How Bregman Divergences Shape Shampoo

Anonymous authors
Paper under double-blind review

Abstract

Understanding the principles behind Shampoo has recently guided the development of more effective neural network optimizers. These methods learn a preconditioner by optimizing the Frobenius or Kullback–Leibler (KL) divergence against the gradient second moment. In this work, we investigate how the choice of divergence shapes preconditioning, which remains unclear and blocks further improvements. To do so, we develop a unified Bregman divergence framework that connects all popular divergences, allowing us to study them jointly. Through spectral analysis and empirical investigation on the gradient second moment, we examine how divergence choice influences approximation behavior and preconditioner estimation. We find that some divergences can compensate for the bias introduced by approximating the gradient second moment with a finite number of gradients, and that this helps explain the behavior of their Shampoo variant in GPT-2 pretraining and synthetic experiments. By connecting divergence choice to practical training behavior, we believe our framework provides principled guidance for understanding the foundations of, and further improving, Shampoo.

1 Introduction

Shampoo's empirical success (Dahl et al., 2023; Kasimbeg et al., 2025) raises a fundamental question: what makes its preconditioner effective? This question has motivated a growing body of work (Morwani et al., 2025; Vyas et al., 2025; An et al., 2025; Eschenhagen et al., 2025; 2026; Lin et al., 2026) with the goal of developing more effective neural network optimizers. A promising approach is to view Shampoo's update scheme as a Kronecker-factored approximation of the full-matrix gradient second moment via Bregman divergence minimization (Lin et al., 2026). This approach recovers existing variants and produces new ones, including F-Shampoo with Frobenius divergence, VN-Shampoo with von Neumann divergence, and KL-Shampoo with Kullback-Leibler divergence. Recent studies have reported strong performance of KL-Shampoo in LLM pretraining (Eschenhagen et al., 2026; Zhang et al., 2026). However, the variants derived from different divergences perform very differently in practice (Figure 1 right). The framework of Lin et al. (2026) does not explain this performance gap. Given that improvements to Shampoo have been driven by understanding its principles, we argue that explaining this gap is essential for guiding the design of better variants, since many other Bregman divergences remain to be explored and evaluated. For example, preconditioned stochastic gradient descent (PSGD) (Li, 2018; 2019) is a related method that has shown promising results (Frans et al., 2025) but whose success remains poorly understood. Surprisingly, we show that PSGD's strong performance can also be understood through its connection to a new Bregman divergence. Thus, our central question is:

How and why do different divergence objectives lead to differences in training performance?

Two effects are mixed together in practice. A Kronecker approximation cannot generally match the second moment exactly, so the divergence determines how approximation error is allocated; with limited stochastic gradients, the Kronecker factors are also estimated with noise. At the same time, the empirical second moment being approximated is also a finite-sample estimate and may be unreliable. We therefore first isolate the effect of the divergence, then study the reliability of the empirical second moment, and finally examine how the two interact in inverse-root preconditioning.

This leads to the following three questions:

(Q1) *How does divergence choice shape approximation errors and sensitivity to sampling noise?*

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.

Table 1: Overview of the divergence properties we study in this work.

Property \ Divergence	Frobenius	von Neumann	Square root	KL
Trace-power exponent q_{D}	2	1 (limit)	$1/2$	0 (limit)
Weight on large eigenvalues	Strong	Moderate	Mild	None
Penalized more	Overestimation	Overestimation	Symmetric	Underestimation
Error with few samples	Highest	Higher	Lower	Lowest

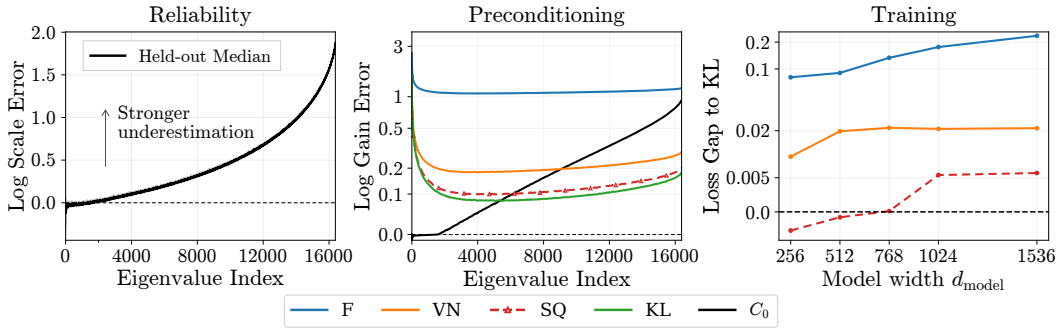


Figure 1: **Divergences shape preconditioning and training.** Left and middle: One GPT-2 checkpoint (output_proj) at step 2500, ordered by decreasing eigenvalue. Left: Empirical second-moment (C_0) eigenvalues become less reliable toward the tail, increasingly underestimating held-out scales (Section 4). Middle (smoothed curves, asinh y-axis for better resolution near zero): KL stays closest to the held-out target (dashed zero line), followed by SQ, compensating C_0's scale errors (Section 5). Right: Final GPT-2 validation loss relative to KL with asinh y-axis; lower is better (Section 6).

(Q2) *How reliable is the empirical gradient second moment being approximated?*

(Q3) *How do divergence properties and second-moment reliability jointly affect preconditioning?*

Our contributions answer these questions in order:

(C1) **A unified analysis of Kronecker approximation.** We unify the three divergences from Lin et al. (2026) into a *trace-power Bregman family* and introduce a fourth one, which recovers the gradient-whitening objective of PSGD and yields a new method, SQ-Shampoo. We show how the family exponent determines which parts of the spectrum are prioritized, whether over- or underestimation is penalized more, and how stable the factors are with limited samples (Table 1).

(C2) **Finite-sample bias in empirical gradient second moments.** On GPT-2, we show that empirical gradient second moments severely underestimate tail eigenvalues even when gradients outnumber dimensions (Figure 1, left). Thus, each divergence approximates an already biased target, leading to an interplay between approximation preferences and this sampling bias.

(C3) **Explaining differences in inverse-root preconditioning.** Last, we investigate how the interplay of the approximation bias of a divergence (C1) and the finite-sample bias (C2) propagates through preconditioning (Figure 1, middle). KL best compensates for the underestimated scales with the lowest gain error, closely followed by SQ (Table 2). The properties of these divergences help explain these empirical differences. Consistently, GPT-2 pretraining yields lower validation losses for SQ and KL than for F and VN, with KL's advantage growing for larger widths.

Together, these results help explain how divergences affect preconditioning and training performance, and provide practical criteria for designing better Kronecker-factored optimizers.

2 Background

Notation. Let $G \in \mathbb{R}^{m \times n}$ be the stochastic gradient of a matrix parameter Θ. We use row-wise vectorization $g = \mathrm{vec}(G)$ and write $C := \mathbb{E}[gg^\top] \in \mathbb{S}^{mn}_{++}$ for its gradient second moment, where $\mathbb{S}^d_{++}$ denotes symmetric positive-definite (SPD) matrices. For simplicity, we omit momentum, as in prior work. Appendix A discusses related work in detail.

Kronecker Approximation and Preconditioning For factors $L \in \mathbb{S}^m_{++}, R \in \mathbb{S}^n_{++}$, we consider methods that approximate C by a Kronecker product $S := L \otimes R$, with inverse-root preconditioning:

$$S^{-1/2} g = (L^{-1/2} \otimes R^{-1/2}) g = \mathrm{vec}(L^{-1/2} G R^{-1/2}).$$

Shampoo. Shampoo (Gupta et al., 2018; Anil et al., 2020) maintains averages L_t, R_t of $G_t G_t^\top$ and $G_t^\top G_t$. Its preconditioning step is $\Theta_{t+1} = \Theta_t - \eta_t L_t^{-p} G_t R_t^{-p}$. Motivated by gradient second-moment approximation, modern Shampoo variants (Shi et al., 2023; Morwani et al., 2025; Eschenhagen et al., 2025) use $p = 1/2$, while original Shampoo (Gupta et al., 2018) uses $p = 1/4$.

PSGD. Preconditioned stochastic gradient descent (PSGD) (Li, 2018; 2019) learns a preconditioner P that acts on the stochastic gradient as Pg. Its Fisher-type gradient-whitening objective is

$$\min_{P \in \mathbb{S}^{mn}_{++}} \ \mathrm{Tr}(PC) + \mathrm{Tr}(P^{-1}), \tag{1}$$

with unique dense solution $P = C^{-1/2}$. PSGD-Kron restricts the preconditioner to a Kronecker product $P = P_L \otimes P_R$ and learns the factors through relative-gradient updates (Li, 2019).

Although PSGD and Shampoo were developed independently, we will show in the next section that PSGD-Kron is closely related to a Shampoo variant under a new Bregman divergence. This allows us to understand them and explain their strong performance in a unified way.

Bregman matrix divergences. For a strictly convex, differentiable $f : (0, \infty) \to \mathbb{R}$, we apply f spectrally to SPD matrices: if $M = U \,\mathrm{diag}(\mu_i) U^\top$, then $f(M) = U \,\mathrm{diag}(f(\mu_i)) U^\top$ and $\Phi(M) = \mathrm{Tr}(f(M))$. For $X, Y \in \mathbb{S}^d_{++}$, the Bregman matrix divergence (Dhillon & Tropp, 2008) is

$$\mathcal{B}_\Phi(X, Y) = \Phi(X) - \Phi(Y) - \langle \nabla\Phi(Y), X - Y \rangle_F. \tag{2}$$

where $\langle A, B \rangle_F = \mathrm{Tr}(A^\top B)$. It is nonnegative and vanishes exactly when $X = Y$. Choosing $f(x) = x^2/2$, $x \log x - x$, or $-\log x$ gives Frobenius (F), von Neumann (VN), or Kullback–Leibler (KL) divergence, respectively. The KL divergence is also known as the LogDet divergence in the literature. Appendix B gives matrix formulas for these divergences.

Deriving Shampoo methods via Bregman minimization. Lin et al. (2026) derive F-, VN-, and KL-Shampoo variants from the stationarity condition of the following Bregman objective:

$$\min_{L \in \mathbb{S}^m_{++},\, R \in \mathbb{S}^n_{++}} \mathcal{B}_\Phi(C, L \otimes R),$$

with different generators Φ. However, these variants perform differently in practice (Figure 1, right) when training neural networks, and these generators differ substantially. The relationship between the choice of Φ and empirical performance remains unknown, as the authors do not explain it. Understanding this connection may guide the design of better Kronecker-factored optimizers.

3 A Trace-Power Family of Kronecker Preconditioners

The objectives in Section 2 look different, although they all construct Kronecker preconditioners from the gradient second moment. We show that they can be unified in a common trace-power Bregman family, with the SQ member connecting Shampoo and PSGD. This lets us study how divergence choice shapes approximation errors and finite-sample stability.

Unified analysis of Bregman objectives. We formulate these objectives through a *trace-power Bregman family*, obtained by extending scalar divergence generators (Cichocki & Amari, 2010) to SPD matrices via spectral functions (Amari, 2014). Our unification further clarifies the choice of the generators and lets us connect PSGD-Kron and study its related variant within the same framework.

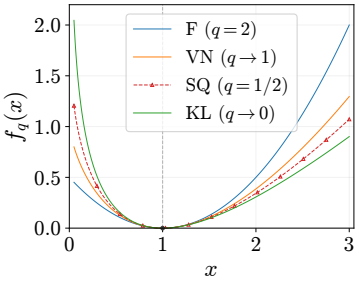


Figure 2: The family of scalar generators $f_q(x)$ inducing trace-power Bregman divergences we study through one unifying lens.

Proposition 1 (Unification of Shampoo and PSGD-Kron via the Trace-Power Bregman Objective)**.** *The objectives of F-, VN-, and KL-Shampoo and PSGD-Kron's gradient-whitening objective belong to the same trace-power Bregman family, generated by* $\Phi_q(M) = \mathrm{Tr}(f_q(M))$*, where* f_q *is applied spectrally to the eigenvalues of* M *as defined in Section 2:*

$$f_q(x) := \frac{x^q - qx + q - 1}{q(q-1)} \ (q \neq 0, 1),\ f_1(x) := x\log x - x + 1,\ f_0(x) := -\log x + x - 1,$$

Let q_{D} *denote the exponent for divergence* D*. Frobenius has* $q_{\mathrm{F}} = 2$*, while von Neumann and KL divergences arise as* $q \to 1$ *and* $q \to 0$*, respectively, for which we denote the exponents by* $q_{\mathrm{VN}} = 1$ *and* $q_{\mathrm{KL}} = 0$*. The square-root (SQ) member,* $q_{\mathrm{SQ}} = 1/2$*, recovers PSGD-Kron's gradient-whitening objective equation 1 under* $P = S^{-1/2}$*, up to a positive scale and an additive constant. Figure 2 shows the four scalar generators.*

Setting $S = L \otimes R$ *gives the conditional minimizers:*

$$L_q(R) = \frac{\mathbb{E}[GR^{q-1}G^\top]}{\mathrm{Tr}(R^q)},\ R_q(L) = \frac{\mathbb{E}[G^\top L^{q-1}G]}{\mathrm{Tr}(L^q)}. \tag{3}$$

Every joint stationary point satisfies these coupled equations. See proof in Appendix B.1.

Square-root (SQ) Shampoo. To better understand how the choice of Bregman divergence shapes optimization, we consider a new Shampoo variant—*square-root (SQ)-Shampoo*—that is induced by the Bregman family and recovers PSGD's objective function by setting $q = 1/2$ in Equation (3). We consider this variant because the existing PSGD method does not employ a Shampoo-like scheme.

We consider four divergences in this work. For a fair comparison, we adopt the estimation technique suggested by Lin et al. (2026) in our study of divergence-induced update schemes, which reduces cost and resembles Shampoo's update scheme. Implementation details are given in Appendix C.

We next examine limited Kronecker capacity and finite-sample stability for these divergences in order to understand their role in preconditioner estimation.

3.1 Spectral Error Allocation under Limited Kronecker Capacity

As our theory (Appendix B.2) and empirical results (Figure 5, left) show, a single Kronecker product generally cannot match C exactly. The divergence therefore determines how the remaining approximation errors are penalized. These penalty properties are not specific to gradient second moments, so we first derive them for arbitrary SPD matrices to isolate the effect of the divergence itself.

Proposition 2 (Spectral Error Decomposition and Scaling)**.** *Let* $A = U\Lambda U^\top, B = V\Omega V^\top \in \mathbb{S}^d_{++}$ *have eigenpairs* (λ_i, u_i) *and* (ω_i, v_i) *ordered by decreasing eigenvalue, and set* $P_{ij} = \langle u_i, v_j\rangle^2$*. Writing* d_f *for the scalar Bregman divergence generated by* f*, we have*

$$\mathcal{B}_\Phi(A, B) = \underbrace{\textstyle\sum_{i=1}^d d_f(\lambda_i, \omega_i)}_{\text{eigenvalue mismatch}} + \underbrace{\Delta_f(\Lambda, \Omega; P)}_{\text{eigenspace misalignment}},\quad \Delta_f(\Lambda, \Omega; P) \geq 0,$$

For the trace-power scalar generator f_q*, write* $d_q := d_{f_q}$ *and* $\Delta_q := \Delta_{f_q}$*. For any* $r > 0$*, both mismatch costs satisfy*

$$d_q(\lambda, r\lambda) = \lambda^q d_q(1, r),\quad \Delta_q(r\Lambda, r\Omega; P) = r^q \Delta_q(\Lambda, \Omega; P). \tag{4}$$

Thus q *controls how strongly both mismatch costs depend on eigenvalue scale.*

Eigenspace mismatch. Keeping the eigenvectors fixed (and hence P unchanged), multiplying all eigenvalues of both matrices by $r > 1$ multiplies the eigenspace-mismatch cost by r^q. Thus larger q places greater relative weight on the same eigenspace misalignment at larger eigenvalue scales.

Eigenvalue (spectral) mismatch. Writing $d_{\mathrm{F}} = d_2$, $d_{\mathrm{VN}} = d_1$, $d_{\mathrm{SQ}} = d_{1/2}$, and $d_{\mathrm{KL}} = d_0$, the scalar eigenvalue penalties are

$$\begin{aligned} d_{\mathrm{F}}(\lambda, r\lambda) &= \tfrac{1}{2}\lambda^2(r-1)^2, & d_{\mathrm{VN}}(\lambda, r\lambda) &= \lambda(r - 1 - \log r), \\ d_{\mathrm{SQ}}(\lambda, r\lambda) &= 2\lambda^{1/2}\left(\sqrt{r} + 1/\sqrt{r} - 2\right), & d_{\mathrm{KL}}(\lambda, r\lambda) &= \log r + 1/r - 1. \end{aligned} \tag{5}$$

For fixed relative mismatch r, the weights $\lambda^2, \lambda, \lambda^{1/2}, 1$ show that larger q puts more emphasis on large eigenvalues (Figure 3, left). We call their associated eigenvectors *leading directions*.

