations. Despite these advances, broadly applicable optimizer design informed by EEG-specific gradient structure remains comparatively underexplored. Existing EEG-oriented optimization efforts instead address diverse aspects of learning, including manifold geometry, training objectives, and gradient coordination: AlphaGrad (Chaisaen et al., 2025) adaptively balances multiple learning objectives for motor imagery classification task, GEEGA (Angkan et al., 2026) aligns gradients across EEG representations. Patient-adversarial learning (Zhang et al., 2024) encourages patient-invariant representations for seizure detection, while distributionally robust decoding (Duan et al., 2023) addresses distribution shifts through Wasserstein-gradient-flow-based data evolution. Manifold-SimCLR (Tibermacine et al., 2026) adapts Adam-style updates to manifold geometry for EEG contrastive learning, whereas PLSRbiGr (Yin et al., 2022) optimizes partial least-squares regression on a bi-Grassmann manifold. AGraSS focuses on a different aspect: an EEG-informed parameter-update rule applicable across neural architectures and tasks. This optimizer-level design is potentially complementary to objective- and data-level optimization strategies.

Matrix-structured and spectral optimization. AdamW combines coordinate-wise adaptive updates with decoupled weight decay, whereas matrix-aware methods such as Shampoo (Gupta et al., 2018) exploit structured gradient statistics. More recently, Muon (Jordan et al., 2024) approximately orthogonalizes momentum-derived updates using Newton–Schulz iterations, thereby approximately flattening their singular-value spectrum. Subsequent developments improve large-scale deployment through update-scale calibration (Liu et al., 2025), polar-factor computation through Polar Express(Amsel et al., 2026), attention stability through MuonClip(Team et al., 2025), and neuron-wise adaptivity through NorMuon(Li et al., 2025). These developments demonstrate the value of structured updates, but do not explicitly incorporate EEG characteristics into spectral scaling such as subject-related variation.

7 Conclusion

This paper investigated the spectral structure of EEG training gradients, revealing persistent spectral anisotropy, concentrated subject-related projections in the spectral head, and elevated relative variance in the tail. Our theoretical analysis explains the phenomenon of subject aware gradient spectral head under a latent EEG model. Building on these findings, AGraSS combines automatic spectral-head detection with subject-aware soft shrinkage, moderating dominant updates while preserving the residual tail during spectral shaping. Experiments across six EEG datasets with task-specific networks and pretrained foundation models demonstrate broad performance benefits. Together, these results highlight the value of incorporating EEG-specific gradient structure into optimizer design.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing, assist with method implementation and refine mathematical proofs. We did not use generative AI tools to the design of research methodologies and experiments, or to support qualitative or thematic data analysis, interpret research results, the remaining disclosure categories are not applicable to this work. We also used generative AI to write code that is used to generate figures and plots. All generated code and figures created by those code were carefully reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, code, figures, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed experimental settings in the main manuscript and in the appendix. We plan to release the code for the experiments as well as related data and results after the review process concludes.

REFERENCES

Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. In *International Conference on Learning Representations*, volume 2026, pp. 138323–138360, 2026.

Prithila Angkan, Amin Jalali, Paul Hungler, and Ali Etemad. Graph-based learning of spectro-topographical eeg representations with gradient alignment for brain-computer interfaces. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3946–3950. IEEE, 2026.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pp. 177–186. Springer, 2010.

Clemens Brunner, Robert Leeb, Gernot Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. Bci competition 2008–graz data set a. *Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology*, 16(1-6):1, 2008.

Rattanaphon Chaisaen, Phairot Autthasan, Apiwat Ditthapron, and Theerawit Wilaiprasitporn. Alphagrad: Normalized gradient descent for adaptive multi-loss functions in eeg-based motor imagery classification. *IEEE Journal of Biomedical and Health Informatics*, 29(10):7116–7128, 2025.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing eeg dataset. *Scientific Data*, 10(1):740, 2023.

Yi Ding, Yong Li, Hao Sun, Rui Liu, Chengxuan Tong, Chenyu Liu, Xinliang Zhou, and Cuntai Guan. Eeg-deformer: A dense convolutional transformer for brain-computer interfaces. *IEEE Journal of Biomedical and Health Informatics*, 29(3):1909–1918, 2024.

