TOWARD A MORE FAVORABLE CURVATURE STRUCTURE: A BASIS PERSPECTIVE ON MATRIX OPTIMIZERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Matrix optimizers have shown promising performance in large language model training. Although prior explanations have linked their effectiveness to curvature, the underlying mechanisms remain incompletely understood. To better understand these mechanisms, we view optimizer updates as basis selection followed by a diagonally preconditioned update in the induced basis. We find that bases induced by matrix optimizers can make curvature structure more favorable for diagonal preconditioning, partly explaining their effectiveness. Empirically, we examine Gauss–Newton curvature during training and find that (i) matrix-optimizer-induced bases generally yield greater diagonal concentration of curvature in hidden parameter blocks than the ambient basis used by optimizers such as Adam; (ii) in the basis induced by each optimizer, the scaling used by its diagonal preconditioner closely corresponds to the diagonal curvature; and (iii) optimizers with higher diagonal concentration of curvature in their induced bases generally achieve better training performance. We further support these findings with an initial theoretical analysis of quadratic problems and neural networks, identifying conditions under which bases induced by matrix optimizers yield more favorable curvature structure and examining how this structure can contribute to improved optimization performance.

1 INTRODUCTION

Matrix optimizers such as Muon (Jordan et al., 2024) and SOAP (Vyas et al., 2025) are moving to the forefront of large language model (LLM) training owing to their promising performance. Compared with Adam (Kingma & Ba, 2015), a widely used diagonally preconditioned optimizer for neural networks, they have shown higher training efficiency in a range of empirical studies (Vyas et al., 2025; Liu et al., 2025b; Shah et al., 2025; Wen et al., 2026; Frans et al., 2025a).

Curvature provides a lens for understanding optimizer behavior and has been used to explain why SGD underperforms Adam in Transformer training, with broader implications for diagonally preconditioned optimizers (Zhang et al., 2024). More recently, studies have taken initial steps toward linking the superior performance of matrix optimizers to curvature (Morwani et al., 2025; Su, 2025; Du & Su, 2026). However, how matrix optimizers exploit curvature information to achieve such gains is not yet fully clear.

Meanwhile, a complementary line of work highlights the role of basis choice in optimization performance (Vyas et al., 2025; Zhang et al., 2025a; Xie et al., 2025; Liu et al., 2025a; Gong et al., 2026; Frans et al., 2025b). Building on these insights, we view optimizer updates as diagonal preconditioning in an optimizer-induced basis:

$$w^+ = w - \eta Q D Q^\top g, \tag{1}$$

where w and w^+ denote the current and updated parameters, respectively, g denotes the gradient or momentum, Q and D are induced by the optimizer, with Q defining an orthogonal basis and D specifying diagonal preconditioning in that basis.[1]

[1] See Section 3.2 for a detailed explanation.

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.

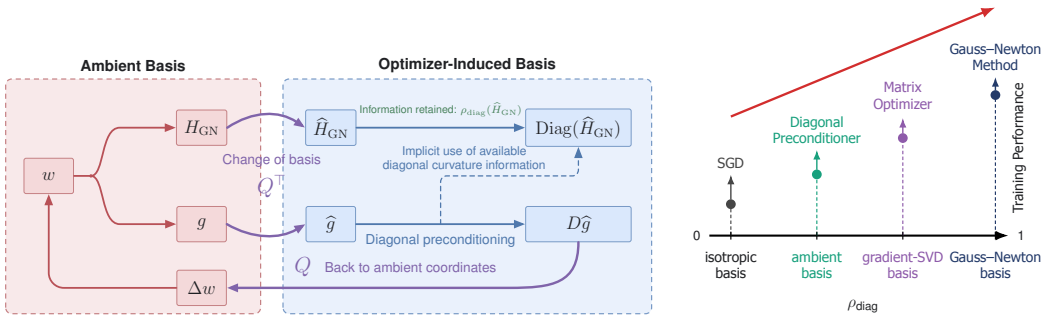


Figure 1: A basis perspective on optimizers for better understanding their behavior in LLM training.

As illustrated in Figure 1, the update transforms g into coordinates in basis Q, applies diagonal preconditioning in these coordinates, and then transforms the resulting direction back to the ambient basis. This motivates us to study curvature across optimizer-induced bases to better understand how optimizers exploit curvature and why matrix optimizers can be more effective. Our central explanation is summarized as follows:

> Bases induced by matrix optimizers lead to a more favorable curvature structure for diagonal preconditioning by producing stronger diagonal concentration of Gauss–Newton curvature.

We support this explanation with the following empirical and theoretical findings on Gauss–Newton (GN) curvature across optimizer-induced bases and its implications for diagonal preconditioning:

1. **GN diagonal concentration varies across optimizer-induced bases.** For hidden blocks, GN curvature exhibits clear diagonal concentration in both the ambient and matrix-optimizer-induced bases, with substantially higher diagonal Frobenius ratios in the latter. To explain this stronger diagonal concentration, we establish sufficient conditions for strict improvement in quadratic models and neural networks.
2. **Diagonal preconditioners implicitly exploit available diagonal curvature information.** In our experiments, both cosine similarity and element-wise agreement indicate a close correspondence between D^{-1} and the GN diagonal entries for Adam and Lion in the ambient basis and Muon in the momentum-SVD basis. These observations suggest that D implicitly exploits the diagonal curvature information available in each optimizer-induced basis.
3. **GN diagonal concentration favors diagonal preconditioning.** Empirically, higher GN diagonal Frobenius ratios in Q generally accompany better training performance. Theoretically, we bound the GN quadratic-model decrease shortfall relative to the optimal GN step using diagonal scaling mismatch in D and residual off-diagonal curvature in Q.
4. **Distinct GN curvature geometry in non-hidden blocks.** Embeddings and the LM head exhibit different GN curvature patterns from hidden blocks. These differences may partly explain why Muon, a representative matrix optimizer, applies its matrix update only to hidden blocks while retaining AdamW for embeddings and the LM head, rather than using a uniform update rule throughout the model (Jordan et al., 2024).

2 Related Work

Basis choice in optimization. Gong et al. (2026) separate rotation selection from within-basis updates, while Liu et al. (2025a) compare Adam and Gauss–Newton diagonal preconditioning across bases and gradient-noise regimes. Empirically, coordinate rotations affect Adam's performance (Zhang et al., 2025a), while SOAP reports improved training efficiency over AdamW by applying Adam in an accumulated gradient-Gram eigenbasis (Vyas et al., 2025). Muon variants with the same basis construction but different within-basis update rules can perform comparably after learning-rate tuning (Shumaylov et al., 2026). However, why some bases lead to better optimizer performance remains unclear. Our curvature-based explanation examines how basis choice affects GN diagonal concentration and its relationship to diagonal preconditioning.

Curvature structure and optimizer design. Neural-network Hessians have been analyzed in terms of their spectral structure (Sagun et al., 2017; Ghorbani et al., 2019; Tang et al., 2025), as well as near-block-diagonal structure and block heterogeneity in the ambient basis (Dong et al., 2025; Zhang et al., 2024). Curvature structure also guides optimizer design (Liu et al., 2024; An et al., 2025), including Hessian-informed parameter grouping (Zhang et al., 2025b), Kronecker-factored Fisher approximations (Martens & Grosse, 2015), diagonal estimation in the resulting product eigenbasis (George et al., 2018), Shampoo's curvature-informed preconditioner (Gupta et al., 2018; Morwani et al., 2025), and curvature-based quadratic surrogates (Du & Su, 2026). While prior studies of Hessian structure primarily focus on the ambient basis, we characterize GN curvature structure across optimizer-induced bases and compare their diagonal concentration.

Perspectives on matrix optimizers. Spectral-norm geometry offers an interpretation of Shampoo (Bernstein & Newhouse, 2024; Eschenhagen et al., 2026) and Muon-style updates (Pethick et al., 2025), while Lin et al. (2026) study structured gradient second-moment estimation from an information-geometric perspective. Su (2025) studies Muon under an isotropic curvature model. We focus on a different geometric object: GN curvature represented in optimizer-induced bases. Wang et al. (2026) link Muon's empirical advantage over Adam to lower normalized directional sharpness, but their theory compares Muon with GD. This leaves unclear which curvature information Muon and Adam exploit and why Muon uses it more effectively. We clarify what curvature information matrix optimizers exploit and why they outperform Adam: their induced bases yield a more diagonal GN curvature structure, allowing diagonal preconditioners to exploit more curvature information.

3 A BASIS PERSPECTIVE ON OPTIMIZERS

3.1 PRELIMINARIES

Notation. Let $f_w(x)$ denote the model logits for input x and parameters $w \in \mathbb{R}^d$. Given data $\{(x_i, y_i)\}_{i=1}^N$, define $\mathcal{L}(w) = \frac{1}{N}\sum_{i=1}^N \ell(f_w(x_i), y_i)$, where ℓ is the cross-entropy loss, convex in its logit argument. Write $g = \nabla_w \mathcal{L}$ and $H = \nabla_w^2 \mathcal{L}$. For a matrix parameter block $W \in \mathbb{R}^{m\times n}$, we set $d = mn$ and write $w = \mathrm{vec}(W)$ and $G = \nabla_W \mathcal{L} = \mathrm{mat}(g)$, with derivatives taken with respect to this block. Here vec denotes column-major vectorization, mat is its inverse, and $\otimes$ denotes the Kronecker product. For matrices of compatible dimensions, $\mathrm{vec}(ABC) = (C^\top \otimes A)\,\mathrm{vec}(B)$.

Gauss–Newton curvature. The Hessian H characterizes local curvature. Assuming the Hessian is invertible, classical Newton's method obtains the update by using H to precondition the gradient

$$w^+ = w - \eta H^{-1} g.$$

However, the Newton step is not guaranteed to be a descent direction since H may be indefinite even for smooth neural network losses (Martens, 2010). A common alternative is the generalized Gauss–Newton matrix H_{GN} (Nocedal & Wright, 2006; Schraudolph, 2002; Martens & Sutskever, 2012). For notational simplicity, we give its definition for a single sample (x, y):

$$H_{\mathrm{GN}} := \nabla_w f_w(x)\, \nabla_z^2 \ell\, \nabla_w f_w(x)^\top \succcurlyeq 0,$$

which serves as a positive semidefinite approximation to the Hessian; see Appendix C for details. Replacing H with this curvature surrogate yields the GN update

$$w^+ = w - \eta H_{\mathrm{GN}}^{-1} g.$$

For large models, the update can only be computed implicitly by approximately solving a quadratic subproblem due to the prohibitive size of H_{GN} at each step. Recently, Meterez et al. (2026) showed that local GN quadratic models offer surprisingly accurate predictions of LLM training dynamics, supporting H_{GN} as a suitable local curvature surrogate. Abreu et al. (2026) further demonstrated several-fold speedups over existing optimizers using the GN method in LLM training.

3.2 OPTIMIZER-INDUCED BASES AND DIAGONAL UPDATES

The general update in (1) unifies widely used optimizers with distinct update formulations and design motivations. Similar formulations have appeared in work on rotation-policy design (Gong et al.,

Table 1: A basis perspective on optimizers. Ignoring momentum and EMA accumulation, the listed matrix optimizers share the gradient SVD basis but differ in their diagonal update rules. Here, $\mathrm{Diag}(H_{\mathrm{GN}}^{Q})$ is the GN diagonal in the Q basis, and $\mathrm{NS5}(\cdot)$ denotes the transformation induced by five Newton–Schulz iterations. See Appendix A.4 for algorithmic details.

Method	Basis	Diagonal preconditioner
SGD	Isotropic Diagonal Basis[2] of H_{GN}	$\mathrm{Diag}(H_{\mathrm{GN}}^{Q})^{-1}(\cdot)$
Adam	Ambient Basis	$\mathrm{Adam}(\cdot)$
Lion		$\mathrm{Sign}(\cdot)$
Muon (Idealized)	SVD Basis of G	$\mathrm{Sign}(\cdot)$
Muon (NS5)		$\mathrm{NS5}(\cdot)$
SOAP		$\mathrm{Adam}(\cdot)$
GN method	Eigenbasis of H_{GN}	$\mathrm{Diag}(H_{\mathrm{GN}}^{Q})^{-1}(\cdot)$

2026) and on diagonal preconditioning across bases and gradient-noise regimes (Liu et al., 2025a). When $Q = I_d$, this update reduces to diagonal preconditioning in the ambient basis, as in Adam.

In contrast, many matrix optimizers, including Muon and SOAP, share the gradient SVD basis when exponential moving averages are omitted. Let $G = U\Sigma V^\top$ be the SVD of the gradient matrix. The update in vectorized form is

$$w^+ = w - \eta(V \otimes U)D(V \otimes U)^\top g, \quad \text{with } g = \mathrm{vec}(G). \tag{2}$$

The classical GN method and SGD also fit this framework. Inverse-diagonal GN preconditioning exactly reproduces their respective updates in the GN eigenbasis and an isotropic diagonal basis, with the scalar factor for SGD absorbed into the learning rate. Table 1 summarizes these cases.