Over- versus under-estimation and asymmetry. For $r > 1$, the fitted values $r\lambda$ and λ/r over- and under-estimate λ by the same multiplicative factor. Their penalties satisfy (Figure 3, right)

$$\begin{aligned} d_{\mathrm{F}}(\lambda, r\lambda) &> d_{\mathrm{F}}(\lambda, \lambda/r), \\ d_{\mathrm{VN}}(\lambda, r\lambda) &> d_{\mathrm{VN}}(\lambda, \lambda/r), \\ d_{\mathrm{SQ}}(\lambda, r\lambda) &= d_{\mathrm{SQ}}(\lambda, \lambda/r), \\ d_{\mathrm{KL}}(\lambda, r\lambda) &< d_{\mathrm{KL}}(\lambda, \lambda/r). \end{aligned} \tag{6}$$

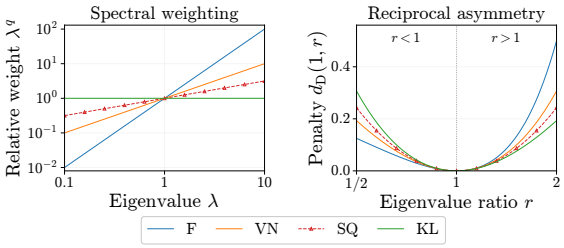


Figure 3: Left: Normalized penalties $\frac{d_{\mathrm{D}}(\lambda, r\lambda)}{d_{\mathrm{D}}(1,r)} = \lambda^{q_{\mathrm{D}}}$ for $r \neq 1$. Smaller q places less emphasis on large eigenvalues; KL is scale-independent. Right: $d_{\mathrm{D}}(1, r)$ versus r on a log scale. F and VN favor underestimation over reciprocal overestimation, KL shows the opposite preference, and SQ is symmetric.

Table 1 summarizes these properties together with the finite-sample comparison. Appendix B.3.4 gives the Kronecker form of the decomposition and shows that it retains these penalty scaling rules.

Takeaway. Smaller q_{D} gives more preference to the tail spectrum and penalizes underestimation more strongly. Since small fitted eigenvalues lead to large inverse-root preconditioning gains, this can help avoid over-amplification and motivates smaller q_{D} in preconditioner design.

3.2 Finite-Sample Stability under Limited Gradient Samples

In practical training, preconditioners are estimated from short histories of stochastic gradients. We therefore study how divergence choice affects finite-sample estimation error.

Model and isolated empirical factor estimation. Even if the gradient second moment is exactly Kronecker, its factors must be estimated from finite gradient samples, introducing noise. We ask how divergence choice affects factor estimation error. To isolate this effect from limited capacity, we consider an exact Kronecker Gaussian model $C = L_\star \otimes R_\star$:

$$G = L_\star^{1/2} Z R_\star^{1/2}, \quad Z_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1). \tag{7}$$

Fixing $R = R_\star$ and replacing the expectation in Equation (3) with the sample expectation over N independent gradients $\{G_b = L_\star^{1/2} Z_b R_\star^{1/2}\}_{b=1}^N$ gives the empirical estimate

$$\widehat{L}_{\mathrm{D}} = \frac{1}{N \operatorname{Tr}(R_\star^{q_{\mathrm{D}}})} \sum_{b=1}^{N} G_b R_\star^{q_{\mathrm{D}}-1} G_b^\top = L_\star^{1/2} \left[\frac{1}{N \operatorname{Tr}(R_\star^{q_{\mathrm{D}}})} \sum_{b=1}^{N} Z_b R_\star^{q_{\mathrm{D}}} Z_b^\top \right] L_\star^{1/2}. \tag{8}$$

Analogously, we define the right-side counterpart $\widehat{R}_{\mathrm{D}}$ of $\widehat{L}_{\mathrm{D}}$ to be the empirically estimated right factor on N samples when fixing the left factor to be $L_\star$.

Proposition 3 (Finite-Sample Factor Stability). *Define the effective rank* $r_{\mathrm{eff}}(A) := {\operatorname{Tr}(A)^2}/{\operatorname{Tr}(A^2)}$ *and the relative factors* $\widetilde{L}_{\mathrm{D}} := L_\star^{-1/2} \widehat{L}_{\mathrm{D}} L_\star^{-1/2}$ *and* $\widetilde{R}_{\mathrm{D}} := R_\star^{-1/2} \widehat{R}_{\mathrm{D}} R_\star^{-1/2}$, *which in the case of perfect empirical estimation simplify to the identity matrix. Under Equation (7), the estimator in Equation (8) and its right-factor counterpart, with the opposite factor fixed at its population value, satisfy*

$$\mathrm{Err}_{\mathrm{D}}^L := \frac{1}{m} \mathbb{E} \|\widetilde{L}_{\mathrm{D}} - I\|_F^2 = \frac{m+1}{N\, r_{\mathrm{eff}}(R_\star^{q_{\mathrm{D}}})}, \quad \mathrm{Err}_{\mathrm{D}}^R := \frac{1}{n} \mathbb{E} \|\widetilde{R}_{\mathrm{D}} - I\|_F^2 = \frac{n+1}{N\, r_{\mathrm{eff}}(L_\star^{q_{\mathrm{D}}})}.$$

Since $r_{\mathrm{eff}}(A^q)$ *is decreasing in* $q \geq 0$, *as shown in Figure 4, left, these errors satisfy*

$$\mathrm{Err}_{\mathrm{KL}}^s \leq \mathrm{Err}_{\mathrm{SQ}}^s \leq \mathrm{Err}_{\mathrm{VN}}^s \leq \mathrm{Err}_{\mathrm{F}}^s, \quad s \in \{L, R\}.$$

with equality when the opposite factor has a flat spectrum. See proof in Appendix B.4.

Why Frobenius fitting has larger estimation error. For a long-tailed spectrum of $R_\star$, F places much of the weight on a few large eigenvalues. Sampling fluctuations in these directions can have a larger effect on the estimate of L. This is reflected in the smaller effective rank of $R_\star^2$. KL instead weights all directions equally, making its estimate less sensitive to sampling noise.

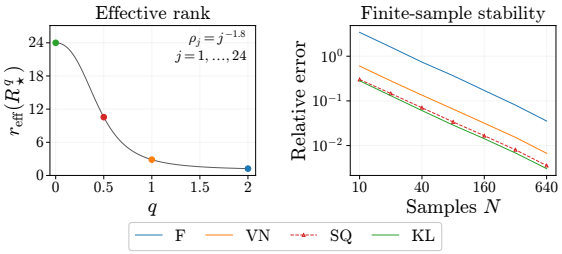


Figure 4: Left: Effective rank for $R_\star$ with eigenvalues $\rho_j = j^{-1.8}$. Right: Joint finite-sample fitting error on non-Kronecker gradients over 20 repeats, relative to each method's 4096-sample fit. Both results favor smaller q_{D}, with KL most stable and F least stable.

Limitations and empirical evidence. Joint factor estimation under general gradients is difficult to analyze directly. To isolate how divergence choice affects sampling error, we use an exact Kronecker Gaussian model with the opposite factor fixed at its population value. We also test the ordering without these assumptions by jointly fitting both factors on synthetic long-tailed non-Kronecker gradients. We extend the relative factor error in Proposition 3 to the jointly fitted $S = L \otimes R$, measured against each method's large-sample fit. The mean errors follow the theoretical ordering at every tested sample size (Figure 4, right).

Takeaway. Under the exact Kronecker Gaussian model with one factor fixed, smaller q_{D} gives lower expected relative factor error, with KL achieving the lowest error. This favors a lower q_{D} when designing preconditioners that must be estimated reliably from limited gradient samples.

4 Reliability of Empirical Gradient Second Moments

We now turn to Q2: *how reliable is the empirical second moment being approximated?* Using GPT-2 gradients, we estimate eigenvectors on one data split and measure the second moment along them on independent splits, giving *held-out scales*. We compare source eigenvalues with these held-out scales (Figure 1, left) to measure reliability and sampling error before any Kronecker approximation.

Shared experimental design (full details in Appendix E). Sections 4 and 5 use the same checkpoints and gradient data. First, we train a model (GPT-2 (Jordan et al., 2024) on FineWeb, 12 layers, hidden size 128) using AdamW and collect a large number of mini-batch gradients (batch size 64, sequence length 512), at selected checkpoints (100, 2500, and 4500) during training. We then split those gradients into four disjoint splits (each of size $N = 20{,}480$). Each split gives rise to an empirical gradient second moment $C_s = {}^1\!/_N \sum_{n=1}^N g_{s,n} g_{s,n}^\top$ with $g_{s,n}$ the nth mini-batch gradient in split s. In what follows, we focus on the query, value, and output projection weights of the model's 5th block which satisfy $N > d = 128^2$, thereby avoiding rank deficiency due to under-sampling. We repeated the analyses on checkpoints generated by training with Shampoo and find similar results (Appendix E.6). Note that this large-sample setting provides a reference for subsequent preconditioning comparisons, not a practical scheme for constructing preconditioners during training.

Are empirical second-moment eigenvalues reliable? Following high-dimensional covariance estimation work (Ledoit & Wolf, 2012), we estimate eigenvectors on a source split s and test their directional scales on a held-out split $s' \neq s$. Let $\{(\lambda_{s,i}, u_{s,i})\}_{i=1}^d$ denote the eigenpairs of the empirical moment C_s on split s. To assess the representativeness of its eigenvalues, we fix its eigenbasis and compare the held-out moments through their *directional scales* $u_{s,i}^\top \boldsymbol{C}_{s'} u_{s,i}$ along these directions. Should $C_{s'}$ share that eigenpair, this directional scale equals the eigenvalue $\lambda_{s,i}$. We measure *held-out scale error* by comparing held-out and source scales:

$$\xi_{s \to s',i} := \log_{10} \frac{u_{s,i}^\top C_{s'} u_{s,i}}{u_{s,i}^\top C_s u_{s,i}} = \log_{10} \frac{u_{s,i}^\top C_{s'} u_{s,i}}{\lambda_{s,i}}. \tag{9}$$

Interpretation. Zero indicates cross-split agreement; positive values indicate underestimation and negative values indicate overestimation. Larger absolute values indicate greater mismatch.

Remark 4. *We focus on eigenvalue (scale) reliability, without eigenvector stability. These directions need not remain eigenvectors across splits. What matters is whether the scales along these directions reproduce on independent data, supporting the scaling applied by the preconditioner is stable.*

Empirical eigenvalues become less reliable toward the tail. In Figure 1 (left), using the AdamW step-2500 output projection, held-out scale error is near zero for large eigenvalues and rises toward smaller eigenvalues, showing increasingly severe underestimation. Held-out scale error remains near zero in leading directions and strongly positive in the tail across checkpoints and projections from both AdamW and Distributed Shampoo (Table 10 and Appendix E.2). We also use directions from a third split s'' for a cross-fit check, which supports sampling noise as the source of the scale differences (See details in Appendix E.1.1).

Held-out scale as the correction. Covariance estimation and mini-batch curvature debiasing consider a correction based on separate data: keep the estimated eigenvectors, but replace their sample eigenvalues with scales measured along those directions on an independent batch (Ledoit & Wolf, 2012; Tatzel et al., 2025). Here, the replacement for $\lambda_{s,i}$ is precisely the held-out scale $u_{s,i}^\top C_{s'} u_{s,i}$ in Equation (9). This motivates using it as the reference for preconditioner gains in Section 5.2.

Takeaway. Even with more gradients than dimensions, leading eigenvalues agree across data splits, while tail eigenvalues are underestimated. Held-out scales measured along the same directions provide corrected scale estimates. Closely fitting the underestimated source eigenvalues can lead to excessive amplification under practical inverse-root preconditioning.

5 From Kronecker Fits to Preconditioning

We now address Q3: *how do divergence properties and second-moment reliability jointly affect preconditioning?* We examine how divergence-dependent approximation and second-moment bias jointly shape inverse-root gains, using the properties from Section 3.1 to explain the differences. We then test their recovery from short gradient histories and show on a stochastic quadratic that the resulting gain differences above can directly affect training loss (Appendix D.3).

5.1 Large-Sample Kronecker Fits

We fit each Kronecker state S_D^{ref} to the empirical second moment C_0, using all gradients in split 0. Figure 5 compares fitted eigenvalues and leading-subspace overlap.

Eigenvalues. We compare empirical and fitted eigenvalues in decreasing order. F underestimates much of the empirical spectrum, while VN follows it; SQ and KL have larger fitted tail eigenvalues. See Figure 13 for more spectrum figures.

Eigenspaces. Let U_k and $V_{D,k}$ contain the leading k eigenvectors of C_0 and S_D^{ref}. Their overlap is $O_D(k) := \|U_k^\top V_{D,k}\|_F^2/k$: One means identical subspaces. F best matches the leading eigenvector but has lower overlap for most larger subspaces.

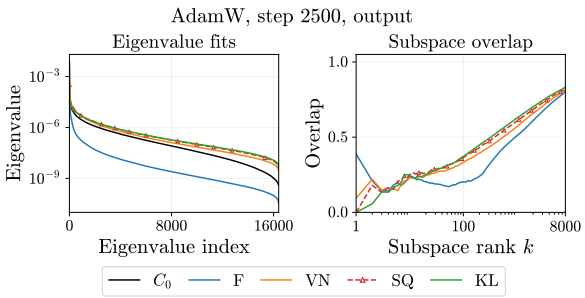


Figure 5: Large-sample Kronecker fits. Left: empirical and fitted eigenvalues. Right: overlap of the leading subspaces. Results align with Section 3.1

Takeaway. The Kronecker fits reflect the spectral preferences predicted in Section 3.1.

5.2 Inverse Root Actions of Large-Sample Fits

We now examine how the Kronecker-fitted second moments act as preconditioners. For each divergence D, the fitted matrix S_D gives the preconditioner $S_D^{-1/2}$. For each source empirical second-moment eigenvector, we compare the amplification induced by the preconditioner with the inverse square root of the corresponding held-out scale. This tests whether the Kronecker preconditioner corrects the scaling errors in the empirical second moment.

Let $(\lambda_{0,i}, u_{0,i})$ be the eigenpairs of the source second moment C_0. We average the three held-out directional scales, $h_i := {}^1\!/_3 \sum_{s'=1}^3 u_{0,i}^\top C_{s'} u_{0,i}$, and use $h_i^{-1/2}$ as the reference gain along $u_{0,i}$. For

a fitted state S_D, we measure the remaining discrepancy by the **held-out gain error**:

$$e_{D,i} := \log_{10} \frac{\|S_D^{-1/2} u_{0,i}\|_2}{h_i^{-1/2}} = \frac{1}{2} \log_{10} \left(h_i u_{0,i}^\top S_D^{-1} u_{0,i} \right). \tag{10}$$

Interpretation. Zero means matching the scaling factor $1/\sqrt{h_i}$ estimated from held-out data; positive and negative values mean larger and smaller factors, respectively. **Our goal is to keep this error close to zero across the whole spectrum.**

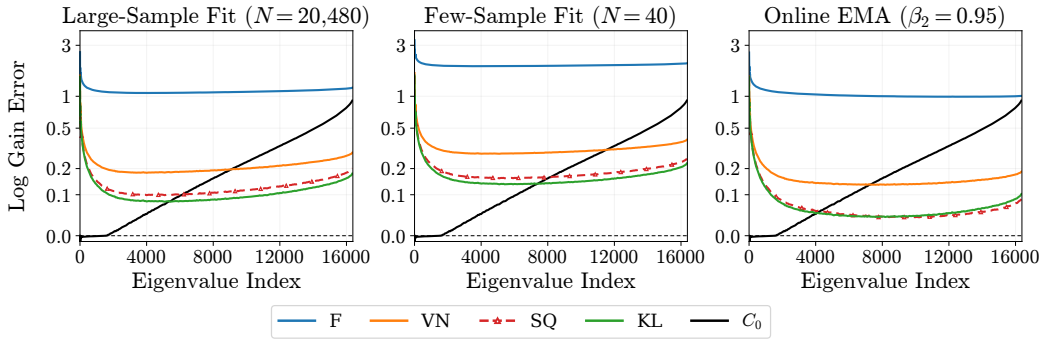


Figure 6: Smoothed log held-out gain error for the output projection at AdamW step 2500. Closer to zero is better; with asinh vertical scale. Left: large-sample fit; middle: 40-sample fits (4 repeats); right: EMA (16 repeats). Black solid: large-sample $C_0^{-1/2}$ in all panels; dashed: the held-out target.

