Rethinking Bregman Divergences in Kronecker-Factored Optimizers

Bing Liu* BING_LIU@ZJU.EDU.CN
College of Control Science and Engineering, Zhejiang University

Wenjie Zhou* ZJ4323005@GMAIL.COM
State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences

Chengcheng Zhao† CHENGCHENGZHAO@ZJU.EDU.CN
College of Control Science and Engineering, Zhejiang University

* Equal contribution † Corresponding author

Abstract

Shampoo-style optimizers approximate gradient covariance matrices using Kronecker-factored structures. Recent work [11] showed that such approximations can be viewed as projections under Bregman matrix divergences, leading to different Kronecker-factored preconditioners. However, it remains unclear what role the choice of divergence plays when the covariance is not exactly Kronecker-factored. We study this question through the spectrum of the covariance matrix. We show that Frobenius, von Neumann, and LogDet divergences distribute the unavoidable Kronecker approximation error differently across the covariance spectrum. We further show that their Kronecker factors are governed by divergence-weighted residuals rather than the raw approximation error, explaining how these spectral preferences are realized in the resulting preconditioners. Empirically, we observe that the top covariance eigenspace is substantially better aligned with the Hessian matrix, while the tail spectrum is much noisier and unreliable. Motivated by these findings, we propose a subspace-aware Kronecker optimizer that applies eigenvalue-based preconditioning in the top subspace and uses an adaptive isotropic acceleration constant in the bottom subspace.

1. Introduction

Shampoo-style optimizers [4] exploit the matrix structure of gradients by replacing dense second-order preconditioners with Kronecker-factored approximations. For a matrix-valued parameter with gradient G, this corresponds to approximating the gradient covariance matrix $C = \mathbb{E}[\mathrm{vec}(G)\mathrm{vec}(G)^\top]$ by a product $L \otimes R$. Lin et al. [11] showed that several bilateral Shampoo variants arise from minimizing Bregman matrix divergences between C and $L \otimes R$. However, this unification does not explain what happens when the covariance is not exactly Kronecker-factored. For a generic dense covariance matrix, nonzero approximation error is unavoidable, so different divergences emphasize different parts of the spectrum. Furthermore, since the covariance significantly differs from the Hessian, meaning that not all its spectrum is suitable for preconditioning [10, 16]. We therefore ask:

Q: When exact Kronecker matching is impossible, which parts of the covariance spectrum should a structured optimizer approximate and trust?

Our work connects two issues: the approximation geometry induced by different Bregman matrix divergences, and the spectral reliability of covariance-based preconditioning. We make this connection through three contributions below:

- **Spectral geometry of Bregman Kronecker approximation.** We characterize the eigenvalue-dependent penalties induced by Frobenius, von Neumann, and LogDet divergences in covariance approximation, and show that the resulting bilateral preconditioners reflect these spectral preferences through divergence-weighted residuals on the Kronecker manifold.
- **Spectral reliability of covariance preconditioning.** We empirically examine the relationship between covariance and Hessian eigenspaces, and observe that top covariance components are substantially more aligned with the Hessian than the lower-eigenvalue components.
- **Subspace-aware Kronecker optimizer.** Motivated by the analysis above, we propose an optimizer that applies eigenvalue-based preconditioning in the top Kronecker eigenspace and uses an adaptive isotropic acceleration constant in the complementary bottom subspace.

2. Preliminaries

2.1. Notations

Let $\Theta \in \mathbb{R}^{m\times n}$ be a matrix-valued parameter and $G = \nabla_\Theta \ell(\Theta) \in \mathbb{R}^{m\times n}$ its stochastic gradient. We write $g = \mathrm{vec}(G) \in \mathbb{R}^{mn}$ and $C := \mathbb{E}[gg^\top] \in \mathbb{S}^{mn}_{++}$ for the gradient covariance matrix, where $\mathbb{S}^{mn}_{++}$ denotes the cone of $mn \times mn$ symmetric positive-definite (SPD) matrices.[1]. We denote the Hessian by $H := \nabla^2_\theta \mathcal{L}(\theta) \in \mathbb{R}^{mn\times mn}$, where $\theta = \mathrm{vec}(\Theta)$. Structured preconditioning approximates C by a Kronecker product $C \approx L \otimes R$, where $L \in \mathbb{S}^m_{++}$ and $R \in \mathbb{S}^n_{++}$. The corresponding preconditioned gradient satisfies $(L \otimes R)^{-1/2} g \iff L^{-1/2} G R^{-1/2}$. For a square matrix A, $\mathrm{Tr}(A)$ and $\det(A)$ denote its trace and determinant. For matrices of the same size, $A \odot B$ denotes the Hadamard product, and $\|A\|_F$ denotes the Frobenius norm. We use standard matrix differential notation: $DF(S)[H]$ denotes the directional derivative of F at S along H, and $\nabla^2 F(S)[H]$ denotes the Hessian operator.

2.2. The original Shampoo algorithm

Shampoo [4] is a matrix-structured optimizer that maintains two marginal second-moment factors, $L = \mathbb{E}[GG^\top]$ and $R = \mathbb{E}[G^\top G]$. In practice, they are updated by EMA: $L_t = \beta_2 L_{t-1} + (1-\beta_2) G_t G_t^\top$ and $R_t = \beta_2 R_{t-1} + (1-\beta_2) G_t^\top G_t$. The update is $\Theta_{t+1} = \Theta_t - \eta L_t^{-p} G_t R_t^{-p}$, with common choices $p = 1/4$ or $p = 1/2$. Thus, Shampoo replaces a dense $mn \times mn$ preconditioner with two smaller factors of sizes $m \times m$ and $n \times n$.

2.3. Bregman matrix divergences

Following [11], we measure the discrepancy between C and $L \otimes R$ using Bregman matrix divergences. Let $F : \mathbb{S}^d_{++} \to \mathbb{R}$ be a strictly convex differentiable generating function. For $X, Y \in \mathbb{S}^d_{++}$,

$$\mathcal{B}_F(X,Y) := F(X) - F(Y) - \mathrm{Tr}\big([\nabla F(Y)](X-Y)\big).$$

Different choices of F induce different notions of covariance approximation.