Tiehang Duan, Zhenyi Wang, Gianfranco Doretto, Fang Li, Cui Tao, and Donald Adjeroh. Distributionally robust cross subject eeg decoding. In *ECAI 2023: 26th European Conference on Artificial Intelligence, September 30–October 4, 2023, Kraków, Poland–Including 12th Conference on Prestigious Applications of Intelligent Systems (PAIS 2023)*, pp. 575–582. SAGE Publications 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2023.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. Reve: A foundation model for eeg-adapting to any setup with large-scale pretraining on 25,000 subjects. *Advances in Neural Information Processing Systems*, 38:22541–22577, 2026.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Wei-Bang Jiang, Liming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous eeg data in bci. In *International Conference on Learning Representations*, volume 2024, pp. 16405–16426, 2024.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *International Conference on Machine Learning*, 2025.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representations*, 2017.

Andreas Miltiadous, Katerina D Tzimourta, Theodora Afrantou, Panagiotis Ioannidis, Nikolaos Grigoriadis, Dimitrios G Tsalikakis, Pantelis Angelidis, Markos G Tsipouras, Euripidis Glavas, Nikolaos Giannakeas, et al. A dataset of scalp eeg recordings of alzheimer's disease, frontotemporal dementia and healthy subjects from routine eeg. *Data*, 8(6):95, 2023.

Wajid Mumtaz, Likun Xia, Syed Saad Azhar Ali, Mohd Azhar Mohd Yasin, Muhammad Hussain, and Aamir Saeed Malik. Electroencephalogram (eeg)-based computer-aided technique to diagnose major depressive disorder (mdd). *Biomedical Signal Processing and Control*, 31:108–115, 2017.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human brain mapping*, 38(11):5391–5420, 2017.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022a.

Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:710–719, 2022b.

Kimi Team, Yifan Bai, Yiping Bao, Y Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Imad Eddine Tibermacine, Samuele Russo, Jacopo Tamarri, Ahmed Tibermacine, and Christian Napoli. Manifold-simclr: Self-supervised contrastive representation learning on riemannian eeg covariance manifolds. *Neurocomputing*, pp. 134096, 2026.

Eva von Weltin, Tameem Ahsan, Vinit Shah, Dawer Jamshed, Meysam Golmohammadi, Iyad Obeid, and Joseph Picone. Electroencephalographic slowing: A primary source of error in automatic seizure detection. In *2017 IEEE signal processing in medicine and biology symposium (SPMB)*, pp. 1–5. IEEE, 2017.

Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. Cbramod: A criss-cross brain foundation model for eeg decoding. In *International conference on learning representations*, volume 2025, pp. 75310–75346, 2025.

Wei Xiong, Jiangtong Li, Jie Li, Kun Zhu, and Changjun Jiang. Eeg-fm-bench: A comprehensive benchmark for the systematic evaluation and diagnostic analyses of eeg foundation models. *International Conference on Machine Learning*, 2025.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Wanguang Yin, Zhichao Liang, Jianguo Zhang, and Quanying Liu. Partial least square regression via three-factor svd-type manifold optimization for eeg decoding. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pp. 778–787. Springer, 2022.

Zongpeng Zhang, Taoyun Ji, Mingqing Xiao, Wen Wang, Guojing Yu, Tong Lin, Yuwu Jiang, Xiaohua Zhou, and Zhouchen Lin. Cross-patient automatic epileptic seizure detection using patient-adversarial neural networks with spatio-temporal eeg augmentation. *Biomedical Signal Processing and Control*, 89:105664, 2024.

Wei Zhao, Xiaolu Jiang, Baocan Zhang, Shixiao Xiao, and Sujun Weng. Ctnet: a convolutional transformer network for eeg-based motor imagery classification. *Scientific reports*, 14(1):20237, 2024.

Yuchen Zhou, Jiamin Wu, Zichen Ren, Zhouheng Yao, Weiheng Lu, Kunyu Peng, Qihao Zheng, Chunfeng Song, Wanli Ouyang, and Chao Gou. Csbrain: A cross-scale spatiotemporal brain foundation model for eeg decoding. *Advances in Neural Information Processing Systems*, 38: 87150–87195, 2026.

Igor Zyma, Sergii Tukaev, Ivan Seleznov, Ken Kiyono, Anton Popov, Mariia Chernykh, and Oleksii Shpenkov. Electroencephalograms during mental arithmetic task performance. *Data*, 4(1):14, 2019.

A Appendix

A.1 Dataset Details

We evaluate AGraSS on six EEG datasets covering motor imagery, emotion recognition, mental stress detection, depression detection, dementia classification, and slowing-event classification. All EEG signals are resampled to 200 Hz. The tasks and window lengths are summarized below.