Divergence choice shapes inverse-root gains. KL and SQ better compensate for empirical second-moment error and stay closer to the held-out target across the spectrum, with KL achieving the lowest aggregate error (Figure 6 and Table 2).

Table 2: Absolute held-out gain error (↓). KL has the lowest mean error, followed by SQ. Each experiment contributes the median $|e_{D,i}|$ over all retained directions and repeats; entries show mean ± sample SD across nine checkpoint–projection pairs per training optimizer.

Optimizer	Method	Large-sample	40-sample	EMA
AdamW	F	0.9759 ± 0.4623	2.4945 ± 0.8265	0.7645 ± 0.3453
	VN	0.3328 ± 0.1769	0.5549 ± 0.2986	0.2897 ± 0.1473
	SQ	0.1467 ± 0.0573	0.2175 ± 0.0893	0.1294 ± 0.0412
	KL	**0.1024** ± 0.0182	**0.1561** ± 0.0297	**0.1071** ± 0.0203
Distributed Shampoo	F	0.8801 ± 0.3234	2.6922 ± 0.5674	0.7008 ± 0.2267
	VN	0.3353 ± 0.1109	0.6207 ± 0.1966	0.2806 ± 0.0935
	SQ	0.1273 ± 0.0285	0.1876 ± 0.0407	0.1257 ± 0.0300
	KL	**0.1003** ± 0.0177	**0.1482** ± 0.0249	**0.1072** ± 0.0204

Link to divergence properties The behavior above follows the spectral preferences in Section 3.1. F and VN emphasize the leading spectrum and tend to underestimate the rest eigenvalues, whereas KL weights errors uniformly across the spectrum and tends to overestimate; SQ lies in between. This yields larger eigenvalues for SQ and KL, which prevents amplifying gradients too much and brings the gains closer to the held-out targets. Appendix E.4 gives further details.

Takeaway. KL has the lowest aggregate held-out gain error, followed by SQ, VN, and F.

5.3 Few-Sample Fits and Online EMA

The large-sample behavior is useful in practice only if it can be well recovered from short gradient histories. We now examine how few-sample fitting and online EMA affect preconditioner gains using two metrics. Recovery error measures how closely each method reproduces its own large-sample gains, as well as the held-out gain error from Section 5.2.

Configuration. All four methods follow the EMA scheme of Lin et al. (2026), detailed in Appendix C.5, with $\beta_2 = 0.95$ and eigendecomposition every five steps. Equal-weight fits use $N = 40$ gradients, close to the nominal EMA effective sample size $(1 + \beta_2)/(1 - \beta_2) = 39$. Each EMA repeat uses 512 gradients to reduce initialization effects. Table 9 summarizes these settings.

Recovery error. For this metric, the reference is each method's large-sample fit S_D^{ref}. Let $\widehat{S}_D$ be the few-sample or EMA fit, and let $\{v_{D,j}, \omega_{D,j}\}_{j=1}^d$ be the eigenpairs of S_D^{ref}. We compare **inverse-root amplification** along the reference directions $v_{D,j}$:

$$\gamma_{D,j}^{\text{stab}} := \log_{10} \frac{\|\widehat{S}_D^{-1/2} v_{D,j}\|_2}{\|(S_D^{\text{ref}})^{-1/2} v_{D,j}\|_2} = \frac{1}{2} \log_{10}(\omega_{D,j} v_{D,j}^\top \widehat{S}_D^{-1} v_{D,j}). \tag{11}$$

Interpretation. The gain mismatch $\gamma_{D,j}^{\text{stab}}$ is zero when the large-sample scaling is recovered; positive and negative values indicate stronger and weaker amplification, respectively.

Table 3: KL most accurately recovers its large-sample scaling, followed by SQ. Recovery error is the median $|\gamma_D^{\text{stab}}|$ over reference directions and repeats; smaller values indicate closer recovery of each method's own inverse-root gains (Equation (11); ↓). Entries are mean ± sample SD across nine checkpoint–projection pairs per training optimizer.

	AdamW		Distributed Shampoo	
Divergence	**Few-sample**	**EMA**	**Few-sample**	**EMA**
F	1.668 ± 0.601	0.173 ± 0.070	1.967 ± 0.430	0.154 ± 0.017
VN	0.259 ± 0.141	0.082 ± 0.032	0.339 ± 0.121	0.087 ± 0.017
SQ	0.089 ± 0.042	0.063 ± 0.020	0.077 ± 0.021	0.048 ± 0.008
KL	**0.066** ± 0.020	**0.044** ± 0.011	**0.060** ± 0.015	**0.040** ± 0.009

Recovery results. KL has the lowest recovery error, followed by SQ, VN, and F, for both 40-sample fits and EMA on both checkpoint sources (Table 3). This ordering agrees with the finite-sample analysis in Section 3.2: lower trace-power exponents give less weight to large sample fluctuations and reduce factor estimation error. Additional results are shown in Appendix E.5.

Held-out gain results. The same ordering holds for aggregate held-out gain error (Table 2). F's overamplification grows sharply with 40-sample fitting; EMA reduces it, but F remains furthest from the held-out target in the main example (Figure 6, middle and right).

Effect on optimization. On a stochastic quadratic model, we further verify that these gain differences can affect optimization; see Appendix D.3.

Takeaway. SQ and KL better compensate for underestimated empirical eigenvalues, yielding more accurate gains and more reliable recovery from short gradient histories. Together, these results show how divergence choice and second-moment reliability jointly shape preconditioning.

6 GPT-2 Training

We compare F-, VN-, SQ-, and KL-Shampoo by their final validation losses after training 12-layer GPT-2 models on FineWeb at several widths. Across widths 256–1536, all methods use $5,100$ updates and spectral EMA ($\beta_2 = 0.95$). Appendix F gives the full settings and training curves. Figure 1 (right) and Table 4 show lower validation losses for SQ and KL than for F and VN at every width. SQ and KL are close at smaller widths, with KL ahead at the two largest widths.

Method / width	**256**	**512**	**768**	**1024**	**1536**
F	3.8281	3.5082	3.3993	3.3528	3.3111
VN	3.7573	3.4377	3.2876	3.1979	3.0972
SQ	**3.7450**	**3.4173**	3.2661	3.1824	3.0817
KL	3.7475	3.4180	**3.2660**	**3.1769**	**3.0758**

Table 4: Final GPT-2 validation loss on FineWeb after $5,100$ updates (↓); one run per configuration. SQ and KL outperform F and VN at every width. Bold marks the lowest loss in each column.

Prediction. We expect preconditioners whose held-out gain errors remain close to zero across the spectrum, even with short gradient histories, to perform better. Held-out gain error in Equation (10) thus provides a criterion for guiding future optimizer design and evaluation.

7 Conclusion

We introduced a trace-power Bregman family that unifies divergence-based Shampoo variants and connects SQ-Shampoo to PSGD. Divergence choice shapes spectral fitting and finite-sample stability, while empirical second moments become increasingly biased in the tail. SQ and KL better compensate for this bias, yielding more accurate inverse-root gains and lower GPT-2 validation losses. These results further provide insights for evaluating new optimizer designs.

AI USE DISCLOSURE

Generative AI tools were used only to improve the clarity, grammar, and readability of the manuscript. They were not used to develop the theoretical framework, formulate mathematical claims, design experiments, interpret results, or produce the scientific conclusions of this work. All AI-assisted edits were reviewed and revised by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Naman Agarwal, Rohan Anil, Elad Hazan, Tomer Koren, and Cyril Zhang. Disentangling adaptive gradient methods from learning rates. *arXiv preprint arXiv:2002.11803*, 2020.

Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

Shun-ichi Amari. Information geometry of positive measures and positive-definite matrices: Decomposable dually flat structure. *Entropy*, 16(4):2131–2145, 2014.

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.

Rohan Anil, Vineet Gupta, Tomer Koren, Kevin Regan, and Yoram Singer. Scalable second order optimization for deep learning. *arXiv preprint arXiv:2002.09018*, 2020.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Daniel Bartz. Cross-validation based nonlinear shrinkage. *arXiv preprint arXiv:1611.00798*, 2016.

Andrzej Cichocki and Shun-ichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

George E Dahl, Frank Schneider, Zachary Nado, Naman Agarwal, Chandramouli Shama Sastry, Philipp Hennig, Sourabh Medapati, Runa Eschenhagen, Priya Kasimbeg, Daniel Suo, et al. Benchmarking neural network training algorithms. *arXiv preprint arXiv:2306.07179*, 2023.

Inderjit S Dhillon and Joel A Tropp. Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2008.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

Pierre Dutilleul. The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, 64(2):105–123, 1999.

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

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.

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.

Kevin Frans, Sergey Levine, and Pieter Abbeel. A stable whitening optimizer for efficient neural network training. *Advances in Neural Information Processing Systems*, 38:174086–174110, 2025.

Kaixin Gao, Xiaolei Liu, Zhenghai Huang, Min Wang, Zidong Wang, Dachuan Xu, and Fan Yu. A trace-restricted kronecker-factored approximation to natural gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7519–7527, 2021.

Roger Grosse and James Martens. A kronecker-factored approximate fisher matrix for convolution layers. In *International Conference on Machine Learning*, pp. 573–582. PMLR, 2016.

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

Tom Heskes. On "natural" learning and pruning in multilayered perceptrons. *Neural Computation*, 12(4):881–901, 2000.

Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024. URL https://github.com/KellerJordan/modded-nanogpt.

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.

Juno Kim, Eshaan Nichani, Denny Wu, Alberto Bietti, and Jason D Lee. Sharp capacity scaling of spectral optimizers in learning associative memory. *arXiv preprint arXiv:2603.26554*, 2026.

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

Brian Kulis, Mátyás A Sustik, and Inderjit S Dhillon. Low-rank kernel learning with bregman matrix divergences. *Journal of Machine Learning Research*, 10(13):341–376, 2009.

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

Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *The Annals of Statistics*, 40(2):1024–1060, 2012. doi: 10.1214/12-AOS989.

Xi-Lin Li. Preconditioned stochastic gradient descent. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1454–1466, 2018.

Xi-Lin Li. Preconditioner on matrix lie group for sgd. In *International Conference on Learning Representations*, 2019.

Wu Lin, Felix Dangel, Runa Eschenhagen, Kirill Neklyudov, Agustinus Kristiadi, Richard E Turner, and Alireza Makhzani. Structured inverse-free natural gradient descent: Memory-efficient & numerically-stable kfac. In *Forty-first International Conference on Machine Learning*, 2024.

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.

Bing Liu, Wenjie Zhou, and Chengcheng Zhao. Rethinking Bregman divergences in Kronecker-factored optimizers. In *ICML Workshop on High-dimensional Learning Dynamics*, 2026.

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

James Martens, Jimmy Ba, and Matt Johnson. Kronecker-factored curvature approximations for recurrent neural networks. In *International Conference on Learning Representations*, 2018.

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.

Yi Ren and Donald Goldfarb. Tensor normal training for deep learning models. *Advances in Neural Information Processing Systems*, 34:26040–26052, 2021.

Hao-Jun Michael Shi, Tsung-Hsien Lee, Shintaro Iwasaki, Jose Gallego-Posada, Zhijing Li, Kaushik Rangadurai, Dheevatsa Mudigere, and Michael Rabbat. A distributed data-parallel pytorch implementation of the distributed shampoo optimizer for training neural networks at-scale. *arXiv preprint arXiv:2309.06497*, 2023.

Lukas Nicola Tatzel, Bálint Mucsányi, Osane Hackel, and Philipp Hennig. Debiasing mini-batch quadratics for applications in deep learning. In *International Conference on Learning Representations*, volume 2025, pp. 41298–41337, 2025.

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*, pp. 3503–3513. PMLR, 2020.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Tan. Muon outperforms adam in tail-end associative memory learning. In *International Conference on Learning Representations*, volume 2026, pp. 37382–37419, 2026.

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.

Ashley Zhang, Ben Keigwin, Dhruv Pai, and Alec Dewulf. Online kl shampoo. 2026. URL https://tilderesearch.com/blog/online-kl-shampoo.

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*, pp. 7654–7663. PMLR, 2019.

APPENDIX CONTENTS

A Related Work **14**

B Proofs **15**
B.1 Proof of Proposition 1 . . . 15
B.2 Limited Kronecker Capacity . . . 17
B.3 Spectral Error Decomposition and Kronecker Approximations . . . 17
B.3.1 Proof of Proposition 2 . . . 17
B.3.2 Divergence-Specific Spectral Penalties . . . 18
B.3.3 Reciprocal Asymmetry . . . 18
B.3.4 Kronecker Approximations . . . 19
B.4 Proof of Proposition 3 . . . 19
B.5 Local Geometry of SQ and KL . . . 20

C Preconditioner Update Details **21**
C.1 Original Shampoo . . . 21
C.2 PSGD-Kron, SQ-Shampoo, and KL-Shampoo . . . 21
C.3 Divergence-Based Factor Statistics . . . 21
C.4 Offline Fixed-Point Fits . . . 22
C.5 Online EMA Updates . . . 22

D Synthetic Validation **22**
D.1 Experimental Setup . . . 22
D.2 Metrics and Results . . . 23
D.3 Directional Gain Interventions on a Quadratic . . . 24

E Real-Gradient Experimental Details **25**
E.1 Empirical Second Moments and Finite-Sample Effects . . . 25
E.1.1 Finite-Sample Tail Selection . . . 25
E.1.2 High-Dimensional Sampling and Tail Signals . . . 26
E.2 Reliability Across Checkpoints . . . 26
E.3 Inverse-Root Gain Across Checkpoints . . . 27
E.4 Kronecker Fits of Different Divergences . . . 29
E.5 Gain Recovery with Short Gradient Histories . . . 30
E.6 Diagnostics on Distributed Shampoo Checkpoints . . . 32

F GPT-2 Training Details and Full Results **33**

A Related Work

Natural gradients and structured curvature. Natural gradient descent uses the Fisher metric to account for parameter-space geometry (Amari, 1998). Layered Fisher approximations (Heskes, 2000) and K-FAC (Martens & Grosse, 2015) make this approach tractable by exploiting network structure. Subsequent work extends Kronecker factorizations to convolutional layers (Grosse & Martens, 2016), recurrent networks (Martens et al., 2018), and general linear layers with weight sharing (Eschenhagen et al., 2023). Tensor Normal Training uses a tensor-normal model of sampled gradients to approximate the Fisher (Ren & Goldfarb, 2021), TKFAC imposes trace constraints (Gao et al., 2021), and SINGD incorporates inverse-free updates to avoid explicit matrix inversion (Lin et al., 2024).

Shampoo and Kronecker approximation. AdaGrad preconditions gradients using their accumulated second moments (Duchi et al., 2011), while Adam uses an exponential moving average of squared gradients for coordinate-wise scaling (Kingma & Ba, 2015). Shampoo exploits tensor structure to construct a Kronecker-factored bound on full-matrix AdaGrad (Gupta et al., 2018). Distributed computation, infrequent inverse-root updates, and learning-rate grafting make Shampoo practical at scale (Anil et al., 2020; Agarwal et al., 2020; Shi et al., 2023). SOAP applies Adam in Shampoo's eigenbasis to improve training performance and stability (Vyas et al., 2025). Morwani et al. (2025) show that the Kronecker product of Shampoo's statistics corresponds to one power-iteration step from identity initialization toward the Frobenius-optimal Kronecker approximation. In their experiments, its cosine similarity to the target closely tracks that of the optimal approximation. They also examine how batch gradients and label sampling affect the estimation.

Matrix divergences and Kronecker estimation. Bregman matrix divergences provide alternatives to Frobenius approximation (Dhillon & Tropp, 2008), with applications to positive-semidefinite kernel learning (Kulis et al., 2009). For Kronecker approximation, the LogDet (KL divergence between two zero-mean multivariate Gaussian distributions) objective corresponds to matrix-normal maximum likelihood, which estimates row and column covariance factors alternately (Dutilleul, 1999). Lin et al. (2026) formulate gradient second-moment estimation through matrix divergences, yielding F-, VN-, and KL-Shampoo. We study how these divergence choices affect spectral error allocation and finite-sample factor stability within a unified trace-power family.