3.3 GN DIAGONAL CONCENTRATION ACROSS BASES

Under the unified view above, each optimizer selects a basis and then applies diagonal preconditioning in that basis. Since diagonal preconditioners rescale each coordinate independently, ignoring off-diagonal curvature coupling, the curvature structure in the chosen basis matters. Experiments on quadratic problems further suggest that Adam's preconditioning tends to be more effective as curvature becomes more diagonal (Zhang et al., 2025b). This motivates examining curvature diagonal concentration across optimizer-induced bases, using the diagonal Frobenius norm ratio.

Definition 1 (Diagonal Frobenius norm ratio)**.** Let $K \in \mathbb{R}^{d\times d}$ be a nonzero matrix. Its diagonal Frobenius norm ratio is defined as

$$\rho_{\mathrm{diag}}(K) := \frac{\|\mathrm{Diag}(K)\|_F}{\|K\|_F},$$

where $\mathrm{Diag}(K)$ denotes the diagonal matrix of K.

Building on this basis perspective, we study curvature through the Gauss–Newton surrogate H_{GN}. In an orthogonal basis $Q \in \mathcal{O}(d)$, the curvature is represented by $H_{\mathrm{GN}}^{Q} := Q^\top H_{\mathrm{GN}} Q$, with

$$\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{Q}) = \frac{\|\mathrm{Diag}(H_{\mathrm{GN}}^{Q})\|_F}{\|H_{\mathrm{GN}}^{Q}\|_F} = \frac{\|\mathrm{Diag}(Q^\top H_{\mathrm{GN}} Q)\|_F}{\|H_{\mathrm{GN}}\|_F},$$

where the second equality follows from the orthogonal invariance of the Frobenius norm. In the orthogonal basis Q, a direct interpretation of $\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{Q})$ is the Frobenius cosine similarity between H_{GN}^{Q} and its diagonal projection. Notably, the rotation bases associated with SGD and the GN method in Table 1 attain its two extremes.

[2]An isotropic diagonal basis Q_{iso} of H_{GN} satisfies $\mathrm{diag}(Q_{\mathrm{iso}}^\top H_{\mathrm{GN}} Q_{\mathrm{iso}}) = \frac{\mathrm{tr}(H_{\mathrm{GN}})}{d}\mathbf{1}_d$, yielding a constant diagonal. Hence, $Q_{\mathrm{iso}}\,\mathrm{Diag}(H_{\mathrm{GN}}^{Q_{\mathrm{iso}}})^{-1}Q_{\mathrm{iso}}^\top g = \frac{d}{\mathrm{tr}(H_{\mathrm{GN}})}g$, recovering SGD after rescaling the learning rate.

Theorem 1 (Extremal diagonal concentration)**.** Let $H_{\mathrm{GN}} \in \mathbb{R}^{d\times d}$ be a nonzero positive semidefinite (PSD) matrix. For any orthogonal matrix $Q \in \mathcal{O}(d)$,

$$\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{Q_{\mathrm{iso}}}) = \frac{\mathrm{tr}(H_{\mathrm{GN}})}{\sqrt{d}\,\|H_{\mathrm{GN}}\|_F} \le \rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{Q}) \le 1 = \rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{Q_{\mathrm{eig}}}).$$

The lower and upper bounds are attained exactly by isotropic diagonal bases and eigenbases of H_{GN}, represented by Q_{iso} and Q_{eig}, respectively.

This ratio also measures the heterogeneity of diagonal curvature in the chosen basis. Since orthogonal transformations preserve both the trace and the Frobenius norm,

$$\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{Q})^2 = \frac{d^2\ \mathrm{Var}_i\left[(H_{\mathrm{GN}}^{Q})_{ii}\right] + \mathrm{tr}(H_{\mathrm{GN}})^2}{d\|H_{\mathrm{GN}}\|_F^2},$$

where the variance is taken uniformly over diagonal entries. For a fixed H_{GN} expressed in different bases, stronger diagonal concentration is therefore equivalent to greater diagonal curvature heterogeneity, making it more amenable to exploitation by diagonal preconditioners.

4 Empirical and Theoretical Observations

In this section, we study the curvature structure in optimizer-induced bases. Given Muon's recent success, we take it as a representative matrix optimizer and focus on the momentum SVD basis used in its actual update, obtained by replacing the gradient matrix G in Eq. (2) with the momentum matrix M. Other matrix optimizers are considered in Section 4.3.

4.1 Diagonal Preconditioners Correspond to Diagonal Curvature

For each optimizer, we first compare the inverse of the diagonal preconditioner, D^{-1}, with $\mathrm{Diag}(H_{\mathrm{GN}}^{Q})^{1/2}$ in its induced basis Q: the ambient basis for Adam and Lion (Chen et al., 2023), and the momentum SVD basis for Muon. High cosine similarity is reported in Table 2 for all methods, with Adam achieving the highest values. Figure 2 further illustrates the element-wise agreement.

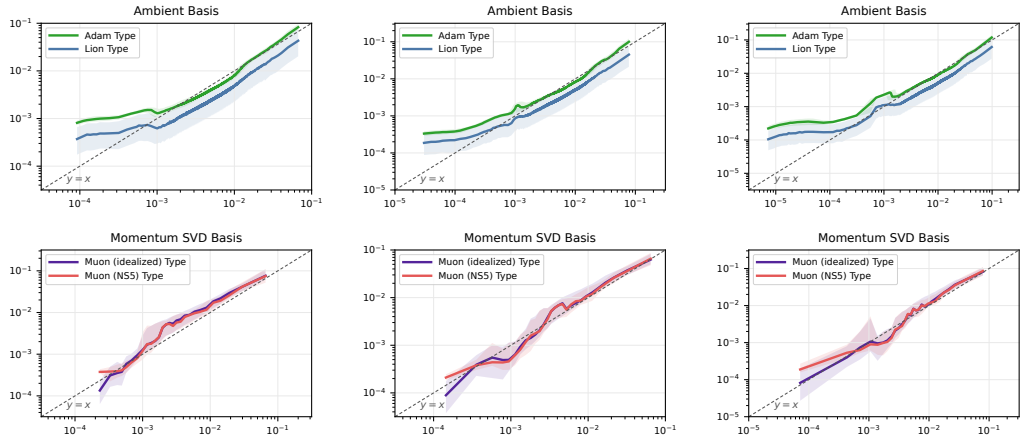
Table 2: Cosine similarity between D^{-1} and $\mathrm{Diag}(H_{\mathrm{GN}}^{Q})^{1/2}$ in each optimizer's induced basis during GPT2-nano training. Each similarity is computed over coordinates with nonzero momentum entries in the corresponding induced basis. All four methods achieve similarities of approximately 0.8 or higher, with Adam approaching 1.

	Ambient Basis		Momentum SVD Basis	
Checkpoint	Adam	Lion	Muon (idealized)	Muon (NS5)
25% steps	0.96	0.79	0.85	0.84
50% steps	0.97	0.79	0.84	0.83
75% steps	0.98	0.79	0.81	0.80

Observation 1: Each optimizer's D^{-1} closely corresponds to $\mathrm{Diag}(H_{\mathrm{GN}}^{Q})^{1/2}$, suggesting that the diagonal preconditioner implicitly exploits the diagonal curvature information available in optimizer's induced basis.

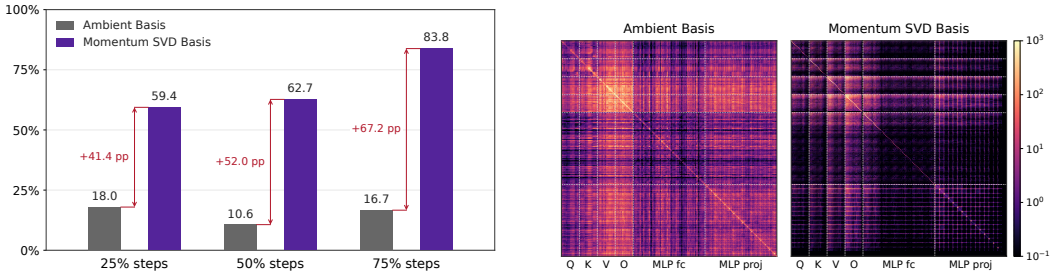
4.2 Stronger GN Diagonal Concentration in the Momentum SVD Basis

Figure 3 shows clear GN diagonal concentration in both the ambient and momentum SVD bases throughout training. In particular, rotating to the momentum SVD basis further strengthens this concentration substantially. We next take a closer look at individual hidden blocks and find that this pattern generally holds across different blocks and layers, as shown in Figure 4.



(a) 25% steps (b) 50% steps (c) 75% steps

Figure 2: Element-wise comparison of D^{-1} with $\mathrm{Diag}(H^Q_{\mathrm{GN}})^{1/2}$ in each optimizer's induced basis during GPT2-nano training. Each D^{-1} is rescaled to account for scale differences, and the comparison is restricted to coordinates with nonzero momentum entries. We find that D^{-1} in Adam and Lion, expressed in the ambient basis, and D^{-1} in Muon (idealized and NS5), expressed in the momentum SVD basis, generally agree with $\mathrm{Diag}(H^Q_{\mathrm{GN}})^{1/2}$, except near the tails. Adam's narrower shaded bands indicate less coordinate-wise variability.



(a) Diagonal concentration $\rho_{\mathrm{diag}}(H^Q_{\mathrm{GN}})$ across bases. (b) Heatmap of the layer-wise H^Q_{GN} across bases.

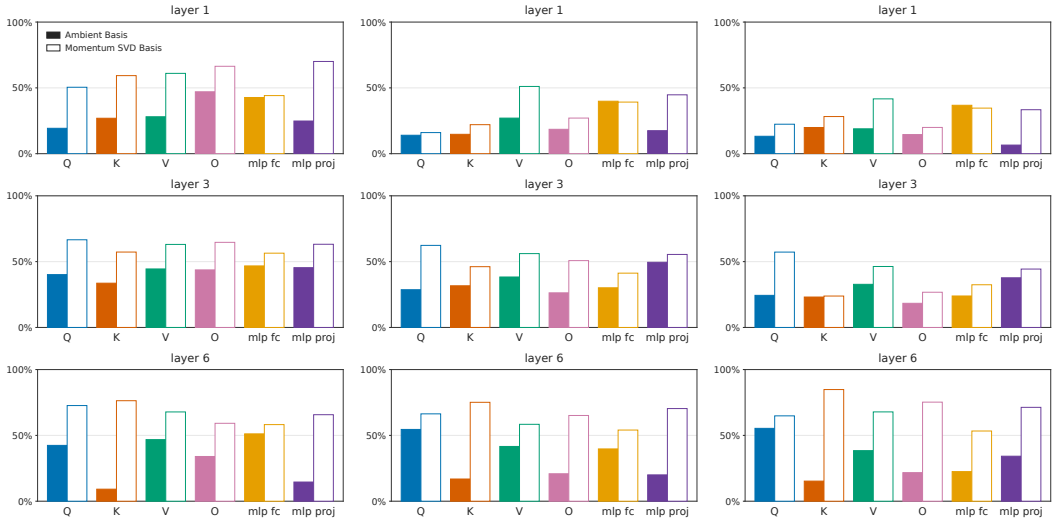
Figure 3: (a) Diagonal concentration $\rho_{\mathrm{diag}}(H^Q_{\mathrm{GN}})$ in the ambient and momentum SVD bases across all hidden parameters of Pythia-70M, where diagonal entries constitute less than 10^{-7} of all entries. (b) Heatmaps of GN for all hidden parameters in the first layer of a small Transformer, in the ambient and momentum SVD bases. The diagonal concentration of H^Q_{GN} is evident in both bases, but stronger in the momentum SVD basis.

To better understand the diagonal concentration in the momentum SVD basis, we establish the gradient–curvature alignment property, supported by extensive empirical observations (Gur-Ari et al., 2018; Fort & Ganguli, 2019; Wang et al., 2026).

Proposition 1 (Gradient alignment with GN curvature)**.** For the average softmax cross-entropy loss, write $H_{\mathrm{GN}} = \sum_{j=1}^{d} \lambda_j q_j q_j^\top$, where $\{q_j\}_{j=1}^{d}$ is an orthonormal basis and $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Define $P_k := \sum_{j=1}^{k} q_j q_j^\top$ and $\gamma := N^{-1} \sum_{i=1}^{N} (e^{\ell_i} - 1)$, where ℓ_i is the loss on sample i at the current parameters. Then:

(1) $g \in \mathrm{Range}(H_{\mathrm{GN}})$, equivalently, $(I - H_{\mathrm{GN}} H_{\mathrm{GN}}^{+})g = 0$.
(2) $\|(I - P_k)g\|_2^2 \leq \gamma \lambda_{k+1}$ for $1 \leq k < d$.