1. Strictly speaking, C is generally positive semidefinite. Following [11], we consider the positive-definite setting, which can be realized by damping $C + \kappa I$ or by using a sufficiently large batch.

3. Rethinking Bregman Divergences in Kronecker-Factored Approximation

Lin et al. [11] introduced a Bregman-divergence-based view of Kronecker-factored approximation to the gradient covariance matrix, providing a unified interpretation of several structured preconditioners. Specifically, given the covariance matrix $C = \mathbb{E}[gg^\top]$, one seeks a Kronecker-factored SPD approximation $S = L \otimes R$ by solving $\min_{L,R} \mathcal{B}_F(C, L \otimes R)$.

Different generating functions F induce different stationary conditions, and thus derive different bilateral preconditioners. Table 1 summarizes the main cases considered in this paper.

Table 1: Bregman-divergence view of bilateral Kronecker preconditioners.

Method	$F(M)$	**Divergence**	**Stationary conditions**
KL-Shampoo	$-\frac{1}{2}\log\det M$	LogDet / KL	$L^* = \frac{1}{n}\mathbb{E}[G(R^*)^{-1}G^\top]$, $R^* = \frac{1}{m}\mathbb{E}[G^\top(L^*)^{-1}G]$
VN-Shampoo	$\mathrm{Tr}(M\log M - M)$	von Neumann	$L^* = \frac{1}{\mathrm{Tr}(R^*)}\mathbb{E}[GG^\top]$, $R^* = \frac{1}{\mathrm{Tr}(L^*)}\mathbb{E}[G^\top G]$
F-Shampoo	$\frac{1}{2}\mathrm{Tr}(M^\top M)$	Frobenius	$L^* = \frac{1}{\mathrm{Tr}((R^*)^2)}\mathbb{E}[GR^*G^\top]$, $R^* = \frac{1}{\mathrm{Tr}((L^*)^2)}\mathbb{E}[G^\top L^*G]$

Next, we will interpret the different bilateral preconditioners obtained with different divergences from three new perspectives.

3.1. Unavoidable Approximation Error

While these optimizers admit a unified Bregman-divergence view, approximating a dense covariance matrix by two low-dimensional factors generally cannot achieve zero divergence, and hence cannot yield perfect spectral matching, which is formalized by the following proposition.

Proposition 1 (Strict Positivity of Kronecker Approximation Error) *Let $C \in \mathbb{S}^{mn}_{++}$ be a SPD matrix drawn from a distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{S}^{mn}$. Let $\mathcal{M} = \{L \otimes R \mid L \in \mathbb{S}^m_{++}, R \in \mathbb{S}^n_{++}\}$ denote the manifold of Kronecker-factored matrices. Then, $C \notin \mathcal{M}$ almost surely (a.s.). Consequently, for any Bregman matrix divergence $\mathcal{B}_F$, the minimum approximation error remains strictly positive:* $\inf_{L,R} \mathcal{B}_F(C, L \otimes R) > 0 \quad$ *a.s.*.

Given that these Kronecker products cannot perfectly approximate the covariance matrix $\mathbb{E}[gg^\top]$, how do these divergences actually differ in matrix approximation? The following subsection shows the spectral preferences of different divergences.

3.2. Spectral Preferences of Matrix Divergences

According to Proposition 1, exact spectrum matching is impossible for a general dense covariance matrix, since $C \notin \mathcal{M}$. Therefore, different divergences differ not only in the value of the approximation error, but also in how this error is distributed across the spectrum.

To characterize this effect, let $A = U\Lambda U^\top$, $B = V\Omega V^\top$ with eigenvalues $\{\lambda_i\}_{i=1}^d$ and $\{\omega_j\}_{j=1}^d$, and define the eigenspace alignment matrix $P_{ij} := \langle u_i, v_j\rangle^2$. For any spectral generating function F, the corresponding Bregman divergence admits the decomposition

$$\mathcal{B}_F(A, B) = \Psi(\Lambda) + \Phi(\Omega) + \sum_{i,j} g_F(\lambda_i, \omega_j) P_{ij},$$

where the first two terms depend only on marginal spectra, while g_F controls the penalty on cross-spectral mismatch. The proof is given in Appendix B.2. For the divergences studied in this paper, $g_{\text{Frob}} = -\lambda_i\omega_j$, $g_{\text{vN}} = -\lambda_i \log \omega_j$, and $g_{\text{LogDet}} = \lambda_i/\omega_j$. Thus, Frobenius emphasizes large-eigenvalue components, von Neumann retains this preference but weakens it logarithmically, and LogDet is more sensitive to small-eigenvalue components. Table 2 summarizes the differences.

Table 2: Spectral preferences of matrix divergences.

Divergence	**Coupling term g_F**	**Error sensitivity**	**Approximation priority**
Frobenius	$-\lambda_i\omega_j$	absolute, bilinear magnitude	strongest emphasis on top spectrum
von Neumann	$-\lambda_i \log \omega_j$	log-damped magnitude	soft preference for top spectrum
LogDet	λ_i/ω_j	relative ratio distortion	broad spectrum, including small-eigenvalue directions

3.3. Stationary Conditions Align with the Spectral Preferences

In this subsection, we show that the bilateral optimizers induced by the stationary conditions of different divergences align with the spectral preferences in Section 3.2.

Theorem 2 (Divergence-induced geometric bias on the Kronecker manifold) *Let C be a target SPD matrix, and let $\mathcal{M}$ denote the manifold of Kronecker-factored SPD matrices. Suppose (L^*, R^*) is a stationary point of $\min_{L,R} \mathcal{B}_F(C, L \otimes R)$, and let $S^* := L^* \otimes R^* \in \mathcal{M}$. Then, for every feasible first-order perturbation in the tangent space of $\mathcal{M}$ at S^*, i.e., $H \in T_{S^*}\mathcal{M}$,*

$$D_S\mathcal{B}_F(C, S)[H]|_{S=S^*} = \left\langle \nabla^2 F(S^*)[S^* - C],\ H \right\rangle = 0,$$

where $\langle A, B\rangle := \operatorname{Tr}(A^\top B)$. Consequently, we have $\nabla^2 F(S^*)[S^* - C] \perp T_{S^*}\mathcal{M}$.