1. **Motor imagery classification:** BCIC-IV-2a is a four-class dataset involving imagined movements of the left hand, right hand, feet, and tongue. Both expert-model and EEG foundation-model experiments use 4-second windows.
2. **Dementia classification:** ADFTD comprises three classes: Alzheimer's disease, frontotemporal dementia, and healthy controls. We use 10-second windows.
3. **Slowing-event classification:** TUSL contains three event categories: seizure, slow wave, and background. We use 10-second windows.
4. **Emotion recognition:** FACED is a nine-class emotion recognition dataset. EEG recordings are segmented into 10-second windows.
5. **Mental stress detection:** Mental Arithmetic distinguishes between arithmetic calculation and resting conditions. Expert-model experiments use 5-second windows following the CBraMod preprocessing protocol, whereas EEG foundation-model experiments use 10-second windows following EEG-FM-Bench, where this dataset is named Workload.
6. **Depression detection:** Mumtaz2016 is a binary classification dataset comprising participants with major depressive disorder and healthy controls. EEG recordings are segmented into 5-second windows.

A.2 Data Preprocessing and Cross-Subject Splits

We use two established data-processing protocols: the CBraMod-based pipeline for expert models and EEG-FM-Bench for pretrained EEG foundation models. Within each model-dataset configuration, preprocessing and data partitions are shared across optimizers. Because the two pipelines differ in their input processing, our comparisons concern optimizer performance within each configuration, rather than direct rankings between expert and foundation models.

All downstream splits are cross-subject: recordings, sessions, and windows from the same participant are assigned to a single partition. Training, validation, and test subjects therefore do not overlap. The partitions are kept fixed across training seeds and hyperparameter configurations.

Expert models. We evaluate EEGNet, EEG-Conformer, ShallowConvNet, and CTNet on BCIC-IV-2a, FACED, Mental Arithmetic, and Mumtaz2016. Recordings are resampled to 200 Hz and segmented into non-overlapping windows. For BCIC-IV-2a, subjects A01–A05, A06–A07, and A08–A09 form the training, validation, and test sets, respectively. FACED uses subjects 1–80, 81–100, and 101–123, while Mental Arithmetic uses subjects 1–28, 29–32, and 33–36. For Mumtaz2016, the training set contains 19 controls and 24 patients, the validation set contains 4 controls and 5 patients, and the test set contains 5 controls and 5 patients.

EEG foundation models. We evaluate REVE, LaBraM, and CBraMod using the EEG-FM-Bench pipeline. This pipeline retains model-compatible resampling, channel handling, and amplitude normalization. EEG inputs are resampled to 200 Hz . BCIC-IV-2a uses the same subject IDs listed above. The target training/validation/test ratios are 70/15/15 for ADFTD, 80/10/10 for TUSL, and 72/14/14 for Mental Arithmetic. Subjects remain indivisible during partitioning.

A.3 Training and Hyperparameter Selection

Training settings. Expert models are trained for 50 epochs with a cosine learning-rate schedule. EEG foundation models are fine-tuned in the single-task, full-parameter setting for 30 epochs with a batch size of 128 across all compared optimizers. Linear warm-up over the first three epochs is followed by cosine decay. Backbone and classifier learning rates are controlled separately through an encoder learning-rate multiplier.

Hyperparameter selection. We perform the same learning-rate grid search for all methods. For each model–dataset–optimizer configuration, hyperparameters are selected using validation performance. The model reported for each run is selected on the validation set, and all test metrics are computed from that model. Neither test performance nor test-subject information is used to select hyperparameters, checkpoints, or subject-aware scaling coefficients.

Computational cost. The one-time ratio-selection stage of AGraSS uses only three calibration epochs, corresponding to 6–10% of the 30–50 epochs used for formal training.

A.4 METHOD DETAILS

Automatic Gradient Spectral Head Detection Let G_m denote the gradient of the m-th weight gradient, the singular values satisfy $\sigma_{m,1} \geq \cdots \geq \sigma_{m,r_m} \geq 0, r_m = \min\{\mathrm{rows}(G_m), \mathrm{cols}(G_m)\}$, As gradient spectra from different layers differ in both scale and rank, directly averaging their raw singular values would be dominated by large-scale matrices. We therefore conduct the normalized log transform: $y_m(k) = \frac{\log(\max(\sigma_{m,k},\epsilon_m)) - \log(\max(\sigma_{m,r_m},\epsilon_m))}{\log(\max(\sigma_{m,1},\epsilon_m)) - \log(\max(\sigma_{m,r_m},\epsilon_m))}$, where $\epsilon_m > 0$ is a scale-aware numerical floor.