Combining this alignment property with an exact Kronecker assumption on GN curvature, motivated by prior work (Eschenhagen et al., 2023; Morwani et al., 2025), we bound the diagonal Frobenius norm ratio in a suitable gradient SVD basis relative to its expectation over random orthogonal bases.



(a) 25% steps (b) 50% steps (c) 75% steps

Figure 4: Diagonal concentration $\rho_{\text{diag}}(H^Q_{\text{GN}})$ for each individual hidden block of Pythia-70M, where diagonal entries comprise only about 10^{-6} of all entries. For different hidden blocks, the momentum SVD basis generally enhances GN diagonal concentration.

Theorem 2. Under the setting of Proposition 1, assume $H_{\text{GN}} = A \otimes B \neq 0$, where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$ are positive semidefinite. Let Q be uniformly distributed on the orthogonal group (with respect to Haar measure). Then there exists a full SVD $G = U\Sigma V^\top$ such that

$$\frac{\rho_{\text{diag}}(H^{V\otimes U}_{\text{GN}})}{\mathbb{E}_Q\left[\rho_{\text{diag}}(H^Q_{\text{GN}})\right]} \geq \sqrt{\frac{d+2}{3\,\text{rank}(H_{\text{GN}})}}.$$

The ratio gap widens with parameter dimension when H_{GN} is low rank. Without momentum accumulation, the momentum SVD basis reduces to the gradient SVD basis, so this result partly explains the observed GN diagonal concentration.

Unlike the comparison with random bases, improvement over the ambient basis cannot be guaranteed in general, since the ambient basis may already diagonalize the GN curvature. We therefore identify sufficient conditions for strict improvement by the momentum SVD basis in a weighted matrix quadratic (Proposition 2). We also establish sufficient conditions for the population SOAP basis to yield strictly stronger GN diagonal concentration than the ambient basis (Proposition 3). These results are presented in Appendices F.1 and F.2, respectively.

Observation 2: GN is diagonally concentrated in both the ambient and momentum SVD bases, but the concentration is consistently stronger in the latter throughout training.

4.3 GN DIAGONAL CONCENTRATION FAVORS DIAGONAL PRECONDITIONING

Is the degree of GN diagonal concentration in optimizer-induced bases closely linked to optimizer performance? We conduct experiments on different models with four optimizers, including Adam, matrix optimizers like Muon and SOAP with their native basis rotations, and randomly rotated Adam as a control. Results in Figure 5 suggest a strong correlation between them.

To examine how curvature representation and diagonal scaling jointly affect local progress, we compare the optimal GN model decrease with the decrease achievable along a diagonally preconditioned gradient direction. At parameter w, define the GN quadratic model $q_{\text{GN}}(s) := \mathcal{L}(w) + g^\top s + \frac{1}{2}s^\top H_{\text{GN}} s$. For an orthogonal basis Q and a positive-definite diagonal matrix D,

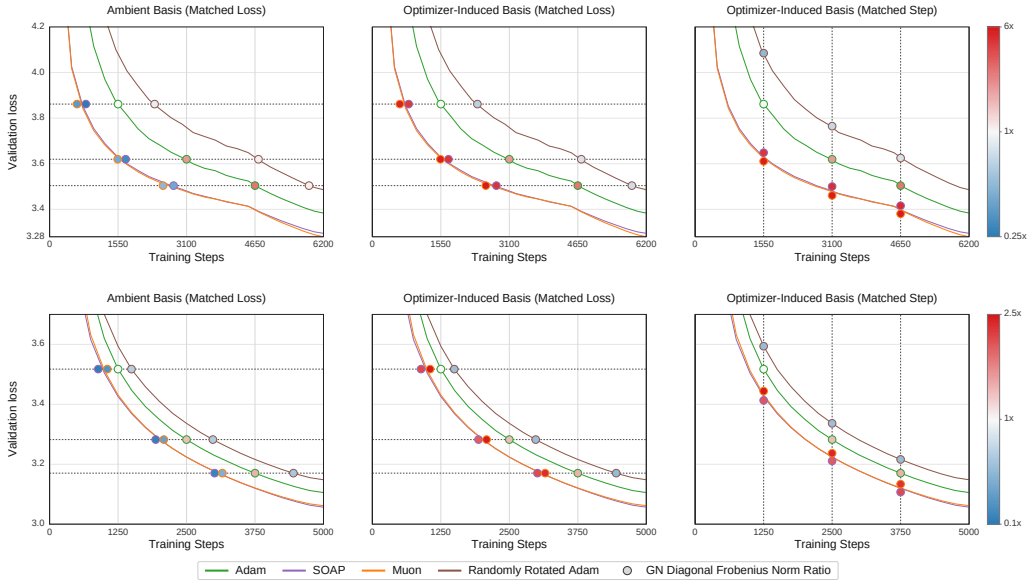


Figure 5: Training curves and GN diagonal Frobenius ratios for modded-nanoGPT 124M (top) and Llama 130M (bottom). All ratios are normalized to the ambient-basis ratio measured after 25% of Adam training ($1\times$) and are encoded by circle colors. The optimizers form three distinct performance tiers that align with the degrees of GN diagonal concentration in their induced bases, with Muon and SOAP exhibiting similar concentration and performance. In the ambient basis, the substantial variation of diagonal Frobenius norm ratios at matched validation loss indicates that different optimizers induce distinct optimization geometries.

we consider

$$\Delta_{\mathrm{GN}} := \mathcal{L}(w) - \min_{s\in\mathbb{R}^d} q_{\mathrm{GN}}(s), \quad \Delta_{Q,D} := \mathcal{L}(w) - \min_{\eta\geq 0} q_{\mathrm{GN}}(-\eta QDQ^\top g).$$

These are the optimal decrease over all steps and the optimal decrease along the preconditioned gradient direction, respectively.

Theorem 3 (GN model descent under diagonal concentration and scaling)**.** Assume $H_{\mathrm{GN}} \succ 0$ and $g \neq 0$. With the definitions above,

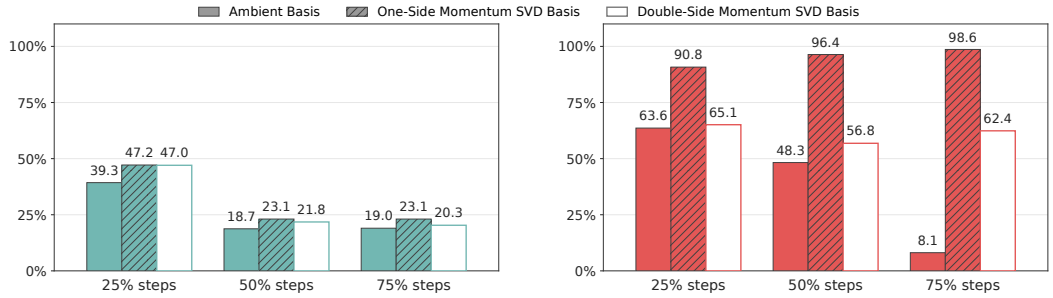
$$1 - \frac{\Delta_{Q,D}}{\Delta_{\mathrm{GN}}} \leq \inf_{\tau>0} \Bigg[\underbrace{\sum_{j=1}^{d} \big(\tau D_{jj}(H^Q_{\mathrm{GN}})_{jj} - 1\big)^2}_{\text{diagonal scaling mismatch}} + \underbrace{\tau^2 \sum_{j\neq k} D_{jj}D_{kk}(H^Q_{\mathrm{GN}})^2_{jk}}_{\text{residual off-diagonal curvature}} \Bigg]$$
$$\leq \inf_{\tau>0} \Bigg[\sum_{j=1}^{d} \big(\tau D_{jj}(H^Q_{\mathrm{GN}})_{jj} - 1\big)^2 + \tau^2 \|D\|_2^2 \|H_{\mathrm{GN}}\|_F^2 \big(1 - \rho_{\mathrm{diag}}(H^Q_{\mathrm{GN}})^2\big) \Bigg]. \tag{3}$$

Here, $\tau > 0$ is a comparison step size that accounts for the overall scale of D. The bound connects GN model decrease to both the curvature representation in Q and the diagonal scaling D. Diagonal concentration enters the bound on the off-diagonal term, together with the scale of D, while the first term measures how well D matches the inverse diagonal curvature. When $D = c\,\mathrm{Diag}(H^Q_{\mathrm{GN}})^{-1}$ for some $c > 0$, choosing $\tau = 1/c$ eliminates the first term, leaving only the off-diagonal curvature weighted by the inverse diagonal scales. If Q also diagonalizes H_{GN}, both terms vanish and the optimally scaled update recovers the full GN step. Appendix E.4 gives the proof.

Observation 3: A key structural factor behind matrix optimizers' advantage is stronger GN diagonal concentration in their induced bases, which favors diagonal preconditioning.

4.4 Extensions: different parameter types and larger models

Non-hidden blocks exhibit distinct curvature geometry. For hidden blocks, we have shown the momentum SVD basis yields stronger GN diagonal concentration than the ambient basis in Figure 3a. Additional experiments in Appendix G (Figure 8) further reveal that rotations on both the input and output sides contribute to this gain, with the strongest concentration when both are applied. Beyond hidden parameter blocks, we extend our analysis to non-hidden parameter matrices, including embeddings and the LM head, and find that their GN curvature geometry differs from that of hidden blocks as shown in Figure 6. This difference may help explain why matrix optimizers often use different update rules for non-hidden blocks (Jordan et al., 2024; Lau & Su, 2026).



(a) Embedding (b) LM head

Figure 6: Diagonal concentration $\rho_{\text{diag}}(H^Q_{\text{GN}})$ for embeddings and the LM head in Pythia-70M under three bases. One-sided momentum SVD rotation acts only on the non-vocabulary dimension. Unlike hidden blocks, these parameter blocks exhibit higher diagonal concentration under one-sided than two-sided rotation, suggesting that adding vocabulary-side rotation reduces diagonal concentration, especially for the LM head.

Scaling to Pythia-410M. We extend our experiments to the larger Pythia-410M model to assess how our findings scale. To our knowledge, this is among the largest models investigated in curvature studies of LLM pretraining (Wang et al., 2026; Zhang et al., 2024; Su, 2025), since curvature matrix size grows quadratically with the parameter dimension. Consistent with our earlier findings, Figure 7's result provides further curvature-based support for Muon's scalability (Liu et al., 2025b).

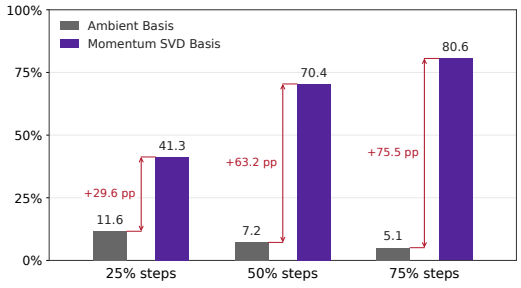


Figure 7: Diagonal concentration $\rho_{\text{diag}}(H^Q_{\text{GN}})$ in the ambient and momentum SVD bases across all hidden blocks of Pythia-410M, where diagonal entries only constitute about 10^{-9} of all entries.

5 Conclusion

We study optimizers from a basis perspective, viewing their updates as basis selection followed by diagonal preconditioning. Our empirical findings and theoretical analysis suggest that matrix optimizers can induce bases with greater Gauss–Newton diagonal concentration, making more curvature information available to diagonal preconditioning. Our experiments are limited to relatively small language models by computational constraints. Future work could extend the analysis to larger models and explore how this perspective can guide the design of more effective optimizers.

AI USE STATEMENT

We used large language models to polish the manuscript, assist with literature searches for related work, and assist with code development. We also used these tools to explore sufficient conditions and assist in developing the proofs in Appendix F.2. The authors checked the AI-assisted proofs and their assumptions and reviewed all AI-assisted content. We take full responsibility for the final manuscript, including its claims, proofs, references, and code.

REPRODUCIBILITY STATEMENT

Appendix D.1 describes the models, datasets, and training configurations. Appendices D.2–D.4 detail the Gauss–Newton measurements and heatmap experiment. Code, experiment configurations, and reproduction instructions are provided in the supplementary material.

REFERENCES

Natalie Abreu, Nikhil Vyas, Sham M. Kakade, and Depen Morwani. The potential of second-order optimization for LLMs: A study with full Gauss-Newton. In *The Fourteenth International Conference on Learning Representations*, 2026.

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.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. Symbolic discovery of optimization algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Ronan Collobert. *Large Scale Machine Learning*. PhD thesis, Université Paris VI, 2004.

Zhaorui Dong, Yushun Zhang, Jianfeng Yao, and Ruoyu Sun. Towards quantifying the Hessian structure of neural networks. *arXiv preprint arXiv:2505.02809*, 2025.

Zhehang Du and Weijie Su. The Newton-Muon optimizer. *arXiv preprint arXiv:2604.01472*, 2026.