Theorem 2 shows that a stationary Kronecker approximation is determined not by the raw residual $S^* - C$, but by the divergence-weighted residual $\nabla^2 F(S^*)[S^* - C]$. Hence, different generating functions generate different local error geometries, and the resulting Kronecker factors align with the corresponding spectral preferences given in Section 3.2, as stated in the following corollary.

Corollary 3 *Consider the problem of minimizing $\mathcal{B}_F(C, L \otimes R)$ over Kronecker factors L and R, where $C \in \mathbb{S}^{mn}_{++}$. Let (L^*, R^*) be a stationary point and set $S^* = L^* \otimes R^*$. If the divergence generated by F induces a specific spectral weighting on the approximation error, then S^* is stationary with respect to that weighted geometry. Consequently, different choices of F yield stationary Kronecker factors governed by different weighted geometries, leading to different spectral preferences.*

This is reflected directly in the stationary conditions. Frobenius yields bilinear terms such as $GR^*G^\top$ and $G^\top L^*G$, which emphasize dominant subspaces. LogDet yields inverse-weighted terms such as $G(R^*)^{-1}G^\top$ and $G^\top(L^*)^{-1}G$, which increase sensitivity to small eigenvalues. The von Neumann case is intermediate: it favors dominant directions through trace-normalized moment matching, but avoids the inverse weighting of LogDet. Thus, the stationary equations directly show the spectral preferences of the underlying divergences.

4. On the Covariance and Hessian Matrix

The empirical Fisher, gradient covariance, true Hessian, and exact FIM are generally different [5, 10, 16]. However, prior work shows that the top eigenspaces of the gradient covariance and Hessian can

still exhibit strong alignment [20, 25, 26]. This suggests that the top spectrum of covariance may contain useful curvature information, even when the full spectrum is unreliable.

We verify this by measuring Hessian–covariance eigenspace alignment on a small Transformer trained on SST-2-1k [15]. The results are reported in Table 3; details are given in Appendix C.

Table 3: Hessian–covariance eigenspace alignment on Q/K/V projection blocks.

Metric	Step 100			Step 1000		
	Q	K	V	Q	K	V
Overlap@5	0.825	0.763	0.896	0.804	0.748	0.855
BandOverlap@180-200	0.091	0.102	0.098	0.097	0.105	0.103

5. Combining Bregman Divergence and Covariance Reliability: New Optimizers

Sections 3 and 4 highlight two observations. First, Kronecker factors cannot perfectly match a dense covariance spectrum, with residuals distributed differently across Bregman divergences. Second, the top covariance eigenspace aligns better with the Hessian, making bottom components unreliable. Prior work also showed that full-covariance preconditioning can perform poorly [10, 16]. Motivated by these observations, we propose a subspace-aware Kronecker preconditioner. The method chooses Kronecker factors via a top-sensitive divergence, such as the von Neumann or Frobenius divergence, applies eigenvalue-based scaling only in the top eigenspace and employs an adaptive isotropic acceleration constant in the bottom subspace. This design uses the reliable top spectral information while avoiding using the noisy bottom space directions.

The full procedure of **BregTop** is given in Algorithms 1 and 2. Algorithm 1 maintains divergence-specific Kronecker factors and their eigenspaces, where STATS returns the factor updates (Δ_L, Δ_R) induced by the chosen Bregman divergence. Algorithm 2 then applies the proposed subspace preconditioning rule. The explicit forms of (Δ_L, Δ_R) for different divergences are given in Appendix B.3.

Algorithm 1: Unified subspace-aware optimizer

Require: $G, M, \gamma, \beta_1, \beta_2, T, \rho, q, c$
$M \leftarrow (1-\beta_1)G + \beta_1 M$
$(\Delta_L, \Delta_R) \leftarrow \textsc{Stats}(G, U_L, U_R, \lambda_L, \lambda_R)$
$L \leftarrow (1-\beta_2)\Delta_L + \beta_2 L, \quad R \leftarrow (1-\beta_2)\Delta_R + \beta_2 R$
$\lambda_L \leftarrow (1-\beta_2)\mathrm{diag}(U_L^\top \Delta_L U_L) + \beta_2 \lambda_L$
$\lambda_R \leftarrow (1-\beta_2)\mathrm{diag}(U_R^\top \Delta_R U_R) + \beta_2 \lambda_R$
if iter $\bmod T = 0$ **then**
  $U_L \leftarrow \mathrm{qr}(LU_L), \quad U_R \leftarrow \mathrm{qr}(RU_R)$
end if
$\widehat{M} \leftarrow \textsc{KronPrecond}(U_L, U_R, \lambda_L, \lambda_R, M, \rho, q, c)$
$\theta \leftarrow \theta - \gamma\, \mathrm{vec}(\widehat{M})$

Algorithm 2: Kronecker-based preconditioning

Require: $U_L, U_R, \lambda_L, \lambda_R, M, \rho, q, c$
$\widetilde{M} \leftarrow U_L^\top M U_R, S \leftarrow \lambda_L \lambda_R^\top, K \leftarrow \lceil \rho mn \rceil$
$\Omega_\rho \leftarrow$ indices of the top-K entries of S
$\mathcal{B}_\rho \leftarrow \{(i,j) : (i,j) \notin \Omega_\rho\}$
$\chi_{\rho,q} \leftarrow c \cdot \mathrm{Quantile}_q\{S_{ij}^{-1/2} : (i,j) \in \mathcal{B}_\rho\}$
for each (i,j) **do**
  $W_{ij} \leftarrow \begin{cases} S_{ij}^{-1/2}, & (i,j) \in \Omega_\rho, \\ \chi_{\rho,q}, & (i,j) \in \mathcal{B}_\rho \end{cases}$
end for
$\widetilde{M}_{\mathrm{pre}} \leftarrow \widetilde{M} \odot W, \widehat{M} \leftarrow U_L \widetilde{M}_{\mathrm{pre}} U_R^\top$
return $\widehat{M}$

In Algorithm 2, the retained set Ω_ρ contains the largest $\lceil \rho mn \rceil$ entries of the joint Kronecker spectrum $S = \lambda_L \lambda_R^\top$. The complementary set $\mathcal{B}_\rho$ is treated isotropically using the adaptive constant $\chi_{\rho,q}$, chosen as c times the q-quantile of $\{S_{ij}^{-1/2} : (i,j) \in \mathcal{B}_\rho\}$. Thus, the top subspace uses eigenvalue-based scaling, while the bottom subspace uses a uniform accelerating scale.