Reliability of empirical second moments. Empirical gradient second moments can differ substantially from curvature matrices like the Fisher and Hessian (Kunstner et al., 2019; Thomas et al., 2020). Furthermore, with finite samples, empirical eigenvalues systematically misestimate population scales along their eigenvectors. Prior work addresses this bias through nonlinear covariance shrinkage or cross-validation (Ledoit & Wolf, 2012; Bartz, 2016) in high-dimensional covariance estimation. A similar selection bias occurs in deep-learning Hessian and GGN estimates: directions with extreme curvature on one mini-batch have less extreme scales on others (Tatzel et al., 2025). We examine these estimation effects in mini-batch gradient second moments and their consequences for divergence-dependent Kronecker preconditioning.

B PROOFS

Throughout the theoretical analysis, following Eschenhagen et al. (2025); Lin et al. (2026), we assume the gradient second moment C to be positive definite.

Matrix notation and divergence formulas. For $M = U \operatorname{diag}(\mu_1, \dots, \mu_d) U^\top \in \mathbb{S}^d_{++}$, matrix functions act on its eigenvalues:

$$f(M) = U \operatorname{diag}(f(\mu_1), \dots, f(\mu_d)) U^\top, \quad \Phi(M) = \sum_{i=1}^{d} f(\mu_i), \quad \nabla\Phi(M) = f'(M).$$

We use $\|A\|_F^2 = \langle A, A \rangle_F$ and write $\nabla^2\Phi(S)[H]$ for the Hessian on a perturbation H. Substituting the generators $f_{\mathrm{F}}(x) = x^2/2$, $f_{\mathrm{VN}}(x) = x \log x - x$, and $f_{\mathrm{KL}}(x) = -\log x$ into Equation (2) gives

$$\mathcal{B}_F(X, Y) = \frac{1}{2}\|X - Y\|_F^2, \quad \mathcal{B}_{\mathrm{VN}}(X, Y) = \operatorname{Tr}(X \log X - X \log Y - X + Y),$$
$$\mathcal{B}_{\mathrm{KL}}(X, Y) = \operatorname{Tr}(Y^{-1}X) - \log\det(Y^{-1}X) - d.$$

In particular, $\mathcal{B}_{\mathrm{KL}}(X, Y) = 2D_{\mathrm{KL}}(\mathcal{N}(0, X) \| \mathcal{N}(0, Y))$. For row-wise vectorization, $\operatorname{vec}(AGB) = (A \otimes B^\top) \operatorname{vec}(G)$.

B.1 PROOF OF PROPOSITION 1

For $q = 2$, $f_2(x) = (x-1)^2/2$ and $f_2'(x) = x - 1$. Substitution into Equation (2) gives

$$\mathcal{B}_{\Phi_2}(X, Y) = \frac{1}{2}\operatorname{Tr}(X^2 - Y^2) - \operatorname{Tr}(Y(X - Y)) = \frac{1}{2}\|X - Y\|_F^2 = \mathcal{B}_F(X, Y).$$

L'Hôpital's rule gives $f_q(x) \to x \log x - x + 1$ as $q \to 1$ and $f_q(x) \to -\log x + x - 1$ as $q \to 0$. The derivative $f_q'(x) = (x^{q-1} - 1)/(q-1)$ converges to $\log x$ and $1 - x^{-1}$, respectively. For fixed $X, Y \in \mathbb{S}^d_{++}$, we can obtain

$$\lim_{q \to 1} \mathcal{B}_{\Phi_q}(X, Y) = \mathcal{B}_{\Phi_1}(X, Y) = \operatorname{Tr}(X \log X - X \log Y - X + Y) = \mathcal{B}_{\mathrm{VN}}(X, Y),$$
$$\lim_{q \to 0} \mathcal{B}_{\Phi_q}(X, Y) = \mathcal{B}_{\Phi_0}(X, Y) = \operatorname{Tr}(Y^{-1}X) - \log\det(Y^{-1}X) - d = \mathcal{B}_{\mathrm{KL}}(X, Y).$$

For $q = \frac{1}{2}$, which we call square root (SQ), $f_{1/2}(x) = -4\sqrt{x} + 2x + 2$ and $f_{1/2}'(x) = 2 - 2x^{-1/2}$. Substituting into Equation (2) yields

$$\frac{1}{2}\mathcal{B}_{\Phi_{1/2}}(C, S) = \operatorname{Tr}(S^{1/2}) + \operatorname{Tr}(CS^{-1/2}) - 2\operatorname{Tr}(C^{1/2}). \tag{12}$$

Under $P = S^{-1/2}$, this becomes

$$\frac{1}{2}\mathcal{B}_{\Phi_{1/2}}(C, S) = \operatorname{Tr}(PC) + \operatorname{Tr}(P^{-1}) - 2\operatorname{Tr}(C^{1/2}).$$

The first two terms are exactly the PSGD objective in Equation (1); the last term is independent of P. Thus SQ and PSGD have the same objective up to a positive scale and an additive constant. For $S = L \otimes R$, the change of variables gives $P = L^{-1/2} \otimes R^{-1/2}$, so the correspondence also holds under the Kronecker constraint.

We now derive the factor equation in Equation (3). For $q \neq 0, 1$,

$$\nabla\Phi_q(S) = \frac{S^{q-1} - I}{q - 1}.$$

By substituting this into the Bregman divergence and keeping only the terms that depend on S, we can derive

$$\mathcal{B}_{\Phi_q}(C, S) = \frac{1}{q}\operatorname{Tr}(S^q) - \frac{1}{q-1}\operatorname{Tr}(CS^{q-1}) + \text{const}.$$

For $S = L \otimes R$,

$$\operatorname{Tr}(S^q) = \operatorname{Tr}(L^q)\operatorname{Tr}(R^q),$$

and the definition of $C = \mathbb{E}[gg^\top]$ gives

$$\mathrm{Tr}(CS^{q-1}) = \mathbb{E}\left[\mathrm{Tr}\left(L^{q-1}GR^{q-1}G^\top\right)\right].$$

Fix R, and write

$$A_R := \mathbb{E}[GR^{q-1}G^\top], \quad a_R := \mathrm{Tr}(R^q).$$

Combining all the equations above, we can write the objective $\mathcal{B}_{\Phi_q}$ that depends on L as

$$J_q(L;R) = \frac{a_R}{q}\,\mathrm{Tr}(L^q) - \frac{1}{q-1}\,\mathrm{Tr}(A_R L^{q-1}).$$

For $q \neq 0, 1$, let $X = L^{q-1}$. Then $L^q = X^{q/q-1}$, and

$$\nabla_X J_q = \frac{1}{q-1}\left(a_R X^{1/q-1} - A_R\right) = \frac{1}{q-1}(a_R L - A_R).$$

Thus the stationary equation gives $L = A_R/a_R = \mathbb{E}[GR^{q-1}G^\top]/\,\mathrm{Tr}(R^q)$. The objective is strictly convex in X, so this stationary point is the unique minimizer.

For the VN defined by $q \to 1$, $\Phi_1(S) = \mathrm{Tr}(S\log S - S + I)$. The limit established above gives

$$\mathcal{B}_{\Phi_1}(C,S) = \mathrm{Tr}(S) - \mathrm{Tr}(C\log S) + \mathrm{const}.$$

Using

$$\mathrm{Tr}(L\otimes R) = \mathrm{Tr}(L)\,\mathrm{Tr}(R), \qquad \log(L\otimes R) = \log L \otimes I_n + I_m \otimes \log R,$$

and the row-wise vectorization identity

$$\mathrm{vec}(G)^\top(\log L \otimes I_n)\,\mathrm{vec}(G) = \mathrm{Tr}(GG^\top \log L),$$

the terms depending on L reduce to

$$J_1(L;R) = \mathrm{Tr}(R)\,\mathrm{Tr}(L) - \mathrm{Tr}(A_R \log L), \qquad A_R := \mathbb{E}[GG^\top].$$

Here R is fixed, so the term involving $I_m \otimes \log R$ is constant in L. Set $X = \log L$. Then

$$J_1(e^X;R) = \mathrm{Tr}(R)\,\mathrm{Tr}(e^X) - \mathrm{Tr}(A_R X),$$

whose gradient with respect to X is

$$\nabla_X J_1 = \mathrm{Tr}(R)e^X - A_R.$$

Setting this to zero gives $L = e^X = A_R/\,\mathrm{Tr}(R)$. Since $\mathrm{Tr}(e^X)$ is strictly convex on symmetric matrices, this is the unique minimizer.

For the KL defined by $q \to 0$, $\Phi_0(S) = -\log\det S + \mathrm{Tr}(S) - mn$. Its affine terms cancel in the divergence, giving

$$\mathcal{B}_{\Phi_0}(C,S) = \mathrm{Tr}(CS^{-1}) + \log\det S + \mathrm{const}.$$

Using $S^{-1} = L^{-1}\otimes R^{-1}$ and $\log\det(L\otimes R) = n\log\det L + m\log\det R$, the terms that depend on L are

$$J_0(L;R) = \mathrm{Tr}(A_R L^{-1}) + n\log\det L, \quad A_R := \mathbb{E}[GR^{-1}G^\top].$$

Setting $X = L^{-1}$, we can derive:

$$J_0(X^{-1};R) = \mathrm{Tr}(A_R X) - n\log\det X, \qquad \nabla_X J_0 = A_R - nX^{-1}.$$

Setting the gradient to zero gives $X^{-1} = A_R/n$, hence

$$L = \frac{A_R}{n} = \frac{\mathbb{E}[GR^{-1}G^\top]}{\mathrm{Tr}(R^0)}.$$

Since $-\log\det X$ is strictly convex, this is the unique minimizer.

The cases $q = 1$ and $q = 0$ give the same formula, so all cases can be written as

$$L_q(R) = \frac{\mathbb{E}[GR^{q-1}G^\top]}{\mathrm{Tr}(R^q)}.$$

Exchanging L and R gives

$$R_q(L) = \frac{\mathbb{E}[G^\top L^{q-1}G]}{\mathrm{Tr}(L^q)}.$$

B.2 Limited Kronecker Capacity

A single Kronecker product generally cannot represent a gradient second moment exactly. The factors L and R have $m(m+1)/2+n(n+1)/2-1$ free parameters after accounting for reciprocal rescaling, fewer than the $mn(mn+1)/2$ parameters of a general symmetric matrix when $m, n \geq 2$. Exact matching is possible for second moments with Kronecker structure; otherwise, the choice of divergence determines how approximation errors are allocated.

B.3 Spectral Error Decomposition and Kronecker Approximations

B.3.1 Proof of Proposition 2

Notation and spectral costs. Write $U = [u_1, \ldots, u_d]$, $V = [v_1, \ldots, v_d]$, $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$, and $\Omega = \operatorname{diag}(\omega_1, \ldots, \omega_d)$, with both spectra ordered decreasingly. The scalar divergence and cost on eigenspace misalignment are

$$d_f(\lambda, \omega) := f(\lambda) - f(\omega) - f'(\omega)(\lambda - \omega),$$
$$\Delta_f(\Lambda, \Omega; P) := -\sum_{i,j=1}^{d} \lambda_i f'(\omega_j)(P_{ij} - \delta_{ij}), \quad P_{ij} = \langle u_i, v_j \rangle^2,$$

where δ_{ij} is the Kronecker delta.

Spectral decomposition. By the definition of the Bregman matrix divergence,

$$\mathcal{B}_\Phi(A, B) = \Phi(A) - \Phi(B) - \operatorname{Tr}\left(\nabla\Phi(B)(A - B)\right). \tag{13}$$

Since $\Phi(M) = \operatorname{Tr}(f(M))$ is a spectral function, its gradient is given by $\nabla\Phi(M) = f'(M)$. Therefore, for $B = V\Omega V^\top$,

$$\nabla\Phi(B) = V f'(\Omega) V^\top.$$

Let $Q = U^\top V$. Then Q is orthogonal and $Q_{ij} = \langle u_i, v_j \rangle$, so that $P_{ij} = Q_{ij}^2$. Using the cyclic property of the trace,

$$\operatorname{Tr}\left(\nabla\Phi(B)A\right) = \operatorname{Tr}\left(V f'(\Omega) V^\top U \Lambda U^\top\right) = \operatorname{Tr}\left(f'(\Omega) Q^\top \Lambda Q\right) = \sum_{i=1}^{d}\sum_{j=1}^{d} \lambda_i f'(\omega_j) P_{ij}. \tag{14}$$

Similarly, using $V^\top V = I$,

$$\operatorname{Tr}\left(\nabla\Phi(B)B\right) = \operatorname{Tr}\left(V f'(\Omega) V^\top V \Omega V^\top\right) = \operatorname{Tr}\left(f'(\Omega) V^\top V \Omega V^\top V\right) = \sum_{j=1}^{d} \omega_j f'(\omega_j). \tag{15}$$

Substituting Equations (14) and (15) into Equation (13) gives

$$\mathcal{B}_\Phi(A, B) = \sum_{i=1}^{d} f(\lambda_i) - \sum_{j=1}^{d} f(\omega_j) + \sum_{j=1}^{d} \omega_j f'(\omega_j) - \sum_{i=1}^{d}\sum_{j=1}^{d} \lambda_i f'(\omega_j) P_{ij}.$$

Adding and subtracting $\sum_{i=1}^{d} \lambda_i f'(\omega_i)$ yields

$$\begin{aligned}
\mathcal{B}_\Phi(A, B) &= \sum_{i=1}^{d} \left[f(\lambda_i) - f(\omega_i) - f'(\omega_i)(\lambda_i - \omega_i)\right] + \sum_{i=1}^{d} \lambda_i f'(\omega_i) - \sum_{i=1}^{d}\sum_{j=1}^{d} \lambda_i f'(\omega_j) P_{ij} \\
&= \sum_{i=1}^{d} d_f(\lambda_i, \omega_i) + \sum_{i=1}^{d}\sum_{j=1}^{d} -\lambda_i f'(\omega_j)\left(P_{ij} - \delta_{ij}\right) \\
&= \sum_{i=1}^{d} d_f(\lambda_i, \omega_i) + \Delta_f(\Lambda, \Omega; P).
\end{aligned}$$

The first term measures eigenvalue mismatch for aligned eigenspaces; Δ_f is the cost of eigenspace misalignment.

Nonnegativity of $\Delta_f(\Lambda, \Omega; P)$. Since Q is orthogonal, $P_{ij} = Q_{ij}^2$ is doubly stochastic. The Birkhoff–von Neumann theorem therefore gives $P = \sum_k \alpha_k \Pi_k$, where Π_k are permutation matrices, $\alpha_k \geq 0$, and $\sum_k \alpha_k = 1$. Both λ_i and $f'(\omega_i)$ are decreasing, since the eigenvalues are ordered and f is convex. The rearrangement inequality says that pairing them in this order maximizes their product sum. Thus, for each permutation σ_k associated with Π_k,

$$\sum_i \lambda_i f'(\omega_{\sigma_k(i)}) \leq \sum_i \lambda_i f'(\omega_i).$$

Taking the weighted average gives

$$\sum_{i,j} \lambda_i f'(\omega_j) P_{ij} \leq \sum_i \lambda_i f'(\omega_i),$$

which proves $\Delta_f(\Lambda, \Omega; P) \geq 0$.