Since the matrices may have different gradient spectral ranks r_m, each normalized curve is interpolated onto a common relative-index grid $x \in [0, 1]$. To obtain a robust spectrum representative of all routed matrices, we compute the pointwise median $\overline{y}(x) = \mathrm{median}_{m \in \mathcal{M}_{\mathrm{valid}}} y_m(x)$,

The gradient spectral head is expected to correspond to the initial rapidly-decaying portion of the representative spectrum. We therefore locate its endpoint using the elbow with the largest perpendicular deviation from the endpoint chord $y = 1 - x$: $\rho^\star = \arg\max_{0 < x \leq \rho_{\max}} \frac{1 - x - \overline{y}(x)}{\sqrt{2}}$.

Because different matrices have different gradient spectral ranks, the shared relative ratio $\rho^\star$ then is converted into a matrix-specific integer head rank. For matrix m, we define $K_m = \min\left(\lfloor \rho_{\max} r_m \rfloor, \mathrm{round}\left(\rho^\star r_m\right)\right)$.

Subject-Aware Gradient Spectral Ratio Detection Subject gradients are averaged within each subject and then equally across subjects to account for unequal sample counts. We then identify consistently conflicting leading modes across matrices and attenuate them according to their normalized projection magnitudes.

To prevent unequal sample counts from dominating the calibration statistics, we average subject-classification gradients within each subject and then equally across subjects. At a fixed calibration state, let $\mathcal{D}_s$ denote the calibration samples from subject s, with $N_s = |\mathcal{D}_s|$. For each analyzed matrix $W^{(m)}$, we compute $G_{\mathrm{inv}}^{(m)} = -\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{1}{N_s} \sum_{i \in \mathcal{D}_s} \nabla_{W^{(m)}} \ell_S(x_i, s)$. This weighting is used for gradient estimation; the subject probe itself is fitted using the sample-mean classification loss. The task gradient retains sample-wise averaging.

Let $\mathcal{M}$ denote the set of matrices subject to gradient spectral shaping, and let K_m be the head rank selected for matrix m. We apply shared coefficients only within the common head $K_c = \min_{m \in \mathcal{M}} K_m$. For $k \leq K_c$, the fraction of matrices with negative projection is $p_k = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \mathbb{I}[b_{m,k} < 0]$.

Motivated by the concentration of conflict in the gradient spectral head, we restrict additional attenuation to a contiguous leading prefix: $P_k = \prod_{j=1}^{k} \left(\mathbb{I}[p_j \geq 0.95]\right)$. Thus, selection stops at the first index that fails either cross-matrix sign consensus or numerical validity. We summarize the conflict magnitude using the median normalized projection: $h_k = -\mathrm{median}_{m \in \mathcal{M}} b_{m,k}$.

We balance fidelity to the baseline update against a conflict-weighted penalty on the retained coefficients:

$$\min_{\boldsymbol{a} \in [0,1]^{K_c}} \frac{1}{2} \sum_{k=1}^{K_c} \left[(a_k - 1)^2 + h_k^2 a_k^2\right], \tag{11}$$

The objective is separable and strictly convex, yielding $a_k = \frac{1}{1 + h_k^2}$. Consequently, modes without qualifying conflict retain $a_k = 1$, whereas larger conflict magnitudes induce stronger attenuation.

Power Iteration SVD For fused query-key-value weights, the transformation is applied separately to the query, key, and value matrices. The resulting matrices are concatenated before computing one RMS value for the original stored parameter:

$$\mathrm{RMS}(\widehat{Z}_t) = \frac{\|\widehat{Z}_t\|_F}{\sqrt{\mathrm{numel}(\widehat{Z}_t)}}. \tag{12}$$

The adjusted step size is

$$\eta_t^{\mathrm{adj}} = \eta_t \frac{\rho_{\mathrm{rms}}}{\mathrm{RMS}(\widehat{Z}_t) + \epsilon}, \tag{13}$$

and the matrix update is

$$W_{t+1}^{(m)} = (1 - \eta_t \lambda)\, W_t^{(m)} - \eta_t^{\mathrm{adj}} \widehat{Z}_t^{(m)}. \tag{14}$$

The attenuation coefficients a_k are therefore relative, pre-normalization coefficients. RMS normalization rescales the complete matrix but does not equalize individual tail modes.

A.5 Numerical Stabilization and Valid-Spectrum Filtering

Let $\sigma_{m,1} \geq \cdots \geq \sigma_{m,r_m} \geq 0$ denote the singular values of the m-th routed gradient matrix. To avoid taking the logarithm of zero while respecting the scale of each matrix, we define the matrix-specific numerical floor

$$\epsilon_m := \max\left\{10^{-12}\sigma_{m,1}, \epsilon_{\mathrm{fp}}\right\}, \tag{15}$$

where ϵ_{fp} denotes the smallest positive normal floating-point value used by the implementation. Thus, ϵ_m is not a tunable hyperparameter: its first term scales with the leading singular value of matrix m, while ϵ_{fp} prevents numerical underflow.