The following theorem formalizes the decomposition implemented by Algorithm 2: it applies eigenvalue-based damping in the top subspace and a uniform acceleration in the bottom subspace.

Theorem 4 (Equivalent decomposition of the preconditioner) *Let $L = U_L \Lambda_L U_L^\top$ and $R = U_R \Lambda_R U_R^\top$, and define $S = \lambda_L \lambda_R^\top$. For a retained ratio $\rho \in (0, 1]$, let Ω_ρ contain the indices of the largest $K = \lceil \rho mn \rceil$ entries of S, and let E be its indicator matrix. Define $\mathcal{P}^\rho_{\mathrm{top}}(M) := U_L\big((U_L^\top M U_R) \odot E\big)U_R^\top$. Then Algorithm 2 returns*

$$\widehat{M} = U_L\Big((U_L^\top M U_R) \odot E \odot S^{-1/2}\Big)U_R^\top + \chi_{\rho,q}\big(M - \mathcal{P}^\rho_{\mathrm{top}}(M)\big).$$

Thus, the top Kronecker eigenspace is scaled by its eigenvalue-based inverse curvature, while the orthogonal bottom space is scaled by the adaptive acceleration constant $\chi_{\rho,q}$ given in Algorithm 2.

Remark 5 *While Song et al. [15] showed that learning primarily occurs in the Hessian's non-dominant (bottom) subspace, our work studies the spectral alignment between the Hessian and its Kronecker covariance approximation. The optimizer design is consistent with the training dynamics observed in [15]: preconditioning in the top subspace dampens noisy updates along high-curvature directions, while isotropic acceleration in the bottom subspace preserves the standard gradient flow, which is necessary for the learning progress.*

6. Experiments

We compare Shampoo, KL-Shampoo, VN-Shampoo, F-Shampoo, and BregTop on full SST-2. The model is a 4-layer Transformer with hidden dimension 128 and 8 attention heads. For all methods, Adam is used for the embedding, bias, normalization, and classification-head parameters, while the Shampoo-type optimizer is applied to the remaining matrix weights. More details are given in Appendix D. BregTop-VN and BregTop-F use the algorithm in Section 5 with VN/F-induced Kronecker factors, respectively. We report the number of steps required for the EMA-smoothed training loss to fall below 0.05 with EMA 0.98. No weight decay is used.

Table 4: Steps to EMA-smoothed train loss < 0.05 on full SST-2.

	Shampoo	KL	VN-v1	VN-v2	F-v1	F-v2	B-VN-v1	B-VN-v2	B-F-v1	B-F-v2
Steps	1119	1092	1141	1127	1598	1602	1018	**1003**	1084	1059

BregTop-VN-v2 reaches the target loss fastest. The BregTop variants improve over their corresponding VN/F baselines, supporting the benefit of our algorithm.

7. Conclusion

We studied Bregman-induced Kronecker preconditioning under unavoidable covariance approximation error. Our analysis shows that different divergences impose different spectral preferences, while our empirical results suggest that only the top eigenspace is reliably aligned with the Hessian. Based on this, we proposed a subspace-aware optimizer that trusts the top eigenspace and treats the bottom space with adaptive isotropic scaling. Experiments show that this design improves the step efficiency.

References

[1] Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. ASGO: Adaptive structured gradient optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[2] Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations*, 2023.

[3] Runa Eschenhagen, Aaron Defazio, Tsung-Hsien Lee, Richard E. Turner, and Hao-Jun Michael Shi. Purifying shampoo: Investigating shampoo's heuristics by decomposing its preconditioner. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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

[5] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

[6] 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/.

[7] Priya Kasimbeg, Frank Schneider, Runa Eschenhagen, Juhan Bae, Chandramouli Shama Sastry, Mark Saroufim, BOYUAN FENG, Less Wright, Edward Z. Yang, Zachary Nado, Sourabh Medapati, Philipp Hennig, Michael Rabbat, and George E. Dahl. Accelerating neural network training: An analysis of the algoperf competition. In *The Thirteenth International Conference on Learning Representations*, 2025.

[8] Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. *arXiv preprint arXiv:2510.16981*, 2025.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[10] Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.

[11] Wu Lin, Scott C. Lowe, Felix Dangel, Runa Eschenhagen, Zikun Xu, and Roger Baker Grosse. Understanding and improving shampoo and SOAP via kullback-leibler minimization. In *The Fourteenth International Conference on Learning Representations*, 2026.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[13] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International Conference on Machine Learning*, pages 2408–2417. PMLR, 2015.

[14] Depen Morwani, Itai Shapira, Nikhil Vyas, Sham M Kakade, Lucas Janson, et al. A new perspective on shampoo's preconditioner. In *The Thirteenth International Conference on Learning Representations*, 2025.

[15] Minhak Song, Kwangjun Ahn, and Chulhee Yun. Does sgd really happen in tiny subspaces? In *13th International Conference on Learning Representations*. International Conference on Learning Representations (ICLR), 2025.

[16] Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 3503–3513. PMLR, 2020.

[17] Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and stabilizing shampoo using adam for language modeling. In *The Thirteenth International Conference on Learning Representations*, 2025.

[18] Jinbo Wang, Mingze Wang, Zhanpeng Zhou, Junchi Yan, Lei Wu, et al. The sharpness disparity principle in transformers for accelerating language model pre-training. In *Forty-second International Conference on Machine Learning*, 2026.

[19] 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*, 2026.

[20] Lei Wu, Mingze Wang, and Weijie Su. The alignment property of sgd noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, 35:4680–4693, 2022.