B.3.2 Divergence-Specific Spectral Penalties

We overlook the affine terms. For Frobenius, $f_{\mathrm{F}}(x) = \frac{1}{2}x^2$ and $f'_{\mathrm{F}}(x) = x$, which gives

$$d_{\mathrm{F}}(\lambda, \omega) = \frac{1}{2}(\lambda - \omega)^2.$$

For von Neumann, $f_{\mathrm{VN}}(x) = x \log x - x$ and $f'_{\mathrm{VN}}(x) = \log x$, and hence

$$d_{\mathrm{VN}}(\lambda, \omega) = \lambda \log \frac{\lambda}{\omega} - \lambda + \omega.$$

For the square-root member, $f_{\mathrm{SQ}}(x) = -4\sqrt{x}$ and $f'_{\mathrm{SQ}}(x) = -2x^{-1/2}$, giving

$$d_{\mathrm{SQ}}(\lambda, \omega) = 2\left(\sqrt{\omega} + \frac{\lambda}{\sqrt{\omega}} - 2\sqrt{\lambda}\right).$$

For LogDet, $f_{\mathrm{KL}}(x) = -\log x$ and $f'_{\mathrm{KL}}(x) = -1/x$, which yields

$$d_{\mathrm{KL}}(\lambda, \omega) = \log \frac{\omega}{\lambda} + \frac{\lambda}{\omega} - 1.$$

Setting $\omega = r\lambda$ gives Equation (5), which establishes the scalar scale dependence in Equation (4). By using the exponents q_{D} in Proposition 1, Frobenius and KL give

$$\Delta_{\mathrm{F}}(r\Lambda, r\Omega; P) = r^2 \Delta_{\mathrm{F}}(\Lambda, \Omega; P), \qquad \Delta_{\mathrm{KL}}(r\Lambda, r\Omega; P) = \Delta_{\mathrm{KL}}(\Lambda, \Omega; P).$$

For von Neumann, using $\sum_{j=1}^d (P_{ij} - \delta_{ij}) = 0$, we obtain

$$\begin{aligned}
\Delta_{\mathrm{VN}}(r\Lambda, r\Omega; P) &= -r \sum_{i=1}^d \sum_{j=1}^d \lambda_i \log(r\omega_j)\,(P_{ij} - \delta_{ij}) \\
&= -r \log r \sum_{i=1}^d \lambda_i \sum_{j=1}^d (P_{ij} - \delta_{ij}) + r\Delta_{\mathrm{VN}}(\Lambda, \Omega; P) \\
&= r\Delta_{\mathrm{VN}}(\Lambda, \Omega; P).
\end{aligned}$$

For SQ, $f'_{\mathrm{SQ}}(\omega) = -2\omega^{-1/2}$ gives $\Delta_{\mathrm{SQ}}(r\Lambda, r\Omega; P) = r^{1/2}\Delta_{\mathrm{SQ}}(\Lambda, \Omega; P)$ directly. Therefore, for $\mathrm{D} \in \{\mathrm{F}, \mathrm{VN}, \mathrm{SQ}, \mathrm{KL}\}$,

$$\Delta_{\mathrm{D}}(r\Lambda, r\Omega; P) = r^{q_{\mathrm{D}}} \Delta_{\mathrm{D}}(\Lambda, \Omega; P), \quad r > 0.$$

B.3.3 Reciprocal Asymmetry

Substituting r and r^{-1} into Equation (5) proves Equation (6). For F, $(r-1)^2 > (r^{-1}-1)^2$ when $r > 1$; for SQ, the two penalties are equal. For VN and KL, the differences are $\lambda h(r)$ and $-h(r)$, respectively, where $h(r) := r - r^{-1} - 2\log r$. Since $h(1) = 0$ and $h'(r) = (r-1)^2/r^2 > 0$ for $r > 1$, VN penalizes overestimation more, while KL penalizes underestimation more.

B.3.4 KRONECKER APPROXIMATIONS

Let $C \in \mathbb{S}^{mn}_{++}$ have eigenpairs (λ_i, u_i) ordered by decreasing eigenvalue, and let L and R have eigenpairs $(\ell_a, v_{L,a})$, $a = 1, \dots, m$, and $(\rho_b, v_{R,b})$, $b = 1, \dots, n$, respectively. Sort the mn eigenvalue products $\ell_a \rho_b$ in decreasing order. If the j-th product comes from the a-th eigenvalue of L and the b-th eigenvalue of R, write $a_j = a$ and $b_j = b$. The corresponding eigenpair of $S = L \otimes R$ is

$$\omega_j = \ell_{a_j} \rho_{b_j}, \qquad v_j = v_{L,a_j} \otimes v_{R,b_j}, \qquad j = 1, \dots, mn.$$

Set $\Lambda = \operatorname{diag}(\lambda_i)$, $\Omega = \operatorname{diag}(\omega_j)$, and $P_{ij} = |\langle u_i, v_j \rangle|^2$ as in Proposition 2. We can obtain:

$$\mathcal{B}_{\Phi_q}(C, L \otimes R) = \underbrace{\sum_{i=1}^{mn} d_q(\lambda_i, \ell_{a_i} \rho_{b_i})}_{\text{eigenvalue mismatch}} + \underbrace{\Delta_q(\Lambda, \Omega; P)}_{\text{eigenspace misalignment}}, \quad \Delta_q(\Lambda, \Omega; P) \geq 0.$$

For Kronecker approximations, both mismatch terms retain the scaling relations in Equation (4), and the scalar penalties still satisfy the reciprocal-error comparisons in Equation (6).

B.4 PROOF OF PROPOSITION 3

We only prove the result for the left factor; the right-factor result follows symmetrically. Let $A := R_\star^{q_D}$, $\widetilde{L}_D := L_\star^{-1/2} \widehat{L}_D L_\star^{-1/2}$. By Equation (8), we have

$$\widetilde{L}_D = \frac{1}{N} \sum_{b=1}^{N} M_b, \quad M_b := \frac{Z_b A Z_b^\top}{\operatorname{Tr}(A)}.$$

Let $z_i^\top$ denote the i-th row of Z_b, so $z_i \sim \mathcal{N}(0, I_n)$. Then

$$\mathbb{E}\left[(Z_b A Z_b^\top)_{ij}\right] = \mathbb{E}[z_i^\top A z_j] = \begin{cases} \operatorname{Tr}(A), & i = j, \\ 0, & i \neq j, \end{cases}$$

where we use $\mathbb{E}[z_i z_i^\top] = I$ and the independence of different rows. Hence, $\mathbb{E}[Z_b A Z_b^\top] = \operatorname{Tr}(A) I$ and $\mathbb{E}[M_b] = I$. Using $\widetilde{L}_D - I = {}^1\!/_N \sum_{b=1}^N (M_b - I)$, we obtain

$$\mathbb{E}\left[\left\|\widetilde{L}_D - I\right\|_F^2\right] = \frac{1}{N^2} \sum_{b,c=1}^{N} \mathbb{E}\left[\langle M_b - I, M_c - I \rangle_F\right].$$

For $b \neq c$, independence together with $\mathbb{E}[M_b - I] = 0$ gives

$$\mathbb{E}\left[\langle M_b - I, M_c - I \rangle_F\right] = 0.$$

Hence, only the N diagonal terms remain, yielding

$$\mathbb{E}\left[\left\|\widetilde{L}_D - I\right\|_F^2\right] = \frac{1}{N} \mathbb{E}\left[\|M_1 - I\|_F^2\right]. \tag{16}$$

Let $z_i^\top$ denote the i-th row of Z_1. For the diagonal entries, $(M_1)_{ii} = \frac{z_i^\top A z_i}{\operatorname{Tr}(A)}$. Since $z_i \sim \mathcal{N}(0, I_n)$ and A is symmetric, the Gaussian quadratic-form identity $\operatorname{Var}(z_i^\top A z_i) = 2 \operatorname{Tr}(A^2)$ gives

$$\mathbb{E}\left[\left((M_1)_{ii} - 1\right)^2\right] = \frac{2 \operatorname{Tr}(A^2)}{\operatorname{Tr}(A)^2}.$$

For $i \neq j$, using the independence of z_i and z_j and $\mathbb{E}[z_j z_j^\top] = I$,

$$\mathbb{E}\left[(z_i^\top A z_j)^2\right] = \mathbb{E}\left[z_i^\top A \, \mathbb{E}[z_j z_j^\top] \, A z_i\right] = \operatorname{Tr}(A^2),$$

and therefore

$$\mathbb{E}\left[(M_1)_{ij}^2\right] = \frac{\operatorname{Tr}(A^2)}{\operatorname{Tr}(A)^2}.$$

Summing over the m diagonal and $m(m-1)$ off-diagonal entries yields

$$\mathbb{E}\left[\|M_1 - I\|_F^2\right] = m(m+1) \frac{\operatorname{Tr}(A^2)}{\operatorname{Tr}(A)^2} = \frac{m(m+1)}{r_{\text{eff}}(A)}.$$

Combining this with Equation (16) and dividing by m gives

$$\mathbb{E}\left[\frac{1}{m}\left\|L_\star^{-1/2}\widehat{L}_{\mathrm{D}}L_\star^{-1/2}-I\right\|_F^2\right]=\frac{m+1}{N\,r_{\mathrm{eff}}(R_\star^{q_{\mathrm{D}}})}.$$

The right-factor identity follows by exchanging $L_\star, m$ with $R_\star, n$.
It remains to compare the four exponents q_{D}. For positive eigenvalues $\{\rho_i\}$, define $h(q) := \log\sum_i \rho_i^q$ and $p_i(q) := \rho_i^q/\sum_j \rho_j^q$. Then

$$h''(q)=\sum_i p_i(q)(\log\rho_i)^2-\left(\sum_i p_i(q)\log\rho_i\right)^2\geq 0,$$

so h' is non-decreasing. Noting that the log effective rank is $\log r_{\mathrm{eff}}(R_\star^q)=2h(q)-h(2q)$, we have

$$\frac{d}{dq}\log r_{\mathrm{eff}}(R_\star^q)=2h'(q)-2h'(2q)\leq 0,\quad q\geq 0.$$

Thus effective rank is decreasing in q_D. Evaluating at $q=q_{\mathrm{D}}$ gives

$$r_{\mathrm{eff}}(R_\star^2)\leq r_{\mathrm{eff}}(R_\star)\leq r_{\mathrm{eff}}(R_\star^{1/2})\leq r_{\mathrm{eff}}(I)=n.$$

Since the expected estimation error is proportional to the reciprocal of effective rank, we obtain

$$\mathrm{Err}_{\mathrm{KL}}\leq\mathrm{Err}_{\mathrm{SQ}}\leq\mathrm{Err}_{\mathrm{VN}}\leq\mathrm{Err}_{\mathrm{F}}\,.$$

The inequalities become equalities for an isotropic opposite-factor spectrum.

B.5 Local Geometry of SQ and KL

We compare the local penalties of SQ and KL through their Hessians. KL matches the Gaussian Fisher geometry up to a constant factor; SQ does not.
For $S\in\mathbb{S}_{++}^d$ and symmetric H, Taylor expansion gives

$$\mathcal{B}_{\Phi_{q_{\mathrm{D}}}}(S+\varepsilon H,S)=\frac{\varepsilon^2}{2}\left\langle H,\nabla^2\Phi_{q_{\mathrm{D}}}(S)[H]\right\rangle_F+o(\varepsilon^2).$$

Thus the Hessian determines the penalty for small perturbations. For eigenvalue perturbations, $f''_{q_{\mathrm{D}}}(s)=s^{q_{\mathrm{D}}-2}$, giving $s^{-3/2}$ for SQ and s^{-2} for KL.
To compare this geometry with Gaussian Fisher information, consider the auxiliary model $x\sim\mathcal{N}(0,S)$ with density function p_S. The directional score for a symmetric perturbation H is

$$s_H(x):=D_S\log p_S(x)[H]=\frac{1}{2}\left(x^\top S^{-1}HS^{-1}x-\mathrm{Tr}(S^{-1}H)\right).$$

Its second moment gives the Fisher quadratic form:

$$\mathbb{E}_{x\sim p_S}[s_H(x)^2]=\frac{1}{2}\mathrm{Tr}(S^{-1}HS^{-1}H)=\langle H,\mathcal{I}_S[H]\rangle_F,\quad \mathcal{I}_S[H]:=\frac{1}{2}S^{-1}HS^{-1}.$$

The KL generator satisfies

$$\nabla^2\Phi_0(S)[H]=S^{-1}HS^{-1}=2\mathcal{I}_S[H].$$

Hence its Hessian equals twice the Gaussian Fisher information operator (Lin et al., 2026). SQ does not share the Gaussian Fisher geometry. At $S=sI$, its Hessian is $s^{-3/2}H$, whereas the Gaussian Fisher operator is $H/(2s^2)$. Their ratio, $2\sqrt{s}$, varies with s, so they cannot agree. Among the four divergences, only KL has a Hessian proportional to the Gaussian Fisher operator for every $S\succ 0$.

C Preconditioner Update Details

We describe the factor updates of original Shampoo, PSGD-Kron, SQ-Shampoo, and KL-Shampoo, then specify the factor statistics, offline fitting procedure, and online EMA scheme used in our experiments.

C.1 Original Shampoo

In the EMA form of Shampoo (Gupta et al., 2018; Anil et al., 2020), the marginal statistics satisfy

$$L_t = \beta_2 L_{t-1} + (1-\beta_2) G_t G_t^\top, \quad R_t = \beta_2 R_{t-1} + (1-\beta_2) G_t^\top G_t.$$

Ignoring momentum, the parameter update is $\Theta_{t+1} = \Theta_t - \eta_t L_t^{-p} G_t R_t^{-p}$. This applies the inverse square root of $S_t = L_t^{2p} \otimes R_t^{2p}$: $p = 1/4$ for original Shampoo, while $p = 1/2$ applies the inverse square root of $L_t \otimes R_t$.

C.2 PSGD-Kron, SQ-Shampoo, and KL-Shampoo

Under $P = S^{-1/2}$, SQ-Shampoo shares PSGD's population gradient-whitening objective (Equation (12)), but uses different factor updates.

Factor updates. PSGD parameterizes the preconditioner as $P = Q^\top Q$ and optimizes Q on a matrix Lie group (Li, 2019). In the Kronecker form,

$$Q = Q_L \otimes Q_R, \quad P = P_L \otimes P_R, \quad P_L := Q_L^\top Q_L, \quad P_R := Q_R^\top Q_R.$$

With row-wise vectorization, the parameter update is $\Theta_{t+1} = \Theta_t - \eta_t P_{L,t} G_t P_{R,t}$. For the original relative-gradient coordinates $dQ = \mathcal{E}Q$, the factors follow

$$Q_{L,t+1} = (I - \mu_t \mathcal{R}_{L,t}) Q_{L,t}, \quad Q_{R,t+1} = (I - \mu_t \mathcal{R}_{R,t}) Q_{R,t},$$

where $\mathcal{R}_{L,t}, \mathcal{R}_{R,t}$ are the factor-wise relative gradients of the whitening objective in Equation (1). The official implementation also provides other update geometries, including $dQ = Q\mathcal{E}Q$.
Our SQ- and KL-Shampoo implementations follow the spectral EMA scheme of Lin et al. (2026). Both average the factor statistics over time, update eigenvalue estimates in the current eigenbases, and periodically refresh bases. The two methods differ in the statistics supplied to this scheme: SQ uses the square-root divergence statistics, while KL uses the KL statistics in Table 6. Algorithm 1 gives the full update.

Table 5: Objectives and update schemes discussed here. Objectives are expressed in S, up to positive scale and additive constants; the PSGD row refers to gradient whitening.

Method	**Population objective**	**Factor updates discussed here**
PSGD-Kron	SQ divergence	Original relative-gradient steps; other geometries are available
SQ-Shampoo	SQ divergence	Spectral EMA of SQ statistics
KL-Shampoo	KL divergence	Spectral EMA of KL statistics

C.3 Divergence-Based Factor Statistics

Table 6 gives the factor statistics for F-, VN-, SQ-, and KL-Shampoo, obtained by setting $q = q_{\mathrm{D}}$ in Equation (3). These statistics define both the offline and online updates that follow. For SQ-Shampoo, $q = 1/2$ gives the bilateral fixed-point iteration

$$L^{(k+1)} = \frac{\mathbb{E}[G(R^{(k)})^{-1/2} G^\top]}{\mathrm{Tr}((R^{(k)})^{1/2})}, \quad R^{(k+1)} = \frac{\mathbb{E}[G^\top (L^{(k)})^{-1/2} G]}{\mathrm{Tr}((L^{(k)})^{1/2})}.$$

Table 6: Instantaneous factor statistics for the Shampoo variants.

Method	$\boldsymbol{\Psi_L(G; L, R)}$	$\boldsymbol{\Psi_R(G; L, R)}$
F-Shampoo	$GRG^\top / \mathrm{Tr}(R^2)$	$G^\top LG / \mathrm{Tr}(L^2)$
VN-Shampoo	$GG^\top / \mathrm{Tr}(R)$	$G^\top G / \mathrm{Tr}(L)$
SQ-Shampoo	$GR^{-1/2}G^\top / \mathrm{Tr}(R^{1/2})$	$G^\top L^{-1/2} G / \mathrm{Tr}(L^{1/2})$
KL-Shampoo	$GR^{-1}G^\top / n$	$G^\top L^{-1} G / m$

C.4 Offline Fixed-Point Fits

Given N gradients, the offline bilateral rule for F-, VN-, SQ-, and KL-Shampoo is

$$\widetilde{L}_{\mathrm{D}}^{(k+1)} := \frac{1}{N}\sum_{b=1}^{N} \Psi_{\mathrm{D},L}(G_b; L_{\mathrm{D}}^{(k)}, R_{\mathrm{D}}^{(k)}), \quad \widetilde{R}_{\mathrm{D}}^{(k+1)} := \frac{1}{N}\sum_{b=1}^{N} \Psi_{\mathrm{D},R}(G_b; L_{\mathrm{D}}^{(k)}, R_{\mathrm{D}}^{(k)}),$$

followed by under-relaxation,

$$L_{\mathrm{D}}^{(k+1)} = (1-\alpha)L_{\mathrm{D}}^{(k)} + \alpha\widetilde{L}_{\mathrm{D}}^{(k+1)}, \quad R_{\mathrm{D}}^{(k+1)} = (1-\alpha)R_{\mathrm{D}}^{(k)} + \alpha\widetilde{R}_{\mathrm{D}}^{(k+1)}.$$

We initialize $L^{(0)} = I_m$ and $R^{(0)} = I_n$, use $\alpha = 0.3$, and iterate to numerical convergence.