We define the stabilized log-spectral range of matrix m as

$$\Delta_m := \log\left(\max\{\sigma_{m,1}, \epsilon_m\}\right) - \log\left(\max\{\sigma_{m,r_m}, \epsilon_m\}\right). \tag{16}$$

The set of valid spectra used to construct the representative spectrum is then

$$\mathcal{M}_{\mathrm{valid}} := \left\{m \in \mathcal{M} \;:\; \sigma_{m,1} > 0 \;\wedge\; \Delta_m > 10^{-12}\right\}, \tag{17}$$

where $\mathcal{M}$ denotes the set of all routed gradient matrices. This criterion excludes zero spectra and spectrally degenerate matrices for which the normalized log-spectrum would have a zero or numerically negligible denominator.

Only curves from $\mathcal{M}_{\mathrm{valid}}$ are interpolated and included in the pointwise median. If $\mathcal{M}_{\mathrm{valid}} = \varnothing$, or if the maximum elbow deviation does not exceed 10^{-12}, the implementation uses the fallback head ratio $\rho_{\mathrm{fb}} = 0.1$.

A.6 Computational Cost and Inference

Calibration trains a disposable model copy for E_{cal} complete passes over the training set and performs one additional full-data gradient pass. It temporarily stores a second model, its AdamW states, and the subject probe. These objects are released before formal training.

Formal optimization performs a one-time singular-value computation for automatic head detection. Subsequent iterations use cached low-rank power iteration. For an $m \times n$ matrix with selected head rank K, each power iteration costs $\mathcal{O}(mnK)$. Applying the precomputed conflict coefficients adds only $\mathcal{O}(K_c)$ scalar operations per matrix.

Calibration is performed independently for each random seed and can be reused across learning-rate configurations initialized from that seed. Subject labels are not required after calibration, and AGraSS introduces no inference-time parameters or computation.

A.7 Proof of Subject-Gradient Concentration

To prove Theorem 1, we first show that the class-dependent component, noise, and changes in ReLU activation gates contribute only an $O(\rho)$ remainder. Singular-value perturbation then separates the leading q modes from the remaining spectrum. Finally, independent initialization of the subject head ensures nonzero limiting head projections, whereas the tail projections vanish after normalization.

Setting and assumptions. Let N_s be the number of subjects and n_c the number of task classes. We write s and y for their one-hot labels and consider

$$X = \rho Q r_s + \xi, \qquad \xi = \mu_y + \varepsilon, \qquad Q^\top Q = I_q, \tag{18}$$

where $Q \in \mathbb{R}^{d\times q}$ and $1 \le q < r := \min(m, d)$. Write

$$r_s = Rz_s, \qquad R = \mathrm{diag}(c_1, \dots, c_q), \qquad c_j \ge 1, \qquad \min_j c_j = 1,$$

with independent $z_s \sim \mathcal{N}(0, I_q)$. Then $a_s = \rho r_s$ has coordinate standard deviations $\sigma^j_{\mathrm{sub}} = \rho c_j$, so that $\rho = \min_j \sigma^j_{\mathrm{sub}}$.

The encoder and linear classifiers are

$$h_W(x) = \mathrm{ReLU}(Wx), \qquad W \in \mathbb{R}^{m\times d}, \quad A \in \mathbb{R}^{n_c\times m}, \quad C \in \mathbb{R}^{N_s\times m}.$$

Their entries are independently initialized as

$$W_{ij} \sim \mathcal{N}(0, \sigma_W^2/d), \qquad A_{ij} \sim \mathcal{N}(0, \sigma_A^2/m), \qquad C_{ij} \sim \mathcal{N}(0, \sigma_C^2/m),$$

with fixed positive scales, independently of the subject vectors. All gradients are evaluated at this initialization; the parameters remain fixed as ρ varies.

Define

$$\begin{aligned} \mathcal{L}_T &= \frac{1}{2}\mathbb{E}\|Ah_W(X) - y\|_2^2, & G_T &= \nabla_W \mathcal{L}_T, \\ \mathcal{L}_S &= \frac{1}{2}\mathbb{E}\|Ch_W(X) - s\|_2^2, & G_{\mathrm{inv}} &= -\nabla_W \mathcal{L}_S. \end{aligned} \tag{19}$$

Let $(u_k, \sigma_k, v_k)_{k=1}^r$ be singular triplets of the full matrix G_T, ordered by decreasing singular value, including zero singular values when necessary. We study

$$b_k = \frac{u_k^\top G_{\mathrm{inv}} v_k}{\|G_{\mathrm{inv}}\|_F + \epsilon_b}, \qquad \epsilon_b > 0, \tag{20}$$

where ϵ_b is fixed independently of ρ.