[21] Shuo Xie, Tianhao Wang, Sashank J. Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured preconditioners in adaptive optimization: A unified analysis. In *Forty-second International Conference on Machine Learning*, 2025.

[22] Tian Xie, Haoming Luo, Haoyu Tang, Yiwen Hu, Jason Klein Liu, Qingnan Ren, Yang Wang, Wayne Xin Zhao, Rui Yan, Bing Su, et al. Controlled llm training on spectral sphere. *arXiv preprint arXiv:2601.08393*, 2026.

[23] Chenrui Xu, Wenjing Yan, and Ying-Jun Angela Zhang. Fismo: Fisher-structured momentum-orthogonalized optimizer. *arXiv preprint arXiv:2601.21750*, 2026.

[24] Wenjie Zhou, Bohan Wang, Wei Chen, and Xueqi Cheng. Bsfa: Leveraging the subspace dichotomy to accelerate neural network training. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18845–18860, 2025.

[25] Shuchen Zhu, Rizhen Hu, Mingze Wang, Mou Sun, Xue Wang, Kun Yuan, and Zaiwen Wen. Accelerating llm pre-training through flat-direction dynamics enhancement. *arXiv preprint arXiv:2602.22681*, 2026.

[26] 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 *International Conference on Machine Learning*, pages 7654–7663. PMLR, 2019.

Appendix A. Related Work

Matrix-Based Optimizers. Element-wise optimizers, such as Adam [9, 12], have long dominated large-scale neural network training, but they ignore the natural matrix structure of gradients. This limitation has motivated a growing line of work on matrix-based optimizers, including bilaterally preconditioned methods such as K-FAC [13], Shampoo [4], SOAP [17], and FISMO [23], as well as one-sided optimizers such as Muon [6, 8], ASGO [1], and SSO [22]. These methods have shown strong empirical performance and increasing practical relevance in large-scale training.

Anisotropic Training Dynamics. A growing body of work has shown that the loss landscape of neural networks is highly anisotropic and ill-conditioned. In particular, the Hessian spectrum typically consists of a small number of large eigenvalues forming a dominant top subspace, while the vast majority of eigenvalues are close to zero and form a low-curvature bulk subspace [18, 19]. Although a substantial part of the gradient energy lies in the dominant eigendirections, these components often exhibit strong oscillation and may slow down optimization [5, 15]. In contrast, some low-curvature directions can be important for sustained progress, but their curvature estimates are often noisy and require careful stabilization. Motivated by this observation, several recent optimizers have explicitly adopted a top-space damping and tiny-subspace acceleration design to improve training efficiency [24, 25].

Understanding Structured Preconditioners. While Shampoo [4] and its variants have achieved remarkable empirical success [7], a recent line of work aims to theoretically explain their effectiveness. In particular, [3, 14, 21] interpret Shampoo-like preconditioners through the lens of Kronecker-product approximations $L \otimes R$ to the gradient covariance matrix $\mathbb{E}[gg^\top]$, typically evaluated under the Frobenius norm. More recently, Lin et al. [11] have further generalized this perspective by studying these preconditioners through the framework of *matrix divergences*. This view shows that different bilateral Shampoo variants can be interpreted as minimizing different Bregman divergences between the dense covariance matrix and its Kronecker approximation. It therefore provides a unified language for comparing Frobenius-, von Neumann-, and LogDet/KL-induced preconditioners.

Appendix B. Proofs and Additional Details

B.1. Proof of Proposition 1

The dimension of the ambient space $\mathbb{S}^{mn}_{++}$ is $\frac{mn(mn+1)}{2} - 1 = \mathcal{O}(m^2n^2)$. The Kronecker manifold $\mathcal{M}$ is parameterized by L and R, yielding at most $\frac{m(m+1)}{2} + \frac{n(n+1)}{2} = \mathcal{O}(m^2 + n^2)$ degrees of freedom. Because $\dim(\mathcal{M}) \ll \dim(\mathbb{S}^{mn}_{++})$, $\mathcal{M}$ is a strictly lower-dimensional submanifold, which has Lebesgue measure zero in $\mathbb{S}^{mn}_{++}$.

Since C admits a density with respect to the Lebesgue measure, we have $\mathbb{P}(C \in \mathcal{M}) = 0$, so $C \notin \mathcal{M}$ almost surely. For the Bregman divergences considered in this paper, $\mathcal{B}_F(C, S) \to 0$ implies $S \to C$. Since the Kronecker-factored SPD family is closed under positive-definite limits, $\inf_{L,R} \mathcal{B}_F(C, L \otimes R) = 0$ would imply $C \in \mathcal{M}$, which contradicts $C \notin \mathcal{M}$ almost surely. Therefore, we can derive

$$\inf_{L,R} \mathcal{B}_F(C, L \otimes R) > 0 \quad \text{a.s..}$$

B.2. Proof of Section 3.2

By the definition of the Bregman matrix divergence generated by a strictly convex, differentiable function F, we have:

$$\mathcal{B}_F(A, B) = F(A) - F(B) - \mathrm{Tr}\big(\nabla F(B)(A - B)\big).$$

For a spectral generating function $F(M) = \sum_k f(\lambda_k(M)) = \mathrm{Tr}(f(M))$, its matrix gradient is computed directly via the scalar derivative applied to its eigenvalues: $\nabla F(M) = f'(M)$. Consequently, given the spectral decomposition $B = V\Omega V^\top$, the gradient can be written as $\nabla F(B) = V f'(\Omega) V^\top$. Using the linearity of the trace, we can expand the cross-term:

$$\mathrm{Tr}\big(\nabla F(B)(A - B)\big) = \mathrm{Tr}\big(\nabla F(B)A\big) - \mathrm{Tr}\big(\nabla F(B)B\big).$$

The second term depends only on the spectrum of B:

$$\mathrm{Tr}\big(\nabla F(B)B\big) = \mathrm{Tr}\big(V f'(\Omega)V^\top V\Omega V^\top\big) = \mathrm{Tr}\big(f'(\Omega)\Omega\big) = \sum_{j=1}^{d} f'(\omega_j)\omega_j.$$