Runa Eschenhagen, Alexander Immer, Richard Turner, Frank Schneider, and Philipp Hennig. Kronecker-factored approximate curvature for modern neural network architectures. In *Advances in Neural Information Processing Systems*, volume 36, pp. 33624–33655, 2023. doi: 10.52202/075280-1461.

Runa Eschenhagen, Anna Cai, Tsung-Hsien Lee, and Hao-Jun Michael Shi. Clarifying Shampoo: Adapting spectral descent to stochasticity and the parameter trajectory. *arXiv preprint arXiv:2602.09314*, 2026.

Stanislav Fort and Surya Ganguli. Emergent properties of the local geometry of neural loss landscapes. *arXiv preprint arXiv:1910.05929*, 2019.

Kevin Frans, Pieter Abbeel, and Sergey Levine. What really matters in matrix-whitening optimizers? In *OPT 2025: Optimization for Machine Learning*, 2025a.

Kevin Frans, Sergey Levine, and Pieter Abbeel. A stable whitening optimizer for efficient neural network training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025b.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast approximate natural gradient descent in a Kronecker factored eigenbasis. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via Hessian eigenvalue density. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2232–2241, 2019.

Wenbo Gong, Javier Zazo, Qijun Luo, Puqian Wang, James Hensman, and Chao Ma. ARO: A new lens on matrix optimization for large models. *arXiv preprint arXiv:2602.09006*, 2026.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1842–1850, 2018.

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

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/.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Tim Tsz-Kit Lau and Weijie Su. Symmetry-compatible principle for optimizer design: Embeddings, LM heads, SwiGLU MLPs, and MoE routers. *arXiv preprint arXiv:2605.18106*, 2026.

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.

Bingbin Liu, Rachit Bansal, Depen Morwani, Nikhil Vyas, David Alvarez-Melis, and Sham M Kakade. Adam or Gauss-Newton? a comparative study in terms of basis alignment and SGD noise. *arXiv preprint arXiv:2510.13680*, 2025a.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International Conference on Learning Representations*, volume 2024, pp. 1621–1650, 2024.

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*, 2025b.

James Martens. Deep learning via Hessian-free optimization. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2408–2417, 2015.

James Martens and Ilya Sutskever. Training deep and recurrent networks with Hessian-free optimization. In *Neural Networks: Tricks of the Trade: Second Edition*, pp. 479–535. Springer, 2012.

Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, Sham Kakade, and Alex Damian. A defense of the quadratic model. *arXiv preprint arXiv:2607.21716*, 2026.

Depen Morwani, Itai Shapira, Nikhil Vyas, Eran Malach, Sham M. Kakade, and Lucas Janson. A new perspective on Shampoo's preconditioner. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 49069–49104, 2025.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the Hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Nicol N Schraudolph. Fast curvature matrix-vector products for second-order gradient descent. *Neural Computation*, 14(7):1723–1738, 2002.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of Muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

Weijie Su. Isotropic curvature model for understanding deep learning optimization: Is gradient orthogonalization optimal? *arXiv preprint arXiv:2511.00674*, 2025.

Qian-Yuan Tang, Yufei Gu, Yunfeng Cai, Mingming Sun, Ping Li, Zhou Xun, and Zeke Xie. Investigating the overlooked Hessian structure: From CNNs to LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 58805–58831, 2025.

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.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Dirk Bergemann, and Zhuoran Yang. Why Muon outperforms Adam: A curvature perspective. *arXiv preprint arXiv:2606.04662*, 2026.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, volume 2026, pp. 144731–144838, 2026.

Shuo Xie, Mohamad Amin Mohamadi, and Zhiyuan Li. Adam exploits ℓ_∞-geometry of loss landscape via coordinate-wise adaptivity. In *The Thirteenth International Conference on Learning Representations*, 2025.

Tianyue H. Zhang, Lucas Maes, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding Adam requires better rotation dependent assumptions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need Adam: A Hessian perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik Durk Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. In *International Conference on Learning Representations*, volume 2025, pp. 28033–28063, 2025b.

A More Preliminaries

A.1 Exact Gauss–Newton Diagonal Computation

Let $\mathcal{D}$ denote the data distribution and $g_{x,y} := \nabla_w \ell(f_w(x), y)$. For cross-entropy loss, the Gauss–Newton matrix can be equivalently expressed as the Fisher information matrix under the model's predictive distribution.

$$H_{\mathrm{GN}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\nabla_w f_w(x)\, \nabla_z^2 \ell\, \nabla_w f_w(x)^\top\right] = \mathbb{E}_{\substack{x\sim\mathcal{D}_x \\ c\sim p}}\left[g_{x,c} g_{x,c}^\top\right],$$

where $\mathcal{D}_x$ is the input marginal distribution and c is drawn from the model's predictive distribution p at x. For a fixed orthogonal basis Q, let $H_{\mathrm{GN}}^Q := Q^\top H_{\mathrm{GN}} Q$. Then

$$[H_{\mathrm{GN}}^Q]_{ii} = \mathbb{E}_{\substack{x\sim\mathcal{D}_x \\ c\sim p}}\left[(Q^\top g_{x,c})_i^2\right] = \mathbb{E}_{x\sim\mathcal{D}_x}\left[\sum_c p_c (Q^\top g_{x,c})_i^2\right],$$

which provides a memory-efficient way to compute the exact Gauss–Newton diagonal on a fixed batch by summing probability-weighted squared gradients over all classes and averaging over inputs, without constructing the full Gauss–Newton matrix.

A.2 Diagonal Frobenius Norm Ratio Estimation

By using the matrix-vector-product approach to Frobenius norm estimation, we sample independent Rademacher probes $v_k \in \{-1, +1\}^d$, $k = 1, \ldots, K$, with $K \geq 2$, and compute

$$h_k = H_{\mathrm{GN}} v_k, \qquad a_k = v_k \odot h_k,$$

where $\odot$ denotes element-wise multiplication. The squared Frobenius norm and squared diagonal norm are estimated by

$$\|H_{\mathrm{GN}}\|_F^2 = \mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \|h_k\|_2^2\right], \qquad \|\operatorname{diag}(H_{\mathrm{GN}})\|_2^2 = \mathbb{E}\left[\frac{\sum_{k\neq\ell} a_k^\top a_\ell}{K(K-1)}\right],$$

where the expectations are over the independent Rademacher probes. Consequently, the diagonal Frobenius norm ratio is estimated as

$$\rho_{\mathrm{diag}}(H_{\mathrm{GN}}) = \frac{\|\operatorname{diag}(H_{\mathrm{GN}})\|_2}{\|H_{\mathrm{GN}}\|_F} \approx \sqrt{\frac{\sum_{k\neq\ell} a_k^\top a_\ell}{(K-1)\sum_{k=1}^K \|h_k\|_2^2}}.$$

The same estimation applies in any fixed orthogonal basis Q, since

$$H_{\mathrm{GN}}^Q v = Q^\top H_{\mathrm{GN}}(Qv), \qquad \|H_{\mathrm{GN}}^Q\|_F = \|H_{\mathrm{GN}}\|_F.$$

Thus, replacing h_k with $Q^\top H_{\mathrm{GN}}(Q v_k)$ in the above estimators gives the diagonal Frobenius norm ratio in the Q-basis without constructing the rotated matrix.

A.3 High-Dimensional Random Rotations

Explicitly generating an orthogonal matrix $Q \sim \mathrm{Haar}(\mathcal{O}(d))$ and applying a random rotation requires $O(d^2)$ memory and computation, which is prohibitive in high dimensions. We instead combine random permutations with rotations within small groups as a tractable alternative (Zhang et al., 2025a).

Step 1: Random permutation. For a vector $v \in \mathbb{R}^d$, sample a random permutation matrix P and reorder its coordinates to obtain Pv.

Step 2: Groupwise random rotation. Partition Pv into $m = \lfloor d/b \rfloor$ groups of size $b \ll d$ and a residual group of size $r = d \bmod b$. Sample a random rotation matrix $R \in SO(b)$ and apply it to each full group. If $r > 0$, sample $R_{\mathrm{residual}} \in SO(r)$ to rotate the remaining coordinates.

Step 3: Inverse permutation. Restore the original coordinate ordering and obtain the randomly rotated vector as follows:

$$\tilde{v} = P^\top \begin{bmatrix} R & 0 & \cdots & 0 & 0 \\ 0 & R & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & R & 0 \\ 0 & 0 & \cdots & 0 & R_{\text{residual}} \end{bmatrix} Pv.$$

The permutation and rotation matrices are sampled once at initiation and held fixed. This construction defines an orthogonal transformation using $O(d + b^2)$ storage and $O(db)$ computation, without forming a dense $d \times d$ matrix. It serves as a structured substitute for a uniformly sampled dense rotation.

A.4 Optimizer Algorithms

Algorithm 1 Stochastic Gradient Descent with Momentum (SGD)

Hyperparameters: $\{\eta_t\}$, $\beta \in [0, 1)$, $\lambda \geq 0$.

State: $m_0 = 0$.

Update Rule $t \geq 1$**:**

$$m_t = \beta m_{t-1} + g_t,$$
$$w_{t+1} = w_t - \eta_t m_t - \eta_t \lambda w_t.$$

Algorithm 2 Lion

Hyperparameters: $\{\eta_t\}$, $\beta_1, \beta_2 \in [0, 1)$, $\lambda \geq 0$.

State: $m_0 = 0$.

Update at step $t \geq 1$**:**

$$u_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t,$$
$$w_{t+1} = (1 - \eta_t \lambda) w_t - \eta_t \operatorname{sign}(u_t),$$
$$m_t = \beta_2 m_{t-1} + (1 - \beta_2) g_t.$$

Algorithm 3 Adam

Hyperparameters: $\{\eta_t\}$, $\beta_1, \beta_2 \in [0, 1)$, $\epsilon > 0$, $\lambda \geq 0$.

State: $m_0 = v_0 = 0$.

Update at step $t \geq 1$**:**

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t,$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2,$$
$$\widehat{m}_t = \frac{m_t}{1 - \beta_1^t}, \qquad \widehat{v}_t = \frac{v_t}{1 - \beta_2^t},$$
$$w_{t+1} = (1 - \eta_t \lambda) w_t - \eta_t \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon}.$$

Algorithm 4 Randomly Rotated Adam

Hyperparameters: $\{\eta_t\}, \beta_1, \beta_2 \in [0,1), \epsilon > 0, \lambda \geq 0$.

State: $m_0 = v_0 = 0$ and a fixed random orthogonal basis $Q \sim \mathrm{Haar}(\mathcal{O}(d))$, sampled once at initialization.

Update at step $t \geq 1$:

$$
\begin{aligned}
g_t^Q &= Q^\top g_t, \\
m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t^Q, \\
v_t &= \beta_2 v_{t-1} + (1-\beta_2)(g_t^Q)^2, \\
\widehat{m}_t &= \frac{m_t}{1-\beta_1^t}, \qquad \widehat{v}_t = \frac{v_t}{1-\beta_2^t}, \\
w_{t+1} &= (1-\eta_t\lambda) w_t - \eta_t Q \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t}+\epsilon}.
\end{aligned}
$$

Algorithm 5.1 Muon (Idealized)

Hyperparameters: $\{\eta_t\}, \beta \in [0,1), \lambda \geq 0$.

State: $M_0 = 0$.

Update Rule:

$$
\begin{aligned}
M_t &= \beta M_{t-1} + G_t, \\
B_t &= \beta M_t + G_t, \\
O_t &= \mathrm{Msign}(B_t), \\
s &= \sqrt{\max\{m,n\}}, \\
O_t &\leftarrow s O_t, \\
W_{t+1} &= W_t - \eta_t O_t - \eta_t \lambda W_t.
\end{aligned}
$$

Algorithm 5.2 Muon (NS5)

Hyperparameters: $\{\eta_t\}, \beta \in [0,1), \lambda \geq 0$.

State: $M_0 = 0$.

Update Rule:

$$
\begin{aligned}
M_t &= \beta M_{t-1} + G_t, \\
B_t &= \beta M_t + G_t, \\
O_t &= \mathrm{NewtonSchulz}(B_t, \mathrm{steps} = 5), \\
s &= \sqrt{\max\{m,n\}}, \\
O_t &\leftarrow s O_t, \\
W_{t+1} &= W_t - \eta_t O_t - \eta_t \lambda W_t.
\end{aligned}
$$

Algorithm 6 SOAP

Hyperparameters: $\{\eta_t\}, \beta_1, \beta_2 \in [0,1), \epsilon > 0$, basis refresh interval K, and block size b.

State: M, V, L, R, Q_L, Q_R.