For a matrix D with m columns, define

$$\begin{aligned} h_s &= \mathrm{ReLU}(WQr_s), \qquad d_s = \mathbf{1}_{\{WQr_s>0\}}, \\ \mathcal{K}_R(D) &= \frac{1}{N_s}\sum_{s=1}^{N_s} \big[(D^\top D h_s) \odot d_s\big] r_s^\top. \end{aligned} \tag{21}$$

Set $K_T = \mathcal{K}_R(A)$ and $K_S = \mathcal{K}_R(C)$. The nondegeneracy condition used in the theorem is

$$\mathrm{rank}(K_T) = q, \qquad \kappa_1 > \cdots > \kappa_q > 0, \tag{22}$$

where $\kappa_1, \dots, \kappa_q$ are the singular values of K_T. This condition requires $N_s \ge q$ and is not implied by $Q^\top Q = I_q$ or random initialization alone. All subsequent statements are conditional on this condition. Constants in asymptotic bounds may depend on the fixed subject vectors and initialization, but not on ρ.

Lemma 1. *For almost every realization of W and the standardized subject vectors,*

$$\begin{aligned} G_T &= \rho^2 K_T Q^\top + E_T, \\ G_{\mathrm{inv}} &= -\rho^2 K_S Q^\top + E_S, \end{aligned} \qquad \|E_T\|_F + \|E_S\|_F = O(\rho). \tag{23}$$

Proof. For a head D and its one-hot target ℓ, write

$$g(x) = \mathbf{1}_{\{Wx>0\}}, \qquad F_D(x, \ell) = \big[D^\top(Dh_W(x) - \ell) \odot g(x)\big] x^\top.$$

Differentiating the squared loss gives $G_T = \mathbb{E}F_A(X, y)$ and $G_{\mathrm{inv}} = -\mathbb{E}F_C(X, s)$. Differentiation under the expectation is justified by finite second moments and almost surely nonzero preactivations.

Let $X_s = \rho Q r_s$. Positive homogeneity gives $h_W(X_s) = \rho h_s$ and $g(X_s) = d_s$, hence

$$\mathbb{E}F_D(X_s, \ell) = \rho^2 \mathcal{K}_R(D) Q^\top - \rho\, \mathbb{E}\big[(D^\top \ell \odot d_s) r_s^\top\big] Q^\top. \tag{24}$$

The second term is $O(\rho)$ in Frobenius norm. It remains to bound the effect of replacing X_s by $X = X_s + \xi$.

Let $w_j^\top$ denote row j of W and define

$$\delta_* = \min_{s,j} |w_j^\top Q r_s|, \qquad \omega = \|W\|_2.$$

There are finitely many subjects and hidden units. Since the subject vectors are nonzero almost surely and W has an independent Gaussian initialization, $\delta_* > 0$ almost surely.

A change in any activation gate requires a perturbation at least as large as its subject-only preactivation. Thus, for $\mathcal{F}_\rho = \{g(X) \neq d_s\}$,

$$\mathcal{F}_\rho \subseteq \{\omega \|\xi\|_2 \geq \rho \delta_*\}, \qquad \mathbb{P}(\mathcal{F}_\rho) \leq \frac{\omega\, \mathbb{E}\|\xi\|_2}{\rho \delta_*} = O(\rho^{-1}). \tag{25}$$

To compare the gradients, first hold the gate at d_s and change the input and hidden activation, then account for the gate change. Using $\|h_W(X) - h_W(X_s)\|_2 \leq \omega \|\xi\|_2$, $\|X_s\|_2 = O(\rho)$, and $\|\ell\|_2 = 1$, the resulting outer-product expansion yields, for $\rho \geq 1$,

$$\|F_D(X, \ell) - F_D(X_s, \ell)\|_F \leq L_D \left(\rho \|\xi\|_2 + \|\xi\|_2^2 + \rho^2 \mathbf{1}_{\mathcal{F}_\rho}\right),$$

where L_D is independent of ρ and ξ. Taking expectations and applying Eq. 25 gives

$$\mathbb{E}\|F_D(X, \ell) - F_D(X_s, \ell)\|_F = O(\rho).$$

Combining this bound with Eq. 24, first for $(D, \ell) = (A, y)$ and then for $(D, \ell) = (C, s)$, proves Eq. 23. □