For the cross-trace term $\mathrm{Tr}(\nabla F(B)A)$, we substitute $A = U\Lambda U^\top$ and apply the cyclic property of the trace:

$$\mathrm{Tr}\big(\nabla F(B)A\big) = \mathrm{Tr}\big(V f'(\Omega)V^\top U\Lambda U^\top\big) = \mathrm{Tr}\big(f'(\Omega)V^\top U\Lambda U^\top V\big).$$

Let $Q = U^\top V$ represent the orthogonal transition matrix between the eigenspaces of A and B. By definition, squaring its entries gives the eigenspace alignment matrix P, such that $P_{ij} = Q_{ij}^2 = \langle u_i, v_j\rangle^2$. Thus, the cross-trace becomes:

$$\mathrm{Tr}\big(\nabla F(B)A\big) = \mathrm{Tr}\big(f'(\Omega)Q^\top\Lambda Q\big) = \sum_{i=1}^{d}\sum_{j=1}^{d} f'(\omega_j)\lambda_i Q_{ij}^2 = \sum_{i=1}^{d}\sum_{j=1}^{d} f'(\omega_j)\lambda_i P_{ij}.$$

Substituting these decoupled components back into the Bregman divergence definition, we obtain:

$$\mathcal{B}_F(A, B) = F(A) - F(B) + \mathrm{Tr}\big(\nabla F(B)B\big) - \mathrm{Tr}\big(\nabla F(B)A\big).$$

We can now clearly see the three distinct parts of the decomposition:

$$\Psi(\Lambda) = F(A) = \sum_{i=1}^{d} f(\lambda_i), \quad \Phi(\Omega) = \mathrm{Tr}\big(\nabla F(B)B\big) - F(B) = \sum_{j=1}^{d}\big(f'(\omega_j)\omega_j - f(\omega_j)\big),$$

$$\sum_{i=1}^{d}\sum_{j=1}^{d} g_F(\lambda_i, \omega_j)P_{ij} = -\mathrm{Tr}\big(\nabla F(B)A\big) \implies g_F(\lambda_i, \omega_j) = -\lambda_i f'(\omega_j).$$

For the three divergences considered in this paper, the coupling terms are:

Frobenius Divergence: Let $F(M) = \frac{1}{2}\mathrm{Tr}(M^2) \implies f(x) = \frac{1}{2}x^2$, yielding $f'(x) = x$.

$$g_{\text{Frob}}(\lambda_i, \omega_j) = -\lambda_i(\omega_j) = -\lambda_i\omega_j.$$

von Neumann Divergence: Let $F(M) = \mathrm{Tr}(M \log M - M) \implies f(x) = x \log x - x$, yielding $f'(x) = \log x$.

$$g_{\mathrm{vN}}(\lambda_i, \omega_j) = -\lambda_i(\log \omega_j) = -\lambda_i \log \omega_j.$$

Log-Determinant Divergence[2]**:** Let $F(M) = -\log\det M = -\mathrm{Tr}(\log M) \implies f(x) = -\log x$, yielding $f'(x) = -1/x$.

$$g_{\mathrm{LogDet}}(\lambda_i, \omega_j) = -\lambda_i \left(-\frac{1}{\omega_j}\right) = \frac{\lambda_i}{\omega_j}.$$

This completes the proof.

Remark 6 *The function g_F serves as the "exchange rate" for approximation errors. By analyzing g_F, we can derive the spectral bias of each divergence:*

- ***Absolute-error emphasis (Frobenius and von Neumann).*** *In both $\mathcal{B}_{Frob}$ and $\mathcal{B}_{vN}$, the cross-spectral term is weighted by the target eigenvalue λ_i, either linearly through $\lambda_i\omega_j$ or through $\lambda_i \log \omega_j$. As a result, mismatches of small eigenvalues contribute relatively little to the divergence, while errors on large eigenvalue components are penalized more strongly. Therefore, these divergences are biased toward the top spectrum: when approximation error is unavoidable, they are more likely to fit the dominant eigenspace.*
- ***Relative-error emphasis (Log-Determinant).*** *For $\mathcal{B}_{\mathrm{LogDet}}$, the coupling term is λ_i/ω_j. The penalty depends on relative scale rather than only on absolute magnitude. As a result, small eigenvalue components are not ignored, and the divergence puts more weight on preserving relative spectral accuracy across the whole spectrum.*

B.3. BregTop Algorithmic Details

The statistics $(\Delta_L, \Delta_R) = \textsc{Stats}(G, U_L, U_R, \lambda_L, \lambda_R)$ in Algorithm 1 are given as follows. These updates follow the divergence-induced Kronecker statistics derived in [11]:

$$(\Delta_L, \Delta_R) := \begin{cases} \left(GG^\top,\ G^\top G\right) & \text{(Original)} \\ \left(GG^\top/\mathrm{Tr}(R),\ G^\top G/\mathrm{Tr}(L)\right) & \text{(VN-v1)} \\ \left(GG^\top/\sum_{i=1}^{n}\lambda_{R,i},\ G^\top G/\sum_{i=1}^{m}\lambda_{L,i}\right) & \text{(VN-v2)} \\ \left(GRG^\top/\mathrm{Tr}(R^2),\ G^\top LG/\mathrm{Tr}(L^2)\right) & \text{(F-v1)} \\ \left(GU_R\mathrm{Diag}(\lambda_R)U_R^\top G^\top/\sum_{i=1}^{n}\lambda_{R,i}^2,\ G^\top U_L\mathrm{Diag}(\lambda_L)U_L^\top G/\sum_{i=1}^{m}\lambda_{L,i}^2\right) & \text{(F-v2)} \end{cases}$$

2. Up to an irrelevant positive scaling constant, we take $F(M) = -\log\det M$ for the LogDet case.

B.4. Proof of Theorem 2

Define the reduced objective

$$\phi(L, R) := \mathcal{B}_F(C, L \otimes R).$$

Stationarity of (L^*, R^*) implies that, for any feasible perturbation $(\Delta L, \Delta R)$, the first-order directional derivative of ϕ is zero:

$$D\phi(L^*, R^*)[\Delta L, \Delta R] = 0 \qquad \forall\, \Delta L \in \mathbb{S}^m,\ \Delta R \in \mathbb{S}^n.$$