Update Rule:

$$
\begin{aligned}
G_t^Q &= Q_L^\top G_t Q_R, \\
M &\leftarrow \beta_1 M + (1-\beta_1) G_t, \\
M_t^Q &= Q_L^\top M Q_R, \\
V &\leftarrow \beta_2 V + (1-\beta_2)(G_t^Q)^{\odot 2}, \\
\alpha_t &= \eta_t \frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t}, \\
W_{t+1} &= W_t - \alpha_t Q_L \left(\frac{M_t^Q}{\sqrt{V}+\epsilon} \right) Q_R^\top, \\
L &\leftarrow \beta_2 L + (1-\beta_2) G_t G_t^\top, \\
R &\leftarrow \beta_2 R + (1-\beta_2) G_t^\top G_t.
\end{aligned}
$$

If $t \bmod K = 0$, refresh the bases:

$$
Q_L \leftarrow \mathrm{QR}(L Q_L), \qquad Q_R \leftarrow \mathrm{QR}(R Q_R).
$$

Algorithm 7.1 Gauss–Newton (Idealized)

Hyperparameters: $\{\eta_t\}$.

State: None.

Update Rule:

$$\widehat{w}_t = w_t - H_{\mathrm{GN},t}^{-1} g_t,$$

$$w_{t+1} = w_t + \eta_t(\widehat{w}_t - w_t).$$

Algorithm 7.2 Gauss–Newton (Implicit)

Hyperparameters: $\{\eta_t\}$.

State: None.

Update Rule: Approximately solve

$$\widehat{w}_t \approx \arg\min_w \Big[g_t^\top (w - w_t) + \tfrac{1}{2}(w - w_t)^\top H_{\mathrm{GN},t}(w - w_t) \Big],$$

$$w_{t+1} = w_t + \eta_t(\widehat{w}_t - w_t).$$

B Connections to the near–block–diagonal curvature structure

The structure of neural network curvature offers a promising avenue for explaining optimizer performance and guiding optimizer design. Recently, the near-block-diagonal structure of neural network curvature has received considerable attention (Zhang et al., 2024; 2025b; Dong et al., 2025; Collobert, 2004). Our findings offer a closer look at this structure, revealing diagonal concentration within curvature blocks in the ambient basis and stronger concentration in optimizer-induced bases. Specifically, the connection to the near-block-diagonal structure can be understood in the following respects:

- **Further evidence for near-block-diagonal structure of curvature.** Optimizer-induced rotations act within hidden parameter blocks, preserving the block-diagonal Frobenius norm ratio. The diagonal Frobenius norm ratio in each induced basis thus provides a lower bound on this invariant quantity. Our quantitative findings therefore lend strong support to the near-block-diagonal structure of curvature.

- **Within-block curvature structure across bases.** Prior analyses of near-block-diagonal curvature focus on the ambient basis, linking heterogeneity across blocks to the benefits of diagonal preconditioners such as Adam. We extend this perspective by examining within-block curvature structure across optimizer-induced bases, using the diagonal Frobenius norm ratio to track diagonal concentration throughout training. This analysis offers a curvature-based explanation for the performance gains of matrix optimizers such as Muon.

- **Near-block-diagonal structure inherited by the Gauss–Newton surrogate.** We use the Gauss–Newton matrix as the curvature surrogate in this paper. As discussed in Section 3.1, it constitutes the positive semidefinite component of the Hessian decomposition. Our findings show that the Gauss–Newton matrix largely retains the near-block-diagonal structure identified in prior studies of the Hessian.

Together, these connections highlight the close link between curvature structure and optimizer performance, offering guidance for future optimizer design.

C Connection between Gauss–Newton and Hessian

We consider the average loss $\mathcal{L}(w) = N^{-1}\sum_{i=1}^{N} \ell(f_w(x_i), y_i)$ over samples $\{(x_i, y_i)\}_{i=1}^{N}$, with $H = \nabla_w^2 \mathcal{L}(w)$. Linearizing the model while retaining the cross-entropy loss gives

$$\mathcal{L}_{\mathrm{prox}}(w) := \frac{1}{N}\sum_{i=1}^{N} \ell\big(f_w^{\mathrm{lin}}(x_i), y_i\big),\ f_w^{\mathrm{lin}}(x_i) := f_{w_{\mathrm{ref}}}(x_i) + \nabla_w f_{w_{\mathrm{ref}}}(x_i)^\top (w - w_{\mathrm{ref}}).$$

The Hessian of this GN–proximal loss at w_{ref} is $H_{\text{GN}}(w_{\text{ref}})$. Equivalently, the Gauss–Newton matrix retains the first term in the Hessian decomposition

$$H = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \nabla_w f_w(x_i)\, \nabla_z^2 \ell(z_i, y_i)\, \nabla_w f_w(x_i)^\top}_{H_{\text{GN}}} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\sum_{c} \frac{\partial \ell(z_i, y_i)}{\partial z_c}\, \nabla_w^2 f_{w,c}(x_i)}_{\text{Residual term}}.$$

Here, $f_{w,c}(x_i)$ denotes the c-th logit for sample i, and loss derivatives are evaluated at $z_i = f_w(x_i)$. Since cross-entropy is convex in the logits, H_{GN} is positive semidefinite.

D Experiment Details

D.1 Training Configurations for Section 4

Unless otherwise specified, we follow the standard model configurations of the referenced implementations. Muon, SOAP, and randomly rotated Adam operate only on hidden parameter blocks, treating query, key, and value projections separately. Non-hidden parameters use Adam. Randomly rotated Adam applies the procedure in Section A.3 independently within each hidden block, with rotations fixed throughout training. All experiments are run on eight A800 80GB GPUs.

GPT2-nano (11M) on Shakespeare. We use the nanoGPT codebase[3] and its standard Shakespeare training configuration, with batches of 64 sequences of length 256. We train the model with AdamW for 5,000 updates, using $(\beta_1, \beta_2) = (0.9, 0.99)$, $\epsilon = 10^{-8}$, weight decay 0.1, and the learning-rate schedule of the reference configuration.

Pythia (70M and 410M) on The Pile. To strictly align with the original Pythia training settings, all of our Pythia experiments are based on the publicly released Pythia-70M and Pythia-410M checkpoints trained on The Pile, following the original training configuration.[4] Both models were trained using Adam with batches of 1,024 sequences of length 2,048, with optimizer hyperparameters and learning-rate schedules specified in the reference configurations.

Modded-nanoGPT (124M) on FineWeb10B. Our experiments strictly follow the early version of the NanoGPT Track 1 benchmark in the modded-nanoGPT repository.[5] We implement Muon, SOAP, Adam, and randomly rotated Adam, training each for 6,200 updates with batches of 512 sequences of length 1,024 per update. For Adam, we fix $(\beta_1, \beta_2) = (0.9, 0.95)$ and $\epsilon = 10^{-8}$, and grid-search the learning rate over $\{5\times10^{-4}, 1\times10^{-3}, 2\times10^{-3}, 4\times10^{-3}\}$. SOAP and randomly rotated Adam adopt the best Adam configuration, with SOAP refreshing its basis every step. For Muon, we adopt the tuned hyperparameters from its public modded-nanoGPT training logs: hidden layers use Muon with a learning rate of 3.6×10^{-4} and Nesterov momentum 0.95, while the embedding layer uses Adam with a learning rate of 3.6×10^{-3} and $(\beta_1, \beta_2) = (0.9, 0.95)$. All runs use zero weight decay and the reference learning-rate schedule, with no warmup and a linear decay to zero over the final 1,800 updates.

Llama (130M) on C4. We use the GaLore codebase[6] and its Llama-130M architecture configuration. We train on C4 using the T5-base tokenizer for 5,000 updates, with batches of 960 sequences of length 1,024. For Adam, SOAP, and randomly rotated Adam, we fix $(\beta_1, \beta_2) = (0.9, 0.98)$ and $\epsilon = 10^{-8}$, and independently grid-search their learning rates over $\{5\times10^{-4}, 1\times10^{-3}, 2\times10^{-3}, 4\times10^{-3}, 8\times10^{-3}\}$, since randomly rotated Adam diverges at the learning rate selected for Adam. SOAP refreshes its basis at every step. For Muon, hidden layers use Nesterov momentum 0.9 and a learning rate chosen by grid search over $\{2\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}, 2\times10^{-3}\}$. Non-hidden parameters use Adam with the same hyperparameters as the Adam baseline. All runs use zero weight decay and 500 warmup updates followed by linear learning-rate decay to zero.

[3] https://github.com/karpathy/nanoGPT
[4] https://github.com/EleutherAI/pythia
[5] https://github.com/KellerJordan/modded-nanogpt
[6] https://github.com/jiaweizzhao/GaLore

D.2 Implementation Details on Exact Gauss–Newton Diagonal Computation

Using the training batch and optimizer state associated with each checkpoint, we compute exact Gauss–Newton diagonal entries in the ambient and momentum SVD bases, following Appendix A.1. For experiments based on the GPT2-nano model, we analyze the Gauss–Newton matrix restricted to all hidden parameter blocks at steps 1,250, 2,500, and 3,750. Dropout is disabled during measurement to eliminate mask randomness and keep the Gauss–Newton matrix fixed across basis comparisons.

D.3 Implementation Details on Diagonal Frobenius Norm Ratio Estimation

Using the training batch and optimizer state associated with each checkpoint, we estimate the diagonal Frobenius norm ratio in different coordinate bases, following Appendix A.2. For experiments based on the Pythia models, we evaluate the ratio in the ambient and momentum SVD bases. For experiments based on the modded-nanoGPT and Llama models, we consider the ambient and optimizer-induced bases. We apply the estimation to both the full hidden-parameter Gauss–Newton matrix and parameter-wise diagonal blocks corresponding to hidden matrices, embeddings, and the language-model head. Since the fused attention kernels used in our setup do not support the required forward-mode automatic differentiation, we use unfused attention implementations for Gauss–Newton vector products.

D.4 Implementation Details on Figure 3b

This experiment is based on a two-layer Transformer with hidden width 16 and four attention heads, following the reference implementation.[7] We train with AdamW using batch size 4,000, learning rate 10^{-3}, $(\beta_1, \beta_2) = (0.9, 0.999)$, and weight decay 10^{-3}, with 500 warmup updates followed by cosine decay. We compute the exact first-layer Gauss–Newton matrix in the ambient and momentum SVD bases at 50% training steps.

E Proofs of Main-Text Results

E.1 Proof of Proposition 1

Proof. For each sample i, let $z_i := f_w(x_i) \in \mathbb{R}^C$, $p_i := \mathrm{softmax}(z_i)$, and $J_i := \nabla_w f_w(x_i) \in \mathbb{R}^{d\times C}$. With one-hot label y_i, the sample loss is $\ell_i = \ell(z_i, y_i) = -y_i^\top \log p_i$. Set $\delta_i := p_i - y_i$ and $S_i := \mathrm{Diag}(p_i) - p_i p_i^\top$. The gradient and GN matrix of the average loss are

$$g = \frac{1}{N}\sum_{i=1}^{N} J_i \delta_i, \qquad H_{\mathrm{GN}} = \frac{1}{N}\sum_{i=1}^{N} J_i S_i J_i^\top.$$

Because p_i has strictly positive entries,

$$\mathrm{Null}(S_i) = \mathrm{span}\{\mathbf{1}_C\}.$$

Moreover, $\mathbf{1}_C^\top \delta_i = 1 - 1 = 0$ and $S_i \succeq 0$, so

$$\delta_i \in \mathbf{1}_C^\perp = \mathrm{Range}(S_i) = \mathrm{Range}(S_i^{1/2}).$$

Let $a_i = S_i^{+1/2}\delta_i$ and define

$$\mathcal{A} = \frac{1}{\sqrt{N}}\begin{bmatrix} S_1^{1/2} J_1^\top \\ \vdots \\ S_N^{1/2} J_N^\top \end{bmatrix}, \qquad a = \frac{1}{\sqrt{N}}\begin{bmatrix} a_1 \\ \vdots \\ a_N \end{bmatrix}.$$

The gradient vector and Gauss–Newton matrix can be reformulated as

$$H_{\mathrm{GN}} = \mathcal{A}^\top \mathcal{A}, \quad g = \frac{1}{N}\sum_{i=1}^{N} J_i S_i^{1/2} a_i = \mathcal{A}^\top a.$$

[7] https://github.com/fjzzq2002/random_transformers

Since $g \in \mathrm{Range}(\mathcal{A}^\top)$ and for any matrix $\mathcal{A}$, $\mathrm{Range}(\mathcal{A}^\top\mathcal{A}) = \mathrm{Range}(\mathcal{A}^\top)$, we obtain

$$g \in \mathrm{Range}(H_{\mathrm{GN}}) \quad \text{and} \quad \|\mathcal{A}q_k\|_2^2 = q_k^\top \mathcal{A}^\top \mathcal{A} q_k = \lambda_k.$$

Hence, it holds that

$$\begin{aligned}\|(I - P_k)g\|_2^2 = \sum_{t \geq k+1} (q_t^\top g)^2 &= \sum_{t \geq k+1} (q_t^\top \mathcal{A}^\top a)^2 \\ &\leq \lambda_{k+1} \sum_{t \geq k+1, \lambda_t > 0} \frac{\big((\mathcal{A}q_t)^\top a\big)^2}{\|\mathcal{A}q_t\|_2^2} \\ &\leq \lambda_{k+1} \|a\|_2^2.\end{aligned}$$

Meanwhile,

$$\|a\|_2^2 = \frac{1}{N}\sum_{i=1}^N \delta_i^\top S_i^+ \delta_i = \frac{1}{N}\sum_{i=1}^N \delta_i^\top \mathrm{Diag}(p_i)^{-1}\delta_i = \frac{1}{N}\sum_{i=1}^N (e^{\ell_i} - 1) = \gamma.$$

□

E.2 Proof of Theorem 1

Proof. The upper bound follows from the definition, since $\|\mathrm{Diag}(H_{\mathrm{GN}}^Q)\|_F \leq \|H_{\mathrm{GN}}^Q\|_F$, while the lower bound follows from the Cauchy–Schwarz inequality and holds exactly in the isotropic diagonal basis of H_{GN}.