Proof of Theorem 1. Fix W, A, and the standardized subject vectors satisfying Eq. 22 and the almost-sure conditions of Lemma 1. Let

$$K_T = \overline{U}\, \mathrm{diag}(\kappa_1, \ldots, \kappa_q) \overline{V}^\top$$

be its compact SVD. Since $Q^\top Q = I_q$, the nonzero singular values of $K_T Q^\top$ are $\kappa_1, \ldots, \kappa_q$. The singular-value perturbation inequality and Eq. 23 therefore imply

$$\begin{aligned} \sigma_k(G_T) &= \rho^2 \kappa_k + O(\rho), \quad 1 \leq k \leq q, \\ \sigma_k(G_T) &= O(\rho), \qquad\qquad q < k \leq r. \end{aligned} \tag{26}$$

Moreover, $\rho^{-2} G_T = K_T Q^\top + O(\rho^{-1})$ in operator norm. Because the nonzero limiting singular values are simple, the leading singular-vector pairs can be chosen so that

$$u_k \longrightarrow \overline{u}_k, \qquad v_k \longrightarrow Q \overline{v}_k, \qquad 1 \leq k \leq q. \tag{27}$$

Simultaneous sign changes of u_k and v_k do not affect b_k.

We next establish that the limiting head projections are nonzero. For each $k \leq q$, define

$$P_k(C) = \overline{u}_k^\top \mathcal{K}_R(C) \overline{v}_k.$$

Conditional on W, A, and the subject vectors, P_k is a quadratic polynomial in the entries of C. To show that it is not identically zero, write the rows of A as $\alpha_1^\top, \ldots, \alpha_{n_c}^\top$. Additivity in $D^\top D$ gives

$$\sum_{i=1}^{n_c} \overline{u}_k^\top \mathcal{K}_R(\alpha_i^\top) \overline{v}_k = \overline{u}_k^\top K_T \overline{v}_k = \kappa_k > 0.$$

At least one summand is nonzero. Placing the corresponding row $\alpha_i^\top$ in one row of C and setting its other rows to zero therefore produces a matrix at which $P_k(C) \neq 0$.

The zero set of a nonzero polynomial has Lebesgue measure zero. Since C has an independent Gaussian density, $P_k(C) \neq 0$ almost surely, simultaneously for all $1 \leq k \leq q$. In particular, $\|K_S\|_F > 0$ almost surely. Using Eq. 23, Eq. 27, and $\|K_S Q^\top\|_F = \|K_S\|_F$, we obtain

$$b_k \longrightarrow \beta_k := -\frac{\overline{u}_k^\top K_S \overline{v}_k}{\|K_S\|_F} \neq 0, \qquad 1 \leq k \leq q. \tag{28}$$

Thus $|b_k| = \Theta(1)$ for every head index.

For the tail, the identity $G_T v_k = \sigma_k(G_T) u_k$ and Eq. 23 imply

$$\rho^2 K_T Q^\top v_k = \sigma_k(G_T) u_k - E_T v_k.$$

Since $\sigma_q(K_T) = \kappa_q > 0$, for every $k > q$,

$$\|Q^\top v_k\|_2 \leq \frac{\sigma_k(G_T) + \|E_T\|_2}{\rho^2 \kappa_q} = O(\rho^{-1}), \tag{29}$$

uniformly over the finitely many tail indices. Consequently,

$$|u_k^\top G_{\text{inv}} v_k| \leq \rho^2 \|K_S\|_2 \|Q^\top v_k\|_2 + \|E_S\|_2 = O(\rho).$$

Meanwhile,

$$\|G_{\text{inv}}\|_F + \epsilon_b = \rho^2 \|K_S\|_F + O(\rho) = \Theta(\rho^2).$$

Dividing gives $\max_{k>q} |b_k| = O(\rho^{-1})$.

Finally, $\min_{k \leq q} |\beta_k| > 0$ because there are finitely many head indices. Eq. 28 and the tail bound therefore yield

$$\min_{1 \leq k \leq q} |b_k| > \max_{k > q} |b_k|$$

for all sufficiently large ρ. □

Extension to Finite Subject-Probe Fitting Let

$$H = \frac{1}{N_s} \sum_{s=1}^{N_s} h_s h_s^\top, \qquad h_s = \text{ReLU}(WQr_s). \tag{30}$$

Choose η independently of C_0 such that $0 < \eta < \|H\|_2^{-1}$. Starting from the Gaussian initialized C_0, perform

$$C_{t+1} = C_t - \frac{\eta}{\rho^2} \nabla_C \mathcal{L}_S(W, C)|_{C=C_t}, \qquad t = 0, \ldots, T-1, \tag{31}$$

with W and A frozen. Here T is finite and independent of ρ. After fitting, G_{inv} and b_k retain their original definitions with $C = C_T$; the derivative with respect to W treats C_T as fixed.