Remark on few-sample fitting. We later consider fits with $N < mn$, where the empirical second moment $\widehat{C}_N := N^{-1}\sum_{b=1}^{N} g_b g_b^\top$ is necessarily singular. This does not affect the iterations above: they require only positive-definite factors. For KL, the updates correspond to the criterion $\mathrm{Tr}(\widehat{C}_N S^{-1}) + \log\det S$, with $S = L \otimes R$, which remains finite for singular $\widehat{C}_N$.

C.5 Online EMA Updates

Following the KL-Shampoo implementation of (Lin et al., 2026), the optimizer maintains factor accumulators (A_L, A_R) for the eigenbases and separate eigenvalue estimates (λ_L, λ_R). Writing $\bar{L} = Q_L \,\mathrm{Diag}(\lambda_L) Q_L^\top$ and $\bar{R} = Q_R \,\mathrm{Diag}(\lambda_R) Q_R^\top$, Algorithm 1 applies to $\mathrm{D} \in \{\mathrm{F}, \mathrm{VN}, \mathrm{SQ}, \mathrm{KL}\}$.

Algorithm 1 Online EMA update for a Shampoo variant

Require: Matrix gradients G_t, current momentum M_t, method D, second-moment EMA decay β_2, basis period T, and initialized state $(A_L, A_R, Q_L, Q_R, \lambda_L, \lambda_R)$

for $t = 1, 2, \ldots$ **do**
  Form $\bar{L} = Q_L \,\mathrm{Diag}(\lambda_L) Q_L^\top$ and $\bar{R} = Q_R \,\mathrm{Diag}(\lambda_R) Q_R^\top$
  Precondition M_t as $\widetilde{M}_t \leftarrow Q_L \,\mathrm{Diag}(\lambda_L^{-1/2}) Q_L^\top M_t Q_R \,\mathrm{Diag}(\lambda_R^{-1/2}) Q_R^\top$
  $\Delta_L \leftarrow \Psi_{\mathrm{D},L}(G_t; \bar{L}, \bar{R})$, $\Delta_R \leftarrow \Psi_{\mathrm{D},R}(G_t; \bar{L}, \bar{R})$
  $A_L \leftarrow \beta_2 A_L + (1-\beta_2)\Delta_L$, $A_R \leftarrow \beta_2 A_R + (1-\beta_2)\Delta_R$
  $\lambda_L \leftarrow \beta_2 \lambda_L + (1-\beta_2)\,\mathrm{diag}(Q_L^\top \Delta_L Q_L)$
  $\lambda_R \leftarrow \beta_2 \lambda_R + (1-\beta_2)\,\mathrm{diag}(Q_R^\top \Delta_R Q_R)$
  if $t \bmod T = 0$ **then**
    Refresh Q_L, Q_R from A_L, A_R by eigendecomposition/QR, while preserving λ_L, λ_R
  end if
  Return the preconditioned statistic $\widetilde{M}_t$
end for

The input $M_t = \beta_1 M_{t-1} + (1-\beta_1) G_t$ is the current momentum, with $M_t = G_t$ when momentum is disabled. It is preconditioned before G_t is added into the preconditioner statistics. The coefficient β_2 controls the EMA of both the factor accumulators and eigenvalue estimates. In the EMA diagnostics, $T = 5$ for all four methods.

D Synthetic Validation

We examine spectral error allocation and finite-sample stability using a smooth, long-tailed second moment that is not exactly Kronecker.

D.1 Experimental Setup

For matrix gradients $G_b \in \mathbb{R}^{m\times n}$, the empirical gradient second moment is

$$\widehat{C} := \frac{1}{N}\sum_{b=1}^{N} g_b g_b^\top, \quad g_b := \mathrm{vec}_{\mathrm{row}}(G_b).$$

F-, VN-, SQ-, and KL-Shampoo share samples and use the offline updates in Appendix C. Table 7 shows the detailed settings.

Table 7: Settings for the synthetic experiments.

Setting	$m \times n$	N	Additional parameters
Smooth long-tailed	16×24	4096	scale $= 10^{-1}$, $\sigma = 0.02$

Smooth long-tailed spectrum. We combine three matrix-normal components with different eigenvalue decay rates and orientations:

$$G_b := 10^{-1} \left(\sum_{k=1}^{3} w_k L_k Z_{bk} R_k^\top + \sigma E_b \right), \tag{17}$$

where Z_{bk} and E_b have i.i.d. standard Gaussian terms. The weights $(w_1, w_2, w_3) = (1, 0.3, 0.15)$ give the components different strengths. We control their eigenvalue decay through

$$L_k := U_k \operatorname{Diag}\big(i^{-p_k/2}\big)_{i=1}^{m}, \quad R_k := V_k \operatorname{Diag}\big(j^{-p_k/2}\big)_{j=1}^{n},$$

where U_k, V_k are independently sampled orthogonal matrices and $(p_1, p_2, p_3) = (1.8, 1.2, 0.8)$. The second-moment factors $L_k L_k^\top$ and $R_k R_k^\top$ have eigenvalues i^{-p_k} and j^{-p_k}, respectively, so larger p_k gives faster decay. The resulting second moment is generally not a single Kronecker product. We also add isotropic noise with $\sigma = 0.02$ and scale the gradients by 10^{-1}.

Fitting. We generate a fixed pool of 4096 gradients. All methods start from $L = I_{16}$ and $R = I_{24}$ and use a relaxation coefficient of 0.3. We stop when the sum of the two factors' relative Frobenius changes falls below 10^{-10}, or after 10,000 iterations.

D.2 Metrics and Results

For $S := L \otimes R$, let $\lambda_i(M)$ denote descending eigenvalues and $U_{M,k}$ the leading k eigenvectors. We measure cumulative eigenvalue error and leading-subspace overlap by

$$E(k) := \frac{1}{k} \sum_{i=1}^{k} \left| \log_{10} \frac{\lambda_i(S)}{\lambda_i(\widehat{C})} \right|, \quad O(k) := \frac{1}{k} \left\| U_{\widehat{C},k}^\top U_{S,k} \right\|_F^2.$$

Finite-sample stability. For each divergence D, let $S_{\mathrm{D}}^{\mathrm{ref}}$ be its fit to the full 4096-sample pool and $S_{N,r}^{\mathrm{D}}$ its fit to N samples in repeat r. We measure the squared relative Frobenius error

$$E_{\mathrm{rel},r}^{\mathrm{D}}(N) := \frac{1}{mn} \left\| (S_{\mathrm{D}}^{\mathrm{ref}})^{-1/2} S_{N,r}^{\mathrm{D}} (S_{\mathrm{D}}^{\mathrm{ref}})^{-1/2} - I_{mn} \right\|_F^2.$$

This error is invariant to reciprocal factor rescaling. If $S_{\mathrm{D}}^{\mathrm{ref}} = L_\star \otimes R_\star$ and $S_{N,r}^{\mathrm{D}} = \widehat{L}_{\mathrm{D}} \otimes R_\star$, it reduces to $1/m \|\widetilde{L}_{\mathrm{D}} - I_m\|_F^2$, whose expectation is $\mathrm{Err}_{\mathrm{D}}^{L}$ in Proposition 3. Here both factors are estimated jointly, and the reference is each method's large-sample Kronecker fit.

We use 20 random permutations of the same pool and fit the first $N \in \{10, 20, 40, 80, 160, 320, 640\}$ gradients. Figure 4 (right) plots $1/20 \sum_{r=1}^{20} E_{\mathrm{rel},r}^{\mathrm{D}}(N)$.

Figure 7 shows that F prioritizes large eigenvalues but severely underestimates small ones; VN follows more mildly. SQ and KL have larger fitted tail eigenvalues. The mean stability errors follow KL $<$ SQ $<$ VN $<$ F at every tested N, consistent with the theoretical ordering despite jointly fitting the non-Kronecker second moment.

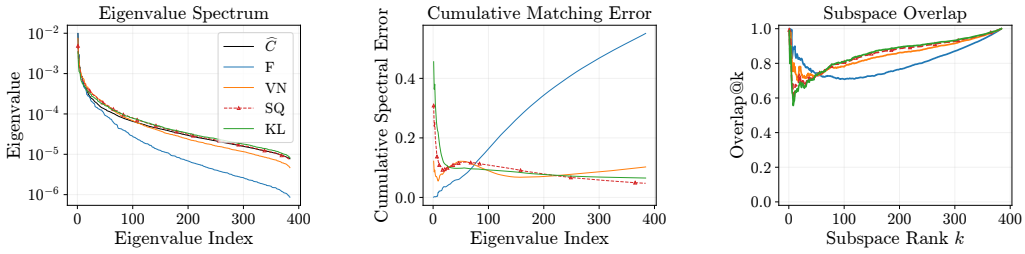


Figure 7: Smooth synthetic setting. From left to right: empirical and fitted spectra, cumulative spectral error, and leading-subspace overlap. Finite-sample stability is shown in Figure 4 (right).

D.3 Directional Gain Interventions on a Quadratic

We test whether gain differences explain the methods' training performance by replacing KL's leading or tail gains with those of F or SQ in a stochastic quadratic.

Setup. To model stochastic gradient noise, we use linear regression with squared loss $\ell(x; z, y) = \frac{1}{2}(z^\top x - y)^2$, where $z \sim \mathcal{N}(0, H)$ and $y \sim \mathcal{N}(0, 1/6)$ are independent. Its population loss is $\mathbb{E}[\ell(x; z, y)] = \frac{1}{2}x^\top Hx + 1/12$, so it is equivalent to the quadratic $f(x) = \frac{1}{2}x^\top Hx$.
We set $d = 384$ and $H = dC_{\text{true}}/\operatorname{Tr}(C_{\text{true}})$, where C_{true} is the population second moment of the generator in Equation (17). Each minibatch gradient averages $B = 64$ sample gradients. Its population second moment is $C(x) = \mathbb{E}[g_B(x)g_B(x)^\top]$, with expectation over minibatch draws:

$$g_B(x) = \frac{1}{B}\sum_{b=1}^{B} z_b(z_b^\top x - y_b), \qquad C(x) = \frac{(x^\top Hx + 1/6)H}{B} + \left(1 + \frac{1}{B}\right)(Hx)(Hx)^\top.$$

At each step, we fit the Kronecker factors to 40 independent mini-batch gradients reshaped into 16×24, then update with a separate mini-batch. We fit the factors using the offline updates in Appendix C, with no EMA.
We run F, VN, SQ, and KL for 1,000 steps at a shared learning rate $\eta = 2 \times 10^{-3}$ (Figure 8). Within each seed, all methods share the same initialization and mini-batches. We initialize $x_0 = H^{-1/2}v$, with a Gaussian vector v normalized to norm 2, giving $f(x_0) = 2$.
Gain replacement. At the current x, fit KL and the comparison method $\mathrm{D} \in \{\mathrm{F}, \mathrm{SQ}\}$ to the same gradients. Write $C(x) = U\Lambda U^\top$ in decreasing eigenvalue order and $P_{\mathrm{D}} = (L_{\mathrm{D}} \otimes R_{\mathrm{D}})^{-1/2}$. For selected directions $\mathcal{I}$, set

$$r_i = \begin{cases} \|P_{\mathrm{D}}u_i\|/\|P_{\mathrm{KL}}u_i\|, & i \in \mathcal{I}, \\ 1, & i \notin \mathcal{I}, \end{cases} \qquad \widetilde{P} = P_{\mathrm{KL}}U\,\mathrm{Diag}(r_i)U^\top.$$

We replace KL's gains in the first 39 (top 10%) or last 96 (bottom 25%) directions by rescaling each $P_{\mathrm{KL}}u_i$ to match $\|P_{\mathrm{D}}u_i\|$. Starting from the same KL iterate at step 900, we continue to step 1,000 using the original training minibatches and learning rate, recomputing the factors and population eigenbasis at each step.
Results. F reaches the highest loss (Figure 8, left). Replacing either region of KL's gains with F's raises loss substantially; replacing them with SQ's leaves loss close to KL's (middle and right). Table 8 summarizes the final losses.

Table 8: Quadratic loss averaged over the last 20 updates and 3 seeds. KL achieves 0.00844.

	Replaced KL gains	
Gain source	**Top**	**Tail**
F	1.441	0.03333
SQ	0.00857	0.00855

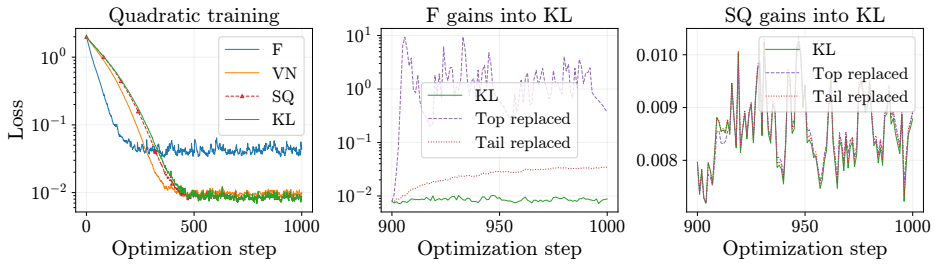


Figure 8: Quadratic training loss averaged over 3 seeds. Left: four methods. Middle and right: replacing KL gains from step 900. F gains raise loss in either region; SQ gains leave it close to KL.

E Real-Gradient Experimental Details

We analyze 128×128 query, value, and attention-output gradients from block 5 of a 12-layer, four-head GPT-2 model trained with AdamW. We freeze the model at steps 100, 2500, and 4500, giving nine checkpoint–projection pairs. Table 9 summarizes the shared settings; Appendix E.6 extends the diagnostics to Distributed Shampoo checkpoints.

Table 9: Shared settings for real-gradient diagnostics.

Setting	**Specification**
Gradient pools	Four splits per experiment; 20,480 FP32 mini-batch gradients per split, assigned by round-robin interleaving
Gradient collection	Batch size 64; sequence length 512
Large-sample fit	All 20,480 gradients in split 0
Few-sample fit	40 gradients from split 0; four repeats
Fixed-point solver	Relaxation $\alpha = 0.3$; tolerance 10^{-8}; at most 1,000 steps
Online EMA	16 disjoint chronological repeats from split 0, each with 512 gradients; $\beta_2 = 0.95$; eigendecomposition every five steps

All four methods retain their eigenvalue EMAs during each eigenbasis refresh. Reliability uses all splits. For gain calibration, split 0 supplies the fits and source eigenvectors; splits 1, 2, and 3 supply the held-out directional scales in Section 5.2.

Numerical details. Second moments and eigendecompositions use FP64. We sort eigenvalues in decreasing order and retain those exceeding $d \times 2^{-52}$ times the largest eigenvalue, where $d = 16{,}384$ and 2^{-52} is FP64 machine precision. For reliability summaries, retained directions are grouped by rank into top 0–10%, middle 10–40%, lower 40–75%, and tail 75–100%.

E.1 Empirical Second Moments and Finite-Sample Effects

E.1.1 Finite-Sample Tail Selection

Cross-fit control. We take eigenvectors $u_{b,i}$ of C_b from a third split b and compare their scales on splits s and t:

$$\xi^{\mathrm{cf}}_{b;s,t,i} := \log_{10} \frac{u_{b,i}^\top C_t u_{b,i}}{u_{b,i}^\top C_s u_{b,i}}, \quad b \notin \{s,t\}.$$

The direction $u_{b,i}$ is therefore selected independently of both second-moment estimates being compared. We compare the six pairs with $s < t$, using each remaining split to select directions. This gives 12 comparisons without reverse pairs that would cancel the signed log-ratios. Cross-fit scale errors have medians and IQRs near zero across the spectrum (Figure 9), supporting sampling-induced selection bias rather than distribution shift.