It remains to show that an isotropic diagonal basis always exists for a positive semidefinite matrix. We prove this by induction on d, with the case $d = 1$ being immediate. For $d > 1$, the mean eigenvalue $\mu = \mathrm{tr}(H_{\mathrm{GN}})/d > 0$ lies between the smallest and largest eigenvalues of H_{GN}. Hence there exists a unit vector q_1 such that $q_1^\top H_{\mathrm{GN}} q_1 = \mu$. Complete q_1 to an orthogonal matrix $[q_1, P]$, where $P \in \mathbb{R}^{d\times(d-1)}$. The compression $P^\top H_{\mathrm{GN}} P$ is positive semidefinite and satisfies

$$\mathrm{tr}(P^\top H_{\mathrm{GN}} P) = \mathrm{tr}(H_{\mathrm{GN}}) - q_1^\top H_{\mathrm{GN}} q_1 = (d-1)\mu > 0.$$

By the induction hypothesis, there is an $R \in \mathcal{O}(d-1)$ such that every diagonal entry of $R^\top P^T H_{\mathrm{GN}} P R$ equals $\mathrm{tr}(P^\top H_{\mathrm{GN}} P)/(d-1) = \mu$. Thus $Q_{\mathrm{iso}} = [q_1, PR]$ is orthogonal and satisfies

$$\mathrm{diag}(Q_{\mathrm{iso}}^\top H_{\mathrm{GN}} Q_{\mathrm{iso}}) = \mu \mathbf{1}_d,$$

which establishes the existence of an isotropic diagonal basis and attainability of the lower bound. □

E.3 Proof of Theorem 2

Proof. We organize the proof into three steps.

Step 1: Calculate the upper bound for the denominator. Let $Q = [q_1, \ldots, q_d]$ be Haar-distributed on $\mathrm{O}(d)$. Each column q_a is uniformly distributed on the unit sphere $\mathbb{S}^{d-1}$. For a uniformly distributed vector $q \in \mathbb{S}^{d-1}$, the fourth-moment identity is

$$\mathbb{E}[q_i q_j q_k q_\ell] = \frac{\delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}}{d(d+2)},$$

where δ_{ij} denotes the Kronecker delta. Since H_{GN} is symmetric, this implies

$$\mathbb{E}\big[(q^\top H_{\mathrm{GN}} q)^2\big] = \sum_{i,j,k,\ell=1}^d (H_{\mathrm{GN}})_{ij} (H_{\mathrm{GN}})_{k\ell}\, \mathbb{E}[q_i q_j q_k q_\ell] = \frac{(\mathrm{tr}\, H_{\mathrm{GN}})^2 + 2\|H_{\mathrm{GN}}\|_F^2}{d(d+2)}.$$

Therefore, by linearity of expectation,

$$\mathbb{E}_Q\big[\|\,\mathrm{diag}(H_{\mathrm{GN}}^Q)\|_F^2\big] = \sum_{a=1}^d \mathbb{E}_Q\big[(q_a^\top H_{\mathrm{GN}} q_a)^2\big] = \frac{(\mathrm{tr}\, H_{\mathrm{GN}})^2 + 2\|H_{\mathrm{GN}}\|_F^2}{d+2}.$$

Using orthogonal invariance of the Frobenius norm, we obtain the exact second moment

$$\mathbb{E}_Q\big[\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^Q)^2\big] = \frac{(\operatorname{tr} H_{\mathrm{GN}})^2 + 2\|H_{\mathrm{GN}}\|_F^2}{(d+2)\|H_{\mathrm{GN}}\|_F^2}.$$

Consequently, the Cauchy–Schwarz inequality gives

$$\mathbb{E}_Q\big[\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^Q)\big] \le \sqrt{\mathbb{E}_Q\big[\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^Q)^2\big]} = \frac{\sqrt{(\operatorname{tr} H_{\mathrm{GN}})^2 + 2\|H_{\mathrm{GN}}\|_F^2}}{\sqrt{d+2}\,\|H_{\mathrm{GN}}\|_F}. \tag{4}$$

Step 2: Calculate the lower bound for the numerator. By Proposition 1 and the Kronecker structure assumption on H_{GN}, we have

$$g = \operatorname{vec}(G) \in \operatorname{Range}(H_{\mathrm{GN}}) = \operatorname{Range}(A \otimes B).$$

Thus, for some $Z \in \mathbb{R}^{m\times n}$,

$$\operatorname{vec}(G) = (A \otimes B)\operatorname{vec}(Z).$$

Under column-major vectorization,

$$(A \otimes B)\operatorname{vec}(Z) = \operatorname{vec}(BZA^\top).$$

Since $A = A^\top$, it follows that $G = BZA$. Hence

$$\operatorname{col}(G) \subseteq \operatorname{Range}(B), \qquad \operatorname{row}(G) \subseteq \operatorname{Range}(A).$$

Let $G = U_r \Sigma V_r^\top$ be the compact SVD of the gradient matrix. These inclusions imply

$$\operatorname{span}(U_r) \subseteq \operatorname{Range}(B), \qquad \operatorname{span}(V_r) \subseteq \operatorname{Range}(A).$$

Because $A, B \succeq 0$,

$$\operatorname{Range}(B) = \ker(B)^\perp, \qquad \operatorname{Range}(A) = \ker(A)^\perp.$$

We can therefore complete U_r to an orthonormal basis $U = [U_r,\ U_+,\ U_0]$ such that

$$\operatorname{span}(U_r, U_+) = \operatorname{Range}(B), \qquad \operatorname{span}(U_0) = \ker(B).$$

Likewise, we can complete V_r to $V = [V_r,\ V_+,\ V_0]$ such that

$$\operatorname{span}(V_r, V_+) = \operatorname{Range}(A), \qquad \operatorname{span}(V_0) = \ker(A).$$

The added vectors lie in the zero-singular-value subspaces of G, so U and V form a valid full SVD completion. For $D_{ij} = u_i v_j^\top$,

$$H_{\mathrm{GN}}\operatorname{vec}(D_{ij}) = (A \otimes B)\operatorname{vec}(u_i v_j^\top) = \operatorname{vec}(Bu_i v_j^\top A^\top).$$

Since $A = A^\top$,

$$H_{\mathrm{GN}}\operatorname{vec}(D_{ij}) = \operatorname{vec}(Bu_i v_j^\top A).$$

If $i > \operatorname{rank}(B)$, then $u_i \in \ker(B)$ and $Bu_i = 0$. Thus,

$$H_{\mathrm{GN}}\operatorname{vec}(D_{ij}) = 0.$$

If $j > \operatorname{rank}(A)$, then $v_j \in \ker(A)$ and $v_j^\top A = 0$. Thus,

$$H_{\mathrm{GN}}\operatorname{vec}(D_{ij}) = 0.$$

Consequently,

$$H_{\mathrm{GN}}\operatorname{vec}(D_{ij}) = 0 \quad \text{whenever } i > \operatorname{rank}(B) \text{ or } j > \operatorname{rank}(A).$$

Taking the quadratic form yields

$$\operatorname{vec}(D_{ij})^\top H_{\mathrm{GN}}\operatorname{vec}(D_{ij}) = 0 \quad \text{whenever } i > \operatorname{rank}(B) \text{ or } j > \operatorname{rank}(A).$$

Equivalently,

$$\big(H_{\mathrm{GN}}^{V\otimes U}\big)_{ij,ij} = 0 \quad \text{whenever } i > \operatorname{rank}(B) \text{ or } j > \operatorname{rank}(A). \tag{5}$$

Since $H_{\mathrm{GN}} \succeq 0$, the diagonal entries of H_{GN}^Q are nonnegative. Therefore,

$$\|\operatorname{diag}(H_{\mathrm{GN}}^Q)\|_1 = \operatorname{tr}(H_{\mathrm{GN}}^Q) = \operatorname{tr}(H_{\mathrm{GN}}).$$

Cauchy–Schwarz gives

$$\|\operatorname{diag}(H_{\mathrm{GN}}^Q)\|_1 \le \sqrt{\|\operatorname{diag}(H_{\mathrm{GN}}^Q)\|_0}\,\|\operatorname{diag}(H_{\mathrm{GN}}^Q)\|_F. \tag{6}$$

Finally, combining $\|H_{\mathrm{GN}}^Q\|_F = \|H_{\mathrm{GN}}\|_F$ with Eqs. (5) and (6) yields

$$\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{V\otimes U}) \ge \frac{\operatorname{tr}(H_{\mathrm{GN}})}{\sqrt{\|\operatorname{diag}(H_{\mathrm{GN}}^{V\otimes U})\|_0}\|H_{\mathrm{GN}}\|_F} \ge \frac{\operatorname{tr}(H_{\mathrm{GN}})}{\sqrt{\operatorname{rank}(H_{\mathrm{GN}})}\|H_{\mathrm{GN}}\|_F}, \tag{7}$$

where $\operatorname{rank}(H_{\mathrm{GN}})$ is the rank of H_{GN} with $\operatorname{rank}(H_{\mathrm{GN}}) = \operatorname{rank}(A)\cdot\operatorname{rank}(B)$.

Step 3: Combine the bounds to obtain the final result. Combining (4) and (7) therefore gives

$$\frac{\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{V\otimes U})}{\mathbb{E}_Q\left[\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{Q})\right]} \geq \frac{\mathrm{tr}(H_{\mathrm{GN}})}{\sqrt{\mathrm{rank}(H_{\mathrm{GN}})}\,\|H_{\mathrm{GN}}\|_F} \frac{\sqrt{d+2}\,\|H_{\mathrm{GN}}\|_F}{\sqrt{\mathrm{tr}(H_{\mathrm{GN}})^2 + 2\|H_{\mathrm{GN}}\|_F^2}}$$
$$= \sqrt{\frac{d+2}{\mathrm{rank}(H_{\mathrm{GN}})\left(1 + \dfrac{2\|H_{\mathrm{GN}}\|_F^2}{\mathrm{tr}(H_{\mathrm{GN}})^2}\right)}}.$$

Finally, the eigenvalues $\lambda_1, \ldots, \lambda_d$ of H_{GN} are nonnegative, so

$$\|H_{\mathrm{GN}}\|_F^2 = \sum_{a=1}^{d} \lambda_a^2 \leq \left(\sum_{a=1}^{d} \lambda_a\right)^2 = \mathrm{tr}(H_{\mathrm{GN}})^2.$$

Hence, we conclude that

$$\frac{\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{V\otimes U})}{\mathbb{E}_Q\left[\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{Q})\right]} \geq \sqrt{\frac{d+2}{3\,\mathrm{rank}(H_{\mathrm{GN}})}} = \sqrt{\frac{mn+2}{3\,\mathrm{rank}(H_{\mathrm{GN}})}}.$$

□

E.4 Proof of Theorem 3

Let $P = QDQ^\top$. Minimizing the model q_{GN} defined before Theorem 3 gives $s_{\mathrm{GN}} = -H_{\mathrm{GN}}^{-1}g$. Along $-Pg$, the decrease is $\eta g^\top Pg - \frac{1}{2}\eta^2 g^\top P H_{\mathrm{GN}} Pg$; differentiation with respect to η gives the optimal positive step and the two decreases:

$$\eta^\star = \frac{g^\top Pg}{g^\top P H_{\mathrm{GN}} Pg}, \qquad \Delta_{\mathrm{GN}} = \frac{1}{2} g^\top H_{\mathrm{GN}}^{-1} g, \qquad \Delta_{Q,D} = \frac{(g^\top Pg)^2}{2 g^\top P H_{\mathrm{GN}} Pg}. \tag{8}$$