We first characterize the tangent space of $\mathcal{M}$ at $S^* = L^* \otimes R^*$. Consider the smooth curve

$$S(t) := (L^* + t\Delta L) \otimes (R^* + t\Delta R).$$

Then

$$\frac{d}{dt} S(t)\Big|_{t=0} = \Delta L \otimes R^* + L^* \otimes \Delta R.$$

Hence, the tangent space is

$$T_{S^*}\mathcal{M} = \{\Delta L \otimes R^* + L^* \otimes \Delta R \ :\ \Delta L \in \mathbb{S}^m,\ \Delta R \in \mathbb{S}^n\}.$$

Next, by the chain rule,

$$D\phi(L^*, R^*)[\Delta L, \Delta R] = D_S\mathcal{B}_F(C, S^*)\big[\Delta L \otimes R^* + L^* \otimes \Delta R\big].$$

Since the left-hand side is zero for all $(\Delta L, \Delta R)$, we obtain

$$D_S\mathcal{B}_F(C, S^*)[H] = 0 \qquad \forall\, H \in T_{S^*}\mathcal{M}.$$

We differentiate the Bregman divergence with respect to its second argument. From the definition,

$$\mathcal{B}_F(C, S) = F(C) - F(S) - \langle \nabla F(S),\, C - S\rangle.$$

Fix any direction $H \in \mathbb{S}^{mn}$. Using Fréchet differentiation, we derive

$$D_S\mathcal{B}_F(C, S)[H] = -\langle \nabla F(S), H\rangle - \big(\langle \nabla^2 F(S)[H],\, C - S\rangle - \langle \nabla F(S), H\rangle\big).$$

The two first-order terms cancel, yielding

$$D_S\mathcal{B}_F(C, S)[H] = \langle \nabla^2 F(S)[H],\, S - C\rangle.$$

Using the self-adjointness of $\nabla^2 F(S)$ under the Frobenius inner product, we equivalently write

$$D_S\mathcal{B}_F(C, S)[H] = \langle \nabla^2 F(S)[S - C],\, H\rangle.$$

Evaluating at $S = S^*$ gives

$$D_S\mathcal{B}_F(C, S^*)[H] = \big\langle \nabla^2 F(S^*)[S^* - C],\, H\big\rangle.$$

Combining all the proof above, for every $H \in T_{S^*}\mathcal{M}$ we have

$$\big\langle \nabla^2 F(S^*)[S^* - C],\, H\big\rangle = 0,$$

which is exactly

$$\nabla^2 F(S^*)[S^* - C] \perp T_{S^*}\mathcal{M}.$$

Thus the divergence-weighted error is orthogonal to all feasible first-order directions on the Kronecker manifold. This completes the proof.

B.5. Proof of Theorem 4

Let $M \in \mathbb{R}^{m\times n}$ be the input momentum matrix. Let $U_L \in \mathbb{R}^{m\times m}$ and $U_R \in \mathbb{R}^{n\times n}$ be the orthogonal matrices from the eigendecomposition, satisfying $U_L U_L^\top = I_m$ and $U_R U_R^\top = I_n$. Let $S \in \mathbb{R}^{m\times n}$ be the eigenvalue matrix where $S_{i,j} = \lambda_i^{(L)}\lambda_j^{(R)}$.

Algorithm 2 rotates M into the Kronecker eigenspace to obtain $\widetilde{M} = U_L^\top M U_R$. We define the top eigenspace set as Ω_ρ, the indices of the largest $K = \lceil \rho mn \rceil$ entries of S. Let $E \in \{0,1\}^{m\times n}$ be the indicator matrix for the top space, where $E_{i,j} = 1$ if $(i,j) \in \Omega_\rho$, and 0 otherwise. Consequently, the indicator matrix for the orthogonal bottom space is $E^\perp = \mathbf{1} - E$, where $\mathbf{1}$ is the matrix of all ones.

By construction, the scaling matrix W decomposes as

$$W = (S^{-\frac{1}{2}} \odot E) + (\chi_{\rho,q}\mathbf{1} \odot E^\perp).$$

Since $\mathbf{1} \odot E^\perp = E^\perp = \mathbf{1} - E$, we have:

$$W = (S^{-\frac{1}{2}} \odot E) + \chi_{\rho,q}(\mathbf{1} - E).$$

The final output of the algorithm $\widehat{M}$, is computed by applying the mask and rotating back:

$$\widehat{M} = U_L(\widetilde{M} \odot W)U_R^\top$$

Substituting the decomposed W into the equation and utilizing the distributive property of the Hadamard product:

$$\widehat{M} = U_L\left(\widetilde{M} \odot \left[(S^{-\frac{1}{2}} \odot E) + \chi_{\rho,q}\mathbf{1} - \chi_{\rho,q}E\right]\right)U_R^\top$$

$$\widehat{M} = U_L\left(\widetilde{M} \odot (S^{-\frac{1}{2}} \odot E)\right)U_R^\top + U_L(\chi_{\rho,q}\widetilde{M} \odot \mathbf{1})U_R^\top - U_L(\chi_{\rho,q}\widetilde{M} \odot E)U_R^\top$$

Notice that $\widetilde{M} \odot \mathbf{1} = \widetilde{M}$. The second term can be simplified by the orthogonality of U_L and U_R:

$$U_L(\chi_{\rho,q}\widetilde{M})U_R^\top = \chi_{\rho,q}U_L(U_L^\top M U_R)U_R^\top = \chi_{\rho,q}(U_L U_L^\top)M(U_R U_R^\top) = \chi_{\rho,q}I_m M I_n = \chi_{\rho,q}M$$

By rearranging the terms, we obtain:

$$\widehat{M} = U_L\left(\widetilde{M} \odot (S^{-\frac{1}{2}} \odot E)\right)U_R^\top + \chi_{\rho,q}M - \chi_{\rho,q}U_L(\widetilde{M} \odot E)U_R^\top.$$

We now verify that $\mathcal{P}_{\text{top}}^\rho(M) := U_L(\widetilde{M} \odot E)U_R^\top$ is an exact orthogonal projection of M onto the top eigenspace.