Selection bias. Another independent split estimates the population second moment along each source eigenvector. The identity below shows that underestimating the source eigenvalue raises the expected held-out ratio, with larger relative errors giving larger ratios. Let $\rho_{s\to t,i} := 10^{\xi_{s\to t,i}}$ denote the ratio before taking the logarithm in Equation (9).

Proposition 5 (Finite-Sample Tail Selection). *Let $C_s = C + E_s$ and $C_t = C + E_t$ be independent empirical gradient second moments drawn from the same distribution, where E_s and E_t denote their zero-mean finite-sample estimation errors. For a source eigenpair $C_s u_{s,i} = \lambda_{s,i} u_{s,i}$ with $\lambda_{s,i} > 0$,*

$$\mathbb{E}[\rho_{s\to t,i} \mid C_s] = \frac{u_{s,i}^\top C u_{s,i}}{\lambda_{s,i}} = 1 - \frac{u_{s,i}^\top E_s u_{s,i}}{\lambda_{s,i}}. \tag{18}$$

Thus, if the source error along this direction is negative, $u_{s,i}^\top E_s u_{s,i} < 0$, its conditional expected held-out ratio is larger than one.

Proof. Conditional on C_s, the vector $u_{s,i}$ and the denominator $\lambda_{s,i}$ are fixed. Independence and $\mathbb{E}[C_t] = C$ give $\mathbb{E}[\rho_{s\to t,i} \mid C_s] = u_{s,i}^\top C u_{s,i}/\lambda_{s,i}$. Substituting $C = C_s - E_s$ yields Equation (18). □

Why ratios rise toward the tail. Selecting directions with small sample eigenvalues can favor downward sampling fluctuations, so their scales may increase when evaluated on an independent split. In Equation (18), the downward estimation error is divided by the source eigenvalue $\lambda_{s,i}$. Small tail eigenvalues can therefore turn downward errors into large held-out ratios.

E.1.2 High-Dimensional Sampling and Tail Signals

Finite-sample spectral distortion is well documented in high-dimensional covariance estimation (Ledoit & Wolf, 2012; Bartz, 2016), and similar selection bias has also been observed in mini-batch Hessian and GGN estimates (Tatzel et al., 2025). With $d/N = 0.8$, the spectral distortion observed here may partly be explained by Marčenko–Pastur-type spreading (Bai & Silverstein, 2010). However, this does not explain why the leading scales are much more consistent across splits than the tail. Prior work has reported alignment between the leading subspaces of the empirical second moment and the Hessian (Zhu et al., 2019; Wu et al., 2022; Liu et al., 2026). This may help explain the stronger reproducibility of the leading scales, although the connection remains to be established.

Tail directions and learning signals. If class k occurs with probability p_k and produces gradient $g = a_k u_k$, with orthonormal u_k, then $C = \sum_k p_k a_k^2 u_k u_k^\top$. For comparable $|a_k|$, infrequent classes correspond to directions with small second-moment scales. This illustrates how tail directions can carry useful signals (Wang et al., 2026; Kim et al., 2026). However, finite-sample noise can still severely underestimate their scales.

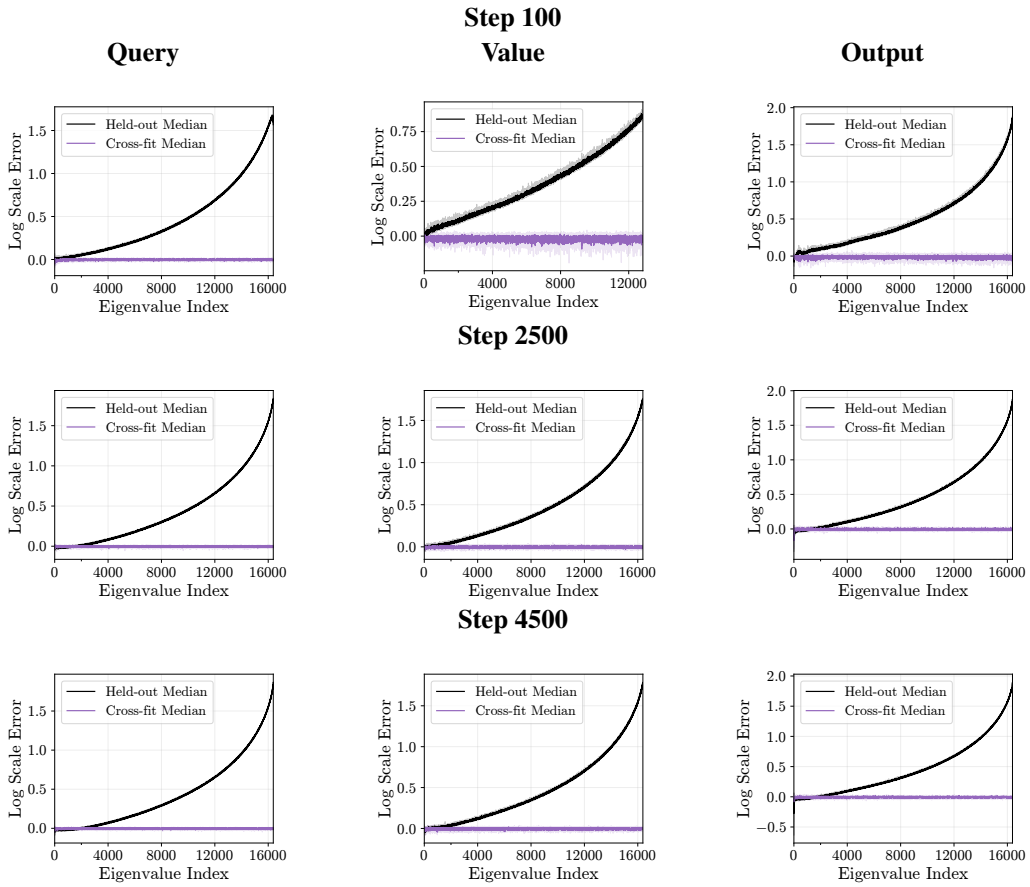
E.2 Reliability Across Checkpoints

Figure 9 shows held-out scale error for all 9 experiments.
Table 10 reports all four regions, including absolute held-out scale error. We take the median over all directions in each region and all 12 comparisons within an experiment, then report the mean and sample standard deviation across the nine checkpoint–projection pairs per training optimizer.

Table 10: Held-out scale error across all four eigenvalue regions. Zero indicates agreement; positive ξ indicates underestimation, and larger $|\xi|$ indicates greater mismatch. Entries show mean ± sample SD across nine checkpoint–projection pairs per training optimizer, using each's regional median.

AdamW checkpoints				
Metric	**Top (0–10%)**	**Middle (10–40%)**	**Lower (40–75%)**	**Tail (75–100%)**
$\lvert\xi\rvert$ (↓)	0.027 ± 0.016	0.124 ± 0.037	0.431 ± 0.030	1.009 ± 0.120
Signed ξ	0.010 ± 0.030	0.124 ± 0.037	0.431 ± 0.030	1.009 ± 0.120
Cross-fit ξ^{cf}	−0.008 ± 0.004	−0.006 ± 0.004	−0.006 ± 0.005	−0.007 ± 0.006
Distributed Shampoo checkpoints				
Metric	**Top (0–10%)**	**Middle (10–40%)**	**Lower (40–75%)**	**Tail (75–100%)**
$\lvert\xi\rvert$ (↓)	0.020 ± 0.009	0.104 ± 0.030	0.429 ± 0.025	1.048 ± 0.017
Signed ξ	0.002 ± 0.020	0.104 ± 0.030	0.429 ± 0.025	1.048 ± 0.017
Cross-fit ξ^{cf}	−0.009 ± 0.005	−0.008 ± 0.004	−0.007 ± 0.004	−0.007 ± 0.004

Figure 9: Held-out scale error. Rows: steps 100, 2500, 4500. Bands show split-comparison IQRs.



E.3 Inverse-Root Gain Across Checkpoints

Figures 10–12 show the gains of large-sample fits, few-sample fits, and EMA states at all three checkpoints. Within each checkpoint–projection pair, all fits use the same split-0 directions and the same scales averaged over splits 1, 2, and 3 (Equation (10)). Evaluations use the numerically retained directions from split 0. The black solid curve shows the source preconditioner $C_0^{-1/2}$, with error $\frac{1}{2}\log_{10}(h_i/\lambda_{0,i})$; the dashed line marks zero.

Large-sample curves use one fit per method. Few-sample and EMA curves average held-out gain error over 4 repeats and 16 repeats, respectively. For Table 2, we take the median absolute held-out gain error over all retained directions and repeats within each experiment, then report the mean and sample standard deviation across nine experiments. This is $\mathrm{median}(|e_{D,i}|)$, not $|\,\mathrm{median}(e_{D,i})|$.

Figure 10: Held-out gain error on AdamW checkpoints at step 100. Rows: large-sample fit, 40-sample fit, EMA with eigendecomposition every five steps.

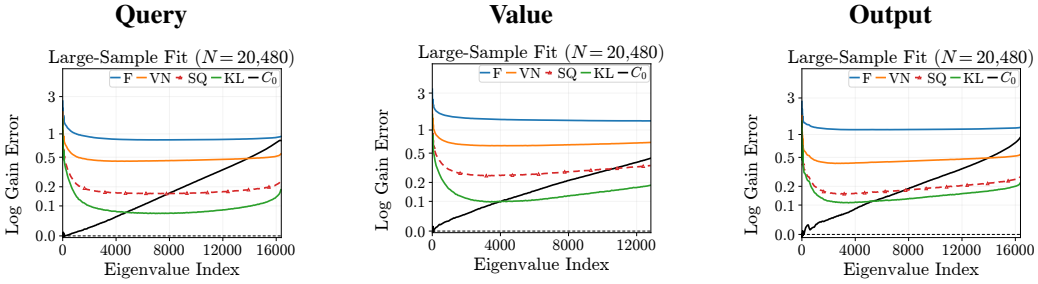


Figure 10: Held-out gain error on AdamW checkpoints at step 100 (continued).

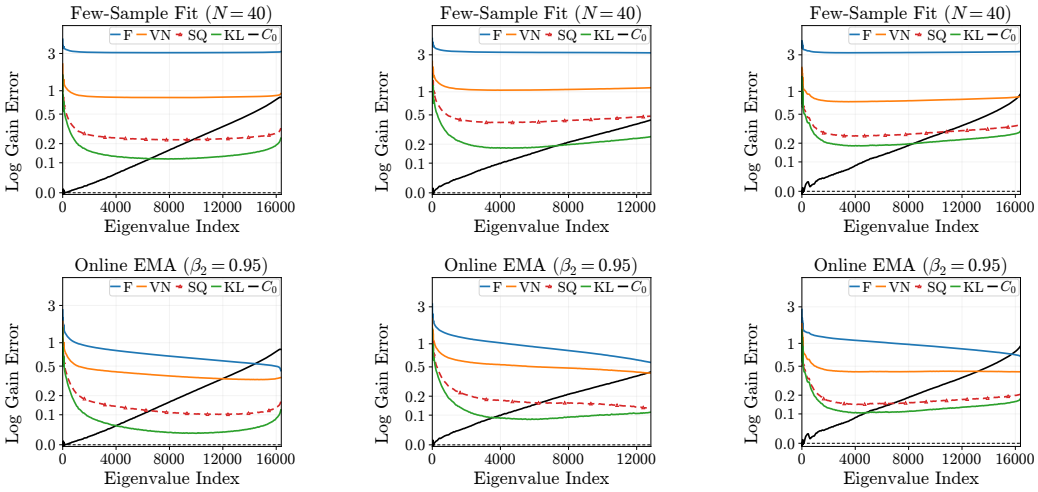


Figure 11: Held-out gain error on AdamW checkpoints at step 2500. Rows: large-sample fit, 40-sample fit, EMA with eigendecomposition every five steps.

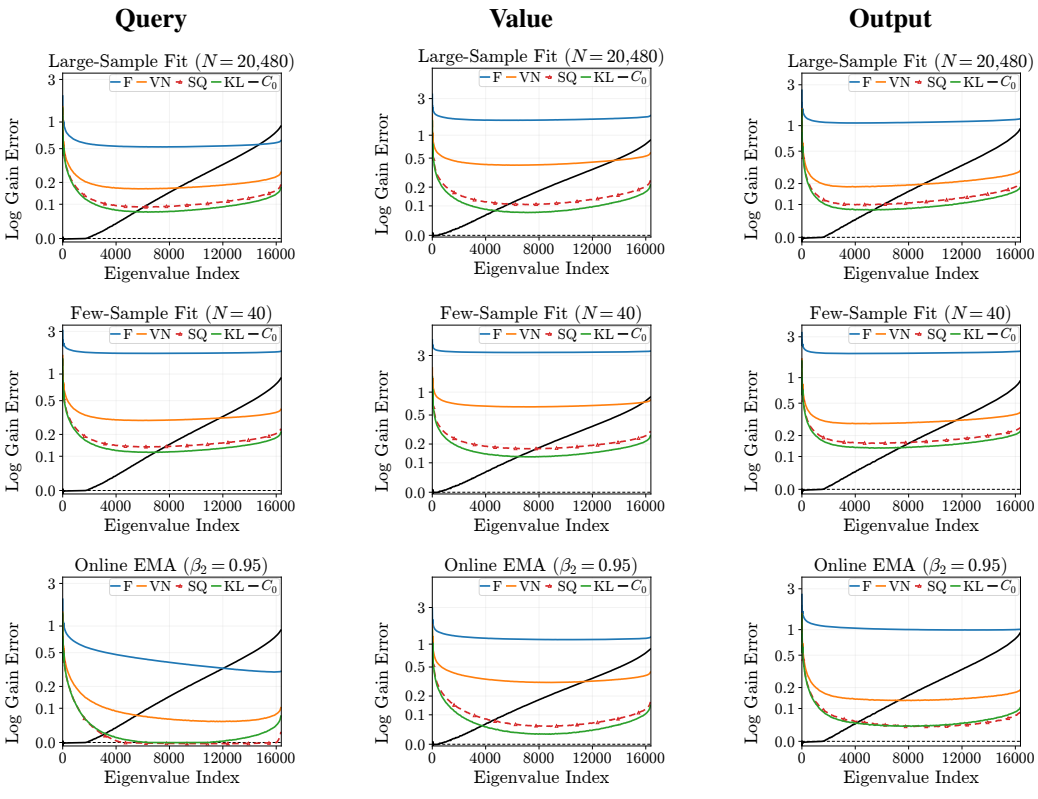
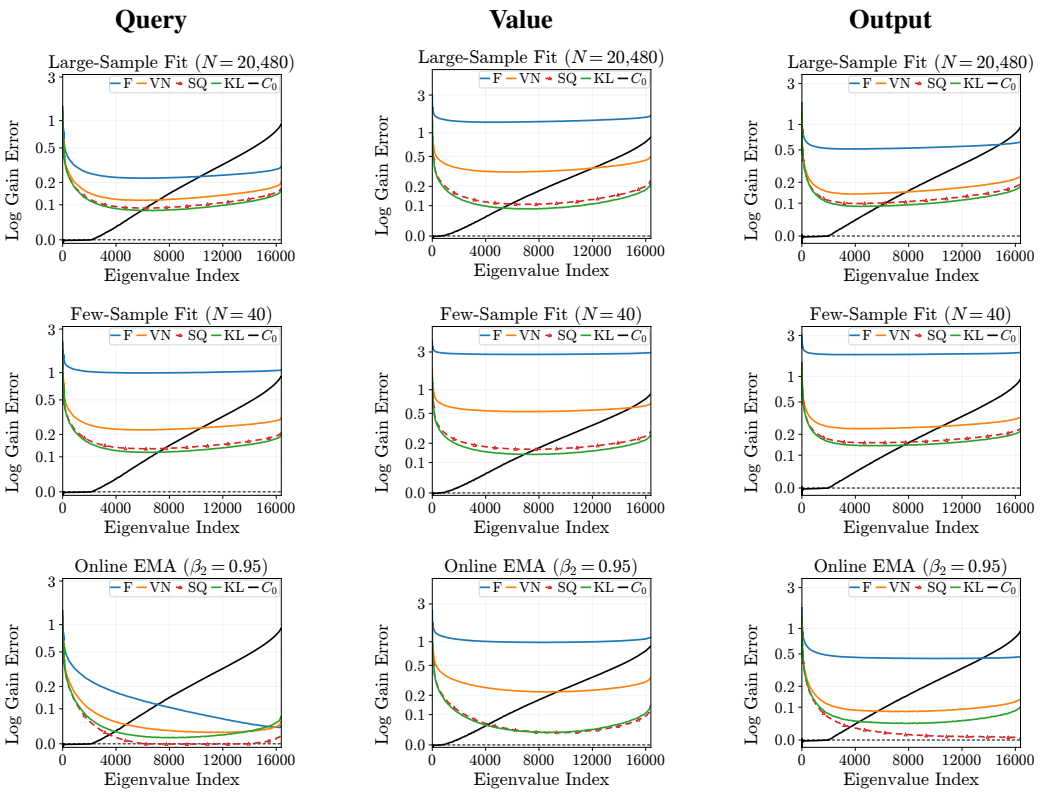


Figure 12: Held-out gain error on AdamW checkpoints at step 4500. Rows: large-sample fit, 40-sample fit, EMA with eigendecomposition every five steps.