Define $\widetilde{g} = H_{\mathrm{GN}}^{-1/2} g$ and $K = H_{\mathrm{GN}}^{1/2} P H_{\mathrm{GN}}^{1/2}$. Substitution in (8) gives the exact identity

$$\frac{\Delta_{Q,D}}{\Delta_{\mathrm{GN}}} = \frac{(\widetilde{g}^\top K \widetilde{g})^2}{(\widetilde{g}^\top K^2 \widetilde{g})(\widetilde{g}^\top \widetilde{g})} \in (0, 1], \tag{9}$$

where the upper bound follows from Cauchy–Schwarz. For any comparison step $\tau > 0$, let $\Delta_{Q,D}(\tau) = \tau g^\top Pg - \frac{1}{2}\tau^2 g^\top P H_{\mathrm{GN}} Pg$. Completing the square yields

$$\Delta_{\mathrm{GN}} - \Delta_{Q,D}(\tau) = \frac{1}{2}(\tau Pg - H_{\mathrm{GN}}^{-1} g)^\top H_{\mathrm{GN}} (\tau Pg - H_{\mathrm{GN}}^{-1} g) = \frac{1}{2}\|(\tau K - I_d)\widetilde{g}\|_2^2. \tag{10}$$

Since $\Delta_{Q,D} = \max_{\tau > 0} \Delta_{Q,D}(\tau)$, division by $\Delta_{\mathrm{GN}} = \frac{1}{2}\|\widetilde{g}\|_2^2$ and the definition of the operator norm imply

$$\frac{\Delta_{\mathrm{GN}} - \Delta_{Q,D}}{\Delta_{\mathrm{GN}}} \leq \inf_{\tau > 0} \|\tau K - I_d\|_2^2. \tag{11}$$

To express this bound in rotated coordinates, set $X = H_{\mathrm{GN}}^{1/2} Q D^{1/2}$. The symmetric matrices $XX^\top = K$ and $X^\top X = D^{1/2} H_{\mathrm{GN}}^{Q} D^{1/2}$ have the same positive eigenvalues. Consequently, their shifted operator norms agree. Using $\|\cdot\|_2 \leq \|\cdot\|_F$ and the Frobenius orthogonality of diagonal and off-diagonal matrices gives

$$\begin{aligned} \|\tau K - I_d\|_2^2 &= \|\tau D^{1/2} H_{\mathrm{GN}}^{Q} D^{1/2} - I_d\|_2^2 \\ &\leq \|\tau D^{1/2} H_{\mathrm{GN}}^{Q} D^{1/2} - I_d\|_F^2 \\ &= \sum_j \left(\tau D_{jj} (H_{\mathrm{GN}}^{Q})_{jj} - 1\right)^2 + \tau^2 \sum_{j \neq k} D_{jj} D_{kk} (H_{\mathrm{GN}}^{Q})_{jk}^2. \end{aligned} \tag{12}$$

Finally, $D_{jj}D_{kk} \le \|D\|_2^2$ and $\sum_{j\neq k}(H^Q_{\mathrm{GN}})^2_{jk} = \|H_{\mathrm{GN}}\|_F^2[1-\rho_{\mathrm{diag}}(H^Q_{\mathrm{GN}})^2]$, by orthogonal invariance of the Frobenius norm. Applying these relations to the off-diagonal term and taking the infimum over $\tau > 0$ gives the final inequality in (3). Together with (11)–(12), this proves Theorem 3.

F Sufficient Conditions for Improved GN Diagonal Concentration

This appendix presents the sufficient conditions for stronger GN diagonal concentration than in the ambient basis discussed in Section 4.2. Appendix F.1 considers momentum-SVD and SOAP-Gram bases for a weighted matrix quadratic, while Appendix F.2 studies the population SOAP basis in a neural-network setting.

Throughout this appendix, write $\mathrm{Off}(C) = C - \mathrm{Diag}(C)$ for the off-diagonal part of a square matrix C. For a nonzero symmetric matrix H and an orthogonal basis Q, define

$$\varepsilon(H,Q) = \frac{\|\mathrm{Off}(Q^\top HQ)\|_F}{\|H\|_F}.$$

The orthogonality of the diagonal and off-diagonal projections gives $\rho_{\mathrm{diag}}(H^Q)^2 = 1-\varepsilon(H,Q)^2$.

F.1 Diagonal concentration for a weighted matrix quadratic

We first give sufficient conditions for exact diagonalization by momentum-SVD and SOAP-Gram bases, then show that strict improvement over the ambient basis persists under small perturbations of an aligned history.

Setup. Consider

$$\mathcal{L}_{\mathrm{quad}}(W) = \frac{1}{2}\mathbb{E}[(Wx-y)^\top B(Wx-y)], \qquad A = \mathbb{E}[xx^\top] \succ 0, \quad B \succ 0,$$

with finite second moments of x, y. The unique optimum is $W_\star = \mathbb{E}[yx^\top]A^{-1}$. Let s index a finite history of parameter matrices W_s. Writing $E_s = W_s - W_\star$, the population gradient and Gauss–Newton matrix are

$$G_s = BE_sA, \qquad H_{\mathrm{GN}} = A \otimes B. \tag{13}$$

For a fixed time t, define the momentum $M_t = \sum_s \alpha_{t,s}G_s$, the SOAP Gram states $R_t^{\mathrm{S}} = \sum_s \beta_{t,s}G_s^\top G_s$ and $L_t^{\mathrm{S}} = \sum_s \beta_{t,s}G_sG_s^\top$ with nonnegative weights $\beta_{t,s}$, and $\bar{E}_t = \sum_s \alpha_{t,s}E_s$.

Exact diagonalization under aligned histories.

Proposition 2 (Exact diagonalization for a weighted matrix quadratic)**.** Consider the weighted quadratic objective above with $m = n = r$, finite second moments of x, y, and $A, B \succ 0$. Suppose $A \otimes B$ is not diagonal in ambient coordinates. Write $A = V\Lambda_A V^\top$, $B = U\Lambda_B U^\top$, and suppose the finite history is aligned with these eigenbases:

$$W_s - W_\star = U\Xi_s V^\top, \qquad \Xi_s \text{ diagonal.}$$

Define $F_s = \Lambda_B\Xi_s\Lambda_A$. Under these common assumptions, choose either of the following bases with its corresponding spectral condition:

(1) **Momentum-SVD basis.** The full singular-vector bases $\widehat{U}, \widehat{V}$ of $M_t = \sum_s \alpha_{t,s}\nabla\mathcal{L}_{\mathrm{quad}}(W_s)$, where the diagonal entries of $\sum_s \alpha_{t,s}F_s$ have nonzero, pairwise distinct absolute values.

(2) **SOAP-Gram basis.** The eigenbases $\widehat{U}, \widehat{V}$ of the left and right SOAP Gram states formed from $\nabla\mathcal{L}_{\mathrm{quad}}(W_s)$ with nonnegative weights $\beta_{t,s}$, where the diagonal entries of $\sum_s \beta_{t,s}F_s^2$ are pairwise distinct.

Then $Q = \widehat{V} \otimes \widehat{U}$ diagonalizes H_{GN}, and

$$\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{Q}) = 1 > \rho_{\mathrm{diag}}(H_{\mathrm{GN}}).$$

Proof. Equation (13) gives

$$\nabla\mathcal{L}_{\mathrm{quad}}(W_s) = UF_sV^\top, \qquad H_{\mathrm{GN}} = A \otimes B.$$

In case 1, $M_t = U(\sum_s \alpha_{t,s}F_s)V^\top$; the singular-value condition makes its singular-vector bases equal to U, V up to signs and permutations. In case 2, the left and right Gram states are $U(\sum_s \beta_{t,s}F_s^2)U^\top$ and $V(\sum_s \beta_{t,s}F_s^2)V^\top$; their simple spectra give the same conclusion for their eigenbases. Thus both $\widehat{V}^\top A\widehat{V}$ and $\widehat{U}^\top B\widehat{U}$ are diagonal, so $Q^\top H_{\mathrm{GN}}Q = (\widehat{V}^\top A\widehat{V}) \otimes (\widehat{U}^\top B\widehat{U})$ is diagonal and has diagonal Frobenius ratio one. Since $H_{\mathrm{GN}} = A \otimes B$ is not diagonal in ambient coordinates, its ambient diagonal Frobenius ratio is strictly less than one, proving the comparison. □

Stability under perturbations. We now allow deviations from exact alignment. The comparison is controlled by the separation between reference-state eigenvalues and the size of the state perturbations. For general histories, the state identities below relate displacement errors to the state perturbations used in the following corollary:

$$M_t = B\bar{E}_tA, \quad M_t^\top M_t = A(\bar{E}_t^\top B^2\bar{E}_t)A, \quad M_tM_t^\top = B(\bar{E}_tA^2\bar{E}_t^\top)B,$$

$$R_t^{\mathrm{S}} = A\left(\sum_s \beta_{t,s}E_s^\top B^2E_s\right)A, \qquad L_t^{\mathrm{S}} = B\left(\sum_s \beta_{t,s}E_sA^2E_s^\top\right)B. \tag{14}$$

Corollary 1 (Diagonal concentration under perturbations of an aligned history)**.** In the setting of Proposition 2, fix an aligned reference history $E_s^0 = U\Xi_sV^\top$ satisfying the spectral condition for one of its two cases. Consider perturbed displacements $E_s = U(\Xi_s + Z_s)V^\top$, using the same finite history indices and fixed weights. Write $G_s^0 = BE_s^0A$ and $G_s = BE_sA$. From the perturbed gradients, form

$$(K_L, K_R) = \begin{cases} (M_tM_t^\top, M_t^\top M_t), & \text{in the momentum-SVD case,} \\ (L_t^{\mathrm{S}}, R_t^{\mathrm{S}}), & \text{in the SOAP-Gram case,} \end{cases}$$

and let K_L^0, K_R^0 denote the corresponding reference matrices. Define the minimum reference-state eigenvalue separation γ and the state perturbation size δ by

$$\gamma = \min_{\ell\in\{L,R\}} \min_{i\neq j} |\lambda_i(K_\ell^0) - \lambda_j(K_\ell^0)| > 0,$$
$$\delta = \max_{\ell\in\{L,R\}} \|K_\ell - K_\ell^0\|_F.$$

Let $\widehat{Q} = \widehat{V} \otimes \widehat{U}$ be the basis from the chosen case, formed using the perturbed states. If $\delta < \gamma/4$, then

$$\varepsilon(H_{\mathrm{GN}}, \widehat{Q}) \le 4\sqrt{2}\,\frac{\delta}{\gamma}. \tag{15}$$

In particular, with $\varepsilon_{\mathrm{amb}} := \varepsilon(H_{\mathrm{GN}}, I) > 0$, the condition

$$\delta < \frac{\gamma\varepsilon_{\mathrm{amb}}}{4\sqrt{2}}$$

guarantees $\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{\widehat{Q}}) > \rho_{\mathrm{diag}}(H_{\mathrm{GN}})$.

Proof. The reference construction gives $[A, K_R^0] = [B, K_L^0] = 0$, where $[X, Y] = XY - YX$. Since $\|K_\ell - K_\ell^0\|_2 \le \delta$, the min–max principle bounds the displacement of each ordered eigenvalue by δ. The eigenvalues of each perturbed state are therefore separated by at least $\gamma - 2\delta > \gamma/2$. In the eigenbasis $\widehat{V}$ of K_R, the off-diagonal entries of the commutator are the corresponding entries of $\widehat{V}^\top A \widehat{V}$ multiplied by eigenvalue differences. Consequently,

$$\begin{aligned}(\gamma - 2\delta)\|\operatorname{Off}(\widehat{V}^\top A \widehat{V})\|_F &\le \|[A, K_R]\|_F \\ &= \|[A, K_R - K_R^0]\|_F \le 2\|A\|_2 \delta.\end{aligned}$$

Using $\|A\|_2 \le \|A\|_F$, and applying the same argument to B, K_L, gives

$$\varepsilon_A := \varepsilon(A, \widehat{V}) \le \frac{4\delta}{\gamma}, \qquad \varepsilon_B := \varepsilon(B, \widehat{U}) \le \frac{4\delta}{\gamma}.$$

Since $H_{\mathrm{GN}} = A \otimes B$, the squared relative off-diagonal norm in $\widehat{Q}$ equals $\varepsilon_A^2 + \varepsilon_B^2 - \varepsilon_A^2 \varepsilon_B^2 \le 32\delta^2/\gamma^2$, proving (15). Finally, $\varepsilon_{\mathrm{amb}} \le 1$, so the stated strict-improvement condition implies $\delta < \gamma/4$ and makes the rotated off-diagonal ratio smaller than its ambient value. Orthogonal invariance of $\|H_{\mathrm{GN}}\|_F$ gives the strict diagonal-ratio comparison. □

The quantity δ measures state perturbations, rather than displacement errors directly. The gradient errors are $G_s - G_s^0 = U \Lambda_B Z_s \Lambda_A V^\top$. With the reference history and weights fixed, the general state identities above imply $\delta = O(\max_s \|Z_s\|_F)$ as $\max_s \|Z_s\|_F \to 0$. Thus a positive reference eigenvalue separation and ambient off-diagonal ratio allow nonzero alignment errors. This result concerns prescribed finite histories; it does not establish that an optimizer trajectory satisfies the error bound.