Proposition 1. *Under the preceding assumptions, conditional on W, A, and the standardized subject vectors satisfying the nondegeneracy condition on $K_T = \mathcal{K}_R(A)$, the conclusions of Theorem 1 hold almost surely over C_0 for every fixed finite T.*

Proof. Define

$$H_\rho = \mathbb{E}[h_W(X) h_W(X)^\top], \qquad B_\rho = \mathbb{E}[s\, h_W(X)^\top].$$

The ReLU Lipschitz bound gives

$$\mathbb{E}\|h_W(X) - \rho h_s\|_2^2 = O(1).$$

Consequently,

$$H_\rho = \rho^2 H + O(\rho), \qquad B_\rho = O(\rho). \tag{32}$$

Since $\nabla_C \mathcal{L}_S(W, C) = C H_\rho - B_\rho$, the update is affine:

$$C_{t+1} = C_t L_\rho + \frac{\eta}{\rho^2} B_\rho, \qquad L_\rho = I_m - \frac{\eta}{\rho^2} H_\rho.$$

Let $L = I_m - \eta H$. Eq. 32 implies $L_\rho = L + O(\rho^{-1})$. For fixed T, iterating the update yields

$$\begin{aligned} C_T &= C_0 L_\rho^T + \frac{\eta}{\rho^2} B_\rho \sum_{j=0}^{T-1} L_\rho^j \\ &= \overline{C}_T + O(\rho^{-1}), \qquad \overline{C}_T := C_0 L^T. \end{aligned} \tag{33}$$

Here L^T denotes a matrix power, not a transpose; the sum is zero when $T = 0$.

The condition $0 < \eta < \|H\|_2^{-1}$ makes L positive definite, hence L^T invertible. Conditional on W, A, and the subject vectors, $\overline{C}_T = C_0 L^T$ therefore has an absolutely continuous distribution on $\mathbb{R}^{N_s \times m}$.

The remainder estimate in the gradient-expansion lemma is uniform over bounded classifier matrices. Since C_T is bounded by Eq. 33, it follows that

$$G_{\text{inv}} = -\rho^2 \mathcal{K}_R(C_T) Q^\top + O(\rho).$$

Moreover, $\mathcal{K}_R$ is quadratic and thus locally Lipschitz, so

$$\mathcal{K}_R(C_T) = \mathcal{K}_R(\overline{C}_T) + O(\rho^{-1}).$$

Therefore,

$$G_{\text{inv}} = -\rho^2 \mathcal{K}_R(\overline{C}_T) Q^\top + O(\rho). \tag{34}$$

All gradient remainders above are bounded in Frobenius norm.

Let $\overline{u}_k, \overline{v}_k$ be the singular vectors of K_T used in the initialization proof. That proof establishes that

$$P_k(D) = \overline{u}_k^\top \mathcal{K}_R(D) \overline{v}_k$$

is a nonzero polynomial for every $k \le q$. Absolute continuity of $\overline{C}_T$ therefore gives $P_k(\overline{C}_T) \neq 0$ almost surely, simultaneously for all $k \le q$. In particular, $\|\mathcal{K}_R(\overline{C}_T)\|_F > 0$ almost surely.

Neither W nor A changes during probe fitting. The task-gradient singular-vector limits thus remain

$$u_k \longrightarrow \overline{u}_k, \qquad v_k \longrightarrow Q\overline{v}_k, \qquad k \le q.$$

Combining these limits with Eq. 34 gives

$$b_k \longrightarrow -\frac{\overline{u}_k^\top \mathcal{K}_R(\overline{C}_T) \overline{v}_k}{\|\mathcal{K}_R(\overline{C}_T)\|_F} \neq 0, \qquad k \le q. \tag{35}$$

Hence $|b_k| = \Theta(1)$ for all head indices.

For $k > q$, the unchanged task-gradient expansion gives $\|Q^\top v_k\|_2 = O(\rho^{-1})$ uniformly over tail indices. Eq. 34 then implies

$$|u_k^\top G_{\text{inv}} v_k| \le \rho^2 \|\mathcal{K}_R(\overline{C}_T)\|_2 \|Q^\top v_k\|_2 + O(\rho) = O(\rho).$$

Meanwhile,

$$\|G_{\text{inv}}\|_F + \epsilon_b = \rho^2 \|\mathcal{K}_R(\overline{C}_T)\|_F + O(\rho) = \Theta(\rho^2).$$

Thus $\max_{k>q} |b_k| = O(\rho^{-1})$. Together with the nonzero limits in Eq. 35, this yields the claimed head–tail separation for sufficiently large ρ. □