Although the retained index set Ω_ρ typically forms a staircase shape rather than a rectangle, the orthogonal projection property holds for any arbitrary subset. For notational convenience and without loss of generality, we assume E isolates a top-left $b \times c$ block:

$$\widetilde{M} \odot E = \begin{bmatrix} \widetilde{M}_{b\times c} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Write $U_L = [\widetilde{U}_L \; U_{L\perp}]$ and $U_R = [\widetilde{U}_R \; U_{R\perp}]$, where $\widetilde{U}_L \in \mathbb{R}^{m\times b}$ and $\widetilde{U}_R \in \mathbb{R}^{n\times c}$ contain the retained eigenvectors. Then block multiplication gives

$$U_L \begin{bmatrix} \widetilde{M}_{b\times c} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} U_R^\top = [\widetilde{U}_L \quad U_{L\perp}] \begin{bmatrix} \widetilde{M}_{b\times c} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \widetilde{U}_R & U_{R\perp} \end{bmatrix}^\top$$
$$= [\widetilde{U}_L \widetilde{M}_{b\times c} \quad \mathbf{0}] \begin{bmatrix} \widetilde{U}_R & U_{R\perp} \end{bmatrix}^\top = \widetilde{U}_L \widetilde{M}_{b\times c} \widetilde{U}_R^\top.$$

Substituting the projection $\widetilde{M}_{b\times c} = \widetilde{U}_L^\top M \widetilde{U}_R$ into the expression yields:

$$\mathcal{P}^\rho_{\text{top}}(M) = (\widetilde{U}_L \widetilde{U}_L^\top) M (\widetilde{U}_R \widetilde{U}_R^\top) = P_L M P_R,$$

where $P_L = \widetilde{U}_L \widetilde{U}_L^\top$ and $P_R = \widetilde{U}_R \widetilde{U}_R^\top$ are the orthogonal projectors onto the top subspaces of L and R, respectively. By the properties of the Kronecker product, this is equivalent to:

$$\text{vec}(\mathcal{P}^\rho_{\text{top}}(M)) = (P_R \otimes P_L)\text{vec}(M).$$

Since the Kronecker product of two orthogonal projectors is itself an orthogonal projector, this confirms that $U_L(\widetilde{M} \odot E)U_R^\top$ is the exact orthogonal projection of M onto the subspace spanned by the top bc eigenvectors of $L \otimes R$. Consequently, the term $U_L \left((\widetilde{M} \odot E) \odot S^{-\frac{1}{2}} \right) U_R^\top$ represents the preconditioned momentum where each eigen-component within this top subspace is scaled by the inverse square root of its corresponding eigenvalue.

By definition, the projection onto the orthogonal bottom space is $\mathcal{P}_{\text{bot}}(M) = M - \mathcal{P}^\rho_{\text{top}}(M)$.

$$\widehat{M} = \underbrace{U_L \left((\widetilde{M} \odot E) \odot S^{-\frac{1}{2}} \right) U_R^\top}_{\text{Precise eigenvalue damping on top space}} + \underbrace{\chi_{\rho,q} \left(M - \mathcal{P}^\rho_{\text{top}}(M) \right)}_{\text{Uniform acceleration on bottom space}}$$

The expression explicitly demonstrates that the global masking operation is mathematically identical to projecting the gradient onto the top eigenspace for precise damping, while independently applying an adaptive uniform acceleration $\chi_{\rho,q}$ to the remaining orthogonal bottom space.

Appendix C. Details of Hessian–Covariance Alignment Experiments

We follow the same setup used in prior Hessian studies [2, 15]. The model is a two-layer Transformer with hidden dimension 64 and 8 attention heads, trained with SGD on the first 1000 samples of SST-2 dataset using the MSE loss. For each checkpoint, we evaluate the Q/K/V matrices of the second Transformer layer.

For each selected block, we collect $N = 200$ mini-batch gradients and form the empirical second-moment matrix $C = N^{-1} \sum_{b=1}^{N} g_b g_b^\top$, where g_b is the batch-averaged gradient vector of the corresponding Q/K/V block. We compute the top covariance eigenspace from the SVD of the stacked gradient matrix, without explicitly forming C.

The Hessian block is computed with respect to the same Q/K/V parameter block. We use Hessian-vector products and the Lanczos routine to obtain the top Hessian eigenvectors. The Hessian loss is evaluated on the same SST-2 subset. We report the top-space alignment

$$\left(\text{Overlap@5} = \|(U_H^{1:5})^\top U_C^{1:5}\|_F^2 / 5 \right)^{1/2},$$

where $U_H^{1:5}$ and $U_C^{1:5}$ denote the top-5 Hessian and covariance eigenspaces, respectively. To probe lower-spectrum covariance directions, we also compute the spectral-band overlap

$$\left(\|(U_H^{180:200})^\top U_C^{180:200}\|_F^2/21\right)^{1/2}.$$

Appendix D. Additional Details for the SST-2 Optimization Experiment

Task and model. We conduct the optimization-speed comparison on the **full** SST-2 training set. The model is a randomly initialized Transformer with 4 layers, hidden dimension 128, and 8 attention heads, containing approximately 4.7M parameters. The classifier is trained with mean-squared error loss on one-hot class targets, following [2, 15].

Parameter grouping. For all Shampoo-type methods, we use the same parameter grouping. The embedding layer, bias terms, normalization parameters, and classification head are optimized by Adam. The remaining matrix weights in the are optimized by the corresponding Shampoo-type optimizer. No weight decay is used.

Selected hyperparameters. All methods use batch size 128, learning rate 10^{-3} for both the matrix optimizer and the Adam group. For BregTop-VN, we use $\rho = 0.06$, $c = 3.0$, and $q = 0.01$. For BregTop-F, we use $\rho = 0.05$, $c = 1.6$, and $q = 0.01$. These values are used for both the v1 and v2 variants of the corresponding BregTop optimizer. For the remaining hyperparameters, such as β_1 and β_2, we follow [11].

Evaluation metric. We compare optimization speed by measuring the number of steps required for the EMA-smoothed training loss to fall below 0.05. The EMA coefficient is set to 0.98.