F.2 Population SOAP formation and strict improvement

Setup. Consider a differentiable neural network with logits $f_w(x) \in \mathbb{R}^K$, class probabilities $p_w(x) = \mathrm{softmax}(f_w(x))$, and cross-entropy loss $\ell(f_w(x), y) = -\log p_w(x)_y$. At the fixed parameter w under study, let $W \in \mathbb{R}^{m \times n}$ be a parameter block that occurs once through $s = Wh$, with h independent of W, and assume $y \mid x \sim \mathrm{Categorical}(p_w(x))$. Write $J = \partial f_w(x)/\partial s$, $p = p_w(x)$, and let $C_s = J^\top [\mathrm{Diag}(p) - pp^\top] J$ denote the curvature in the layer-output coordinates s, where $\mathrm{Diag}(p)$ is the diagonal matrix with entries p. The population Gauss–Newton block and the Gram matrices of the single-example gradient $\widehat{G} = \nabla_W \ell(f_w(x), y)$ satisfy

$$H_{\mathrm{GN}} = \mathbb{E}[hh^\top \otimes C_s], \quad R^{\mathrm{S}} := \mathbb{E}[\widehat{G}^\top \widehat{G}] = \mathbb{E}[\mathrm{tr}(C_s) hh^\top], \quad L^{\mathrm{S}} := \mathbb{E}[\widehat{G}\widehat{G}^\top] = \mathbb{E}[\|h\|_2^2 C_s]. \tag{16}$$

Proposition 3 (Strict improvement in the population SOAP basis)**.** Under the population setup above, suppose the following conditions hold:

(1) **Gaussian inputs and curvature dependence.** Let $h = A^{1/2}\xi$, where $\xi \sim N(0, I_n)$, $A \succeq 0$, and $A \neq 0$. For some $T \in \mathbb{R}^{n \times d_c}$ with $T^\top T = I_{d_c}$, C_s is a measurable positive-semidefinite function of $u = T^\top \xi$ alone.

(2) **Moment and dimension conditions.** The matrix $B := \mathbb{E}[C_s] \neq 0$ is finite and, for some $\kappa_c \ge 0$,

$$\mathbb{E}[\|u\|_2^2 C_s] \preceq \kappa_c B, \qquad s_A := \frac{\|A\|_F^2}{\|A\|_2^2} > d_c.$$

Let V, U be any orthogonal eigenbases of $R^{\mathrm{S}}, L^{\mathrm{S}}$, respectively, and set $Q_{\mathrm{S}} = V \otimes U$. Define

$$b_c = \frac{(1+\sqrt{2})\kappa_c}{\sqrt{s_A - d_c}}, \qquad e_c = \frac{\kappa_c + \sqrt{d_c}}{\sqrt{s_A}}, \qquad a_A = \varepsilon(A, I).$$

Here b_c bounds the relative off-diagonal norm in the SOAP basis, e_c bounds the Frobenius error of the approximation $H_{\mathrm{GN}} \approx A \otimes B$ relative to $\|A\|_F \|B\|_F$, and a_A measures the relative off-diagonal norm of A in ambient coordinates. If

$$a_A > e_c + (1 + e_c) b_c, \tag{17}$$

then

$$\rho_{\mathrm{diag}}(H_{\mathrm{GN}}^{Q_{\mathrm{S}}}) > \rho_{\mathrm{diag}}(H_{\mathrm{GN}}).$$

Population gradient identities. In the single-occurrence setting of (16), let e_y be the one-hot label vector and $S = \mathrm{Diag}(p) - pp^\top$. The chain rule gives $\widehat{G} = J^\top(p - e_y)h^\top$. Under the stated conditional label law, $\mathbb{E}[(p-e_y)(p-e_y)^\top \mid x] = S$, so the conditional second moment of $J^\top(p-e_y)$ is $C_s = J^\top S J$. Taking the two gradient Gram expectations yields (16); the identity for H_{GN} follows from the block Jacobian $h^\top \otimes J$. For a shared matrix, the block Jacobian is a sum over occurrences and its Gauss–Newton matrix also contains cross-occurrence terms.

Theorem 4 (Population SOAP formation under Gaussian curvature coupling)**.** Let $h = A^{1/2}\xi$, $\xi \sim N(0, I_n)$, with $A \succeq 0$, $A \neq 0$. Suppose $T^\top T = I_{d_c}$, $C_s = C_s(T^\top \xi) \succeq 0$ is measurable, $B = \mathbb{E}[C_s] \neq 0$ is finite, and $\mathbb{E}[\|T^\top \xi\|_2^2 C_s] \preceq \kappa_c B$. Define $H_{\mathrm{GN}} = \mathbb{E}[hh^\top \otimes C_s]$, $R = \mathbb{E}[\mathrm{tr}(C_s) hh^\top]$, and $L = \mathbb{E}[\|h\|_2^2 C_s]$. Let V, U be arbitrary orthogonal eigenbases of R, L and $Q = V \otimes U$. If $s_A = \|A\|_F^2/\|A\|_2^2 > d_c$, then

$$\varepsilon(H_{\mathrm{GN}}, Q) \le \frac{(1+\sqrt{2})\kappa_c}{\sqrt{s_A - d_c}}, \qquad \frac{\|H_{\mathrm{GN}} - A \otimes B\|_F}{\|A\|_F \|B\|_F} \le \frac{\kappa_c + \sqrt{d_c}}{\sqrt{s_A}}. \tag{18}$$

Proof. Put $u = T^\top \xi$, $C = A^{1/2}T$, $A_\parallel = CC^\top$, and $A_\perp = A - A_\parallel$. The Gaussian decomposition $h = Cu + \zeta$ has ζ independent of u, with mean zero and covariance $A_\perp$. Hence

$$\begin{aligned} H_{\mathrm{GN}} &= A_\perp \otimes B + F, & F &= \mathbb{E}[Cuu^\top C^\top \otimes C_s], \\ R &= \mathrm{tr}(B) A_\perp + F_R, & F_R &= \mathbb{E}[\mathrm{tr}(C_s) Cuu^\top C^\top], \\ L &= \mathrm{tr}(A_\perp) B + F_L, & F_L &= \mathbb{E}[\|Cu\|_2^2 C_s]. \end{aligned}$$

Write $a = \|A\|_2$. For an independent copy (u', C_s'), positivity of C_s, C_s' and $\|C^\top C\|_2 \le a$ give

$$\begin{aligned} \|F\|_F^2 &= \mathbb{E}\mathbb{E}'[(u^\top C^\top C u')^2 \,\mathrm{tr}(C_s C_s')] \\ &\le a^2 \mathbb{E}\mathbb{E}'[\|u\|_2^2 \|u'\|_2^2 \,\mathrm{tr}(C_s C_s')] = a^2 \|\mathbb{E}[\|u\|_2^2 C_s]\|_F^2. \end{aligned}$$

The moment assumption, monotonicity of the Frobenius norm on positive-semidefinite matrices, and $\|F_R\|_F \le \mathrm{tr}(F_R)$ imply

$$\|F\|_F \le a\kappa_c \|B\|_F, \quad \|F_R\|_F \le a\kappa_c \,\mathrm{tr}(B), \quad 0 \preceq F_L \preceq a\kappa_c B.$$

Since V, U diagonalize R, L, respectively,

$$\|\,\mathrm{Off}(V^\top A_\perp V)\|_F \le a\kappa_c, \qquad \|\,\mathrm{Off}(U^\top B U)\|_F \le \frac{a\kappa_c \|B\|_F}{\mathrm{tr}(A_\perp)}.$$

The denominator is positive: $\mathrm{rank}(A_\parallel) \le d_c$ and the best rank-d_c approximation bound imply

$$\|A_\perp\|_F^2 \ge \sum_{i > d_c} \lambda_i(A)^2 \ge \|A\|_F^2 - d_c a^2 = a^2(s_A - d_c) > 0.$$

For any square X, Y,

$$\|\,\mathrm{Off}(X \otimes Y)\|_F^2 = \|\,\mathrm{Off}(X)\|_F^2 \|Y\|_F^2 + \|\,\mathrm{Diag}(X)\|_F^2 \|\,\mathrm{Off}(Y)\|_F^2.$$

Applying this identity and $\|A_\perp\|_F \le \mathrm{tr}(A_\perp)$ yields

$$\|\,\mathrm{Off}(Q^\top (A_\perp \otimes B) Q)\|_F \le \sqrt{2}\, a\kappa_c \|B\|_F.$$

Thus $\|\,\mathrm{Off}(Q^\top H_{\mathrm{GN}} Q)\|_F \le (1+\sqrt{2}) a\kappa_c \|B\|_F$. Both summands of $H_{\mathrm{GN}} = A_\perp \otimes B + F$ are positive semidefinite, so $\|H_{\mathrm{GN}}\|_F \ge \|A_\perp\|_F \|B\|_F \ge a\sqrt{s_A - d_c}\|B\|_F$, proving the first bound. Finally, $H_{\mathrm{GN}} - A \otimes B = F - A_\parallel \otimes B$ and $\|A_\parallel\|_F \le a\sqrt{d_c}$ give the second bound. □

Proof of Proposition 3. The population identities identify its Gram matrices with those of Theorem 4. Write $H_{\mathrm{GN}} = A \otimes B + E$. The theorem gives $\varepsilon(H_{\mathrm{GN}}, Q_{\mathrm{S}}) \le b_c$ and $\|E\|_F \le e_c \|A\|_F \|B\|_F$. Since $\|\,\mathrm{Off}(A \otimes B)\|_F \ge \|\,\mathrm{Off}(A)\|_F \|B\|_F$, the triangle inequalities imply

$$\varepsilon(H_{\mathrm{GN}}, I) \ge \frac{[a_A - e_c]_+}{1 + e_c}.$$

Condition (17) therefore gives $\varepsilon(H_{\mathrm{GN}}, Q_{\mathrm{S}}) < \varepsilon(H_{\mathrm{GN}}, I)$. The identity relating ε and ρ_{diag} proves the conclusion. □

G Additional Experiments

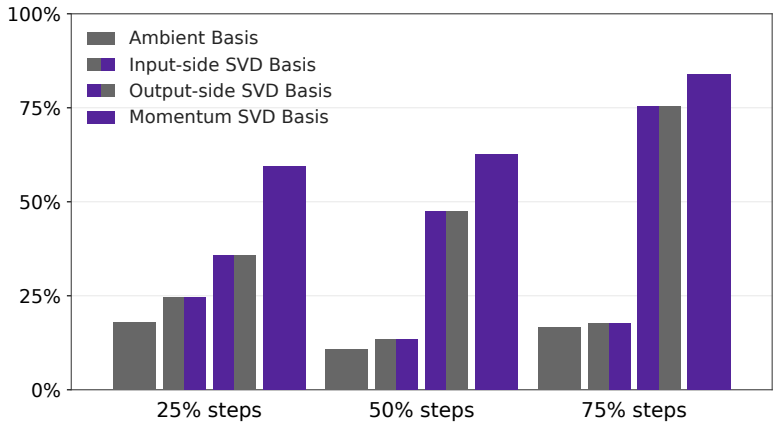


Figure 8: Effect of momentum-SVD rotations on GN diagonal concentration in Pythia-70M. Alongside the ambient and two-sided momentum-SVD bases, we show input-only and output-only rotations. Both one-sided rotations strengthen diagonal concentration, particularly on the output side; double-sided rotation yields the strongest effect.

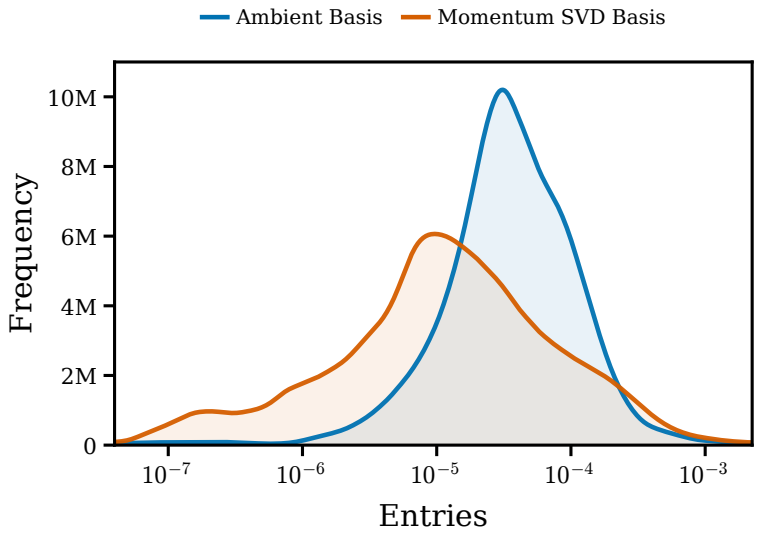


Figure 9: Distribution of GN diagonal entries in the ambient and momentum-SVD bases for GPT-2 Nano (11M) at 50% of training. The greater variation in the momentum-SVD basis is consistent with stronger diagonal concentration.