E.4 Kronecker Fits of Different Divergences

We compare fitted spectra across all nine checkpoint–projection pairs and show how fitted eigenvalues and eigenvectors determine gains. Each fit uses the same split-0 gradients as the empirical second moment (Figure 13).

Spectral preferences. F and VN put more weight on large eigenvalues and penalize overestimation more than reciprocal underestimation (Table 1). SQ gives large eigenvalues less weight and penalizes reciprocal errors equally. KL weights relative errors equally across scales and penalizes underestimation more strongly. These preferences help explain the larger fitted tail eigenvalues of SQ and KL in Section 5.1.

How small eigenvalues increase gains. Let $(\omega_{D,j}, v_{D,j})$ be the eigenpairs of S_D^{ref} in decreasing order, and write $P_{D,ij} := \langle u_{0,i}, v_{D,j} \rangle^2$. Then

$$10^{2e_{D,i}} = h_i \sum_j \frac{P_{D,ij}}{\omega_{D,j}}.$$

Each overlap is divided by a fitted eigenvalue, so even a small overlap can matter when that eigenvalue is tiny. Since $\sum_j P_{D,ij} = 1$, we can also write

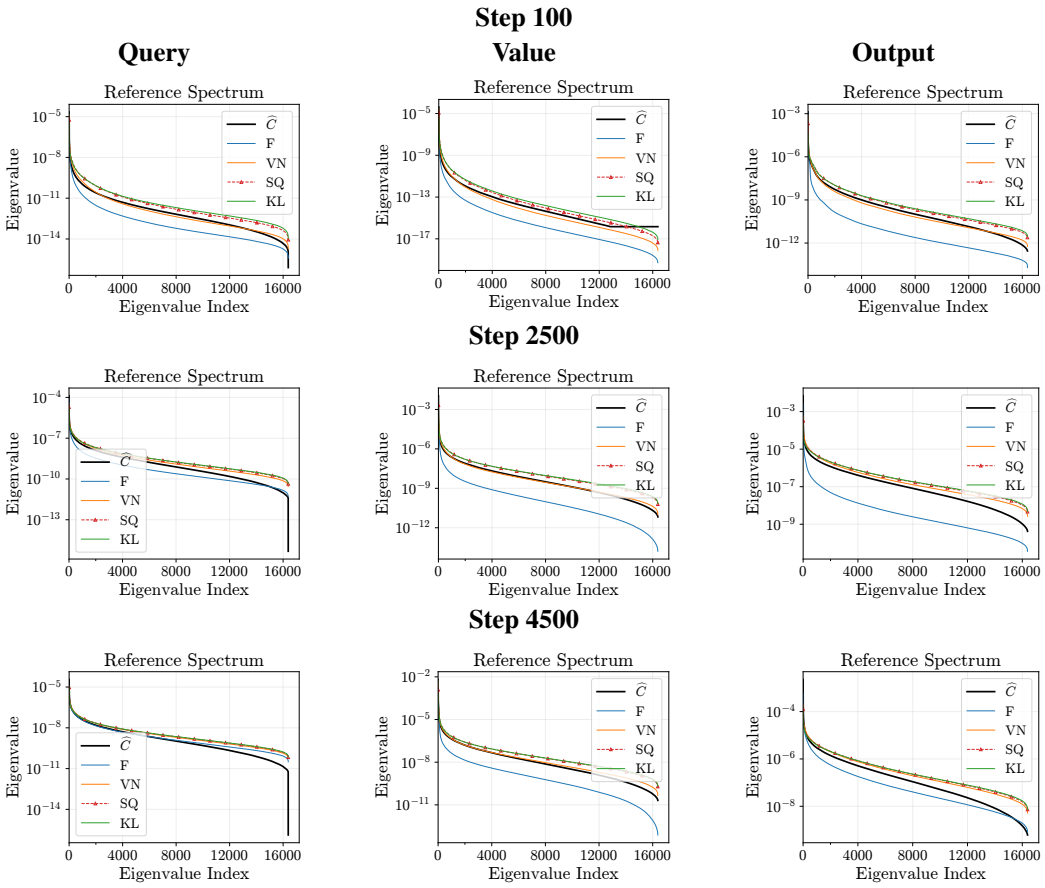
$$10^{2e_{D,i}} = \frac{h_i}{\omega_{D,i}} + h_i \sum_{j \neq i} P_{D,ij} \left(\frac{1}{\omega_{D,j}} - \frac{1}{\omega_{D,i}} \right).$$

With aligned eigenvectors, only the first term remains. The second term shows how gains change when the fitted eigenvectors differ from the source eigenvectors.
Let $\varepsilon_{D,i}$ be the total squared overlap with fitted eigenvectors whose eigenvalues satisfy $\omega_{D,j} \leq \tau$. Then

$$10^{2e_{D,i}} \geq h_i \varepsilon_{D,i} / \tau.$$

If $h_i \varepsilon_{D,i} > \tau$, the gain exceeds the held-out target, even if the leading eigenvalues are fitted well.

Figure 13: Large-sample spectra. Rows: steps 100, 2500, 4500.



E.5 Gain Recovery with Short Gradient Histories

Recovery is measured against each method's own large-sample fit over all 16,384 reference directions (Equation (11)). For Figure 14, we first average the log gain ratios across 4 offline or 16 EMA repeats, and then average every 64 consecutive directions. For Table 3, we compute the median absolute log gain ratio over all directions and repeats without binning, and report the mean and sample SD across nine checkpoint–projection pairs.

Figure 14: Large-sample fit recovery on AdamW checkpoints at step 2500. Rows: F, VN, SQ, KL. Solid: 40-sample fits; dashed: EMA with eigendecomposition every five steps.

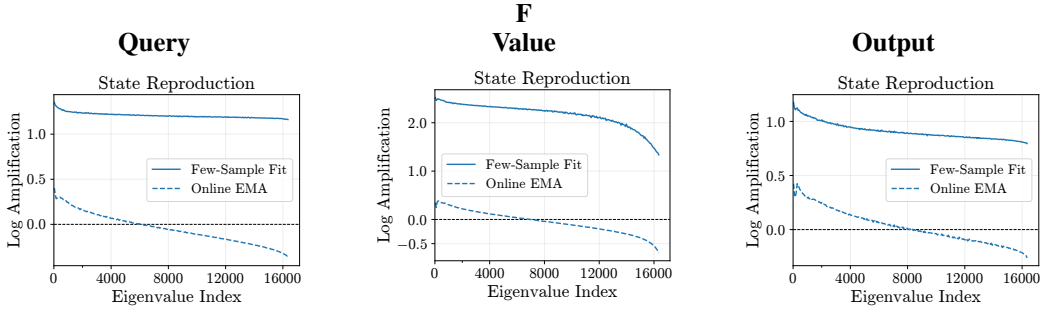
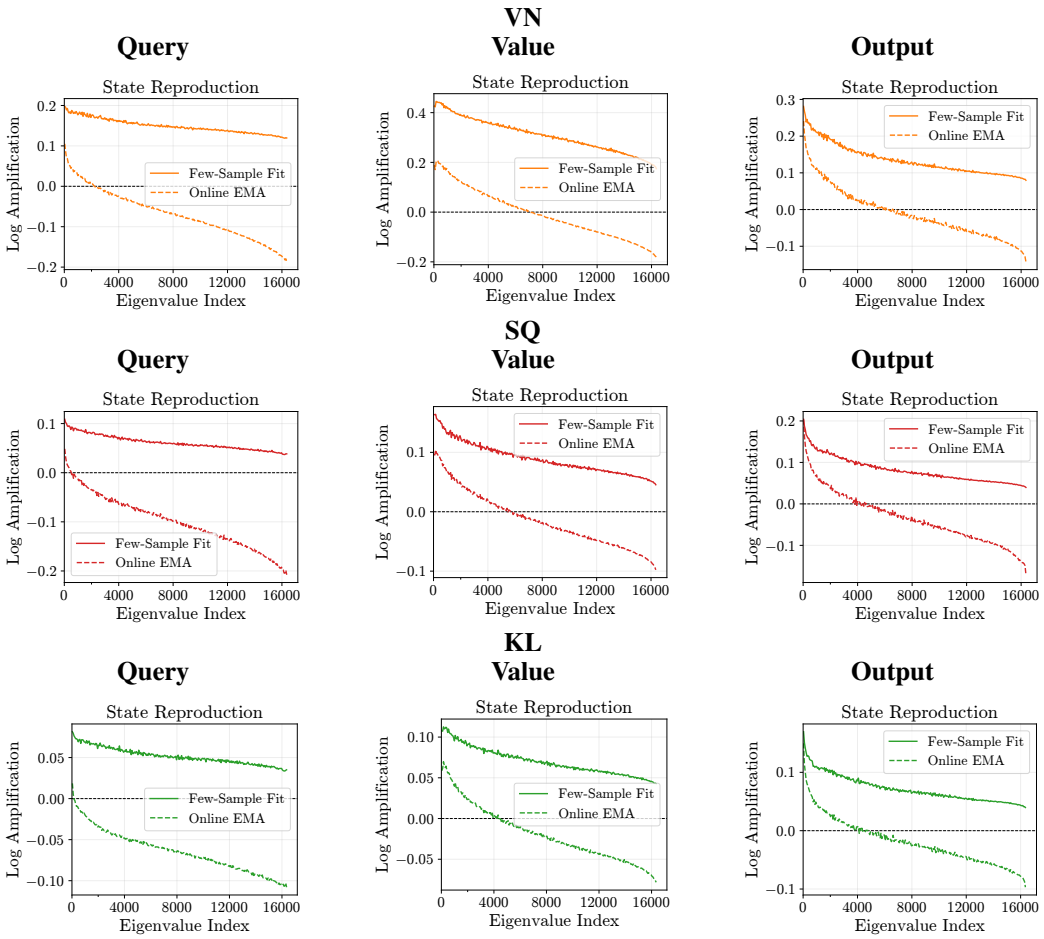


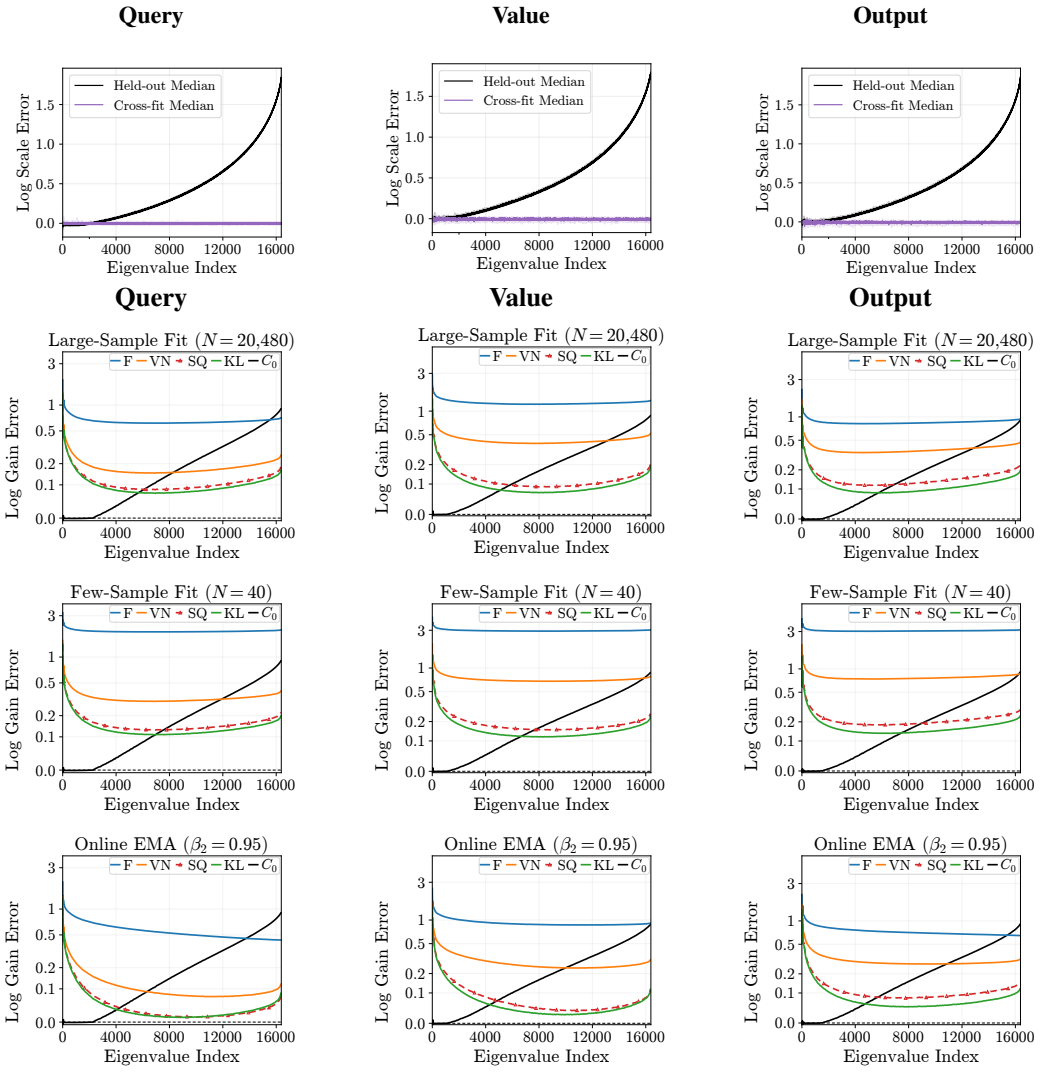
Figure 14: Large-sample fit recovery on AdamW checkpoints at step 2500 (continued).



E.6 Diagnostics on Distributed Shampoo Checkpoints

We train the same GPT-2 model with Distributed Shampoo (Shi et al., 2023), using inverse-square-root factors and Adam grafting.
We use the same checkpoint steps, projections, gradient collection, fitting settings, and aggregation as for AdamW (Appendix E). All four divergences are fitted to the new gradients. Figure 15 shows step 2500; the main-text tables summarize all nine checkpoint–projection pairs.

Figure 15: Distributed Shampoo diagnostics at step 2500. Top: held-out scale error and cross-fit controls. Remaining rows: held-out gain error for large-sample fits, 40-sample fits (four repeats), and EMA (512 gradients, 16 repeats, eigendecomposition every five steps).



F GPT-2 Training Details and Full Results

Models and data. The 12-layer models use the architecture from (Jordan et al., 2024). FineWeb provides 3B training tokens. Each run processes 2.67B tokens with sequence length 1024 and a global batch of 512 sequences.

Training settings. Grid searches at model widths up to 768 found nearly identical optimal learning rates across the tested optimizers. We therefore use a shared learning rate across optimizers at each larger width without further searches. Learning-rate tuning can absorb an overall gain offset, but cannot correct gain errors that vary across directions, layers, or checkpoints. Within each width, methods share initialization, data order, and learning rate (Table 11). We follow (Lin et al., 2026) for the other hyperparameters. The schedule uses 250 warmup steps, a constant-rate phase, and 1450 steps of linear decay. The embedding and other non-matrix valued parameters use AdamW with learning rate 0.0036, $\beta = (0.9, 0.95)$, and no weight decay.

Table 11: Final validation loss after $5,100$ steps; one run per configuration.

Width	Params (M)	LR	F	VN	SQ	KL
256	22.3	0.0024	3.8281	3.7573	**3.7450**	3.7475
512	63.5	0.0024	3.5082	3.4377	**3.4173**	3.4180
768	123.6	0.0018	3.3993	3.2876	3.2661	**3.2660**
1024	202.5	0.0012	3.3528	3.1979	3.1824	**3.1769**
1536	417.0	0.0008	3.3111	3.0972	3.0817	**3.0758**

Training results. Table 11 supplements Table 4 with parameter counts and learning rates. Figure 16 shows the validation-loss trajectories from step 2,000 to the end of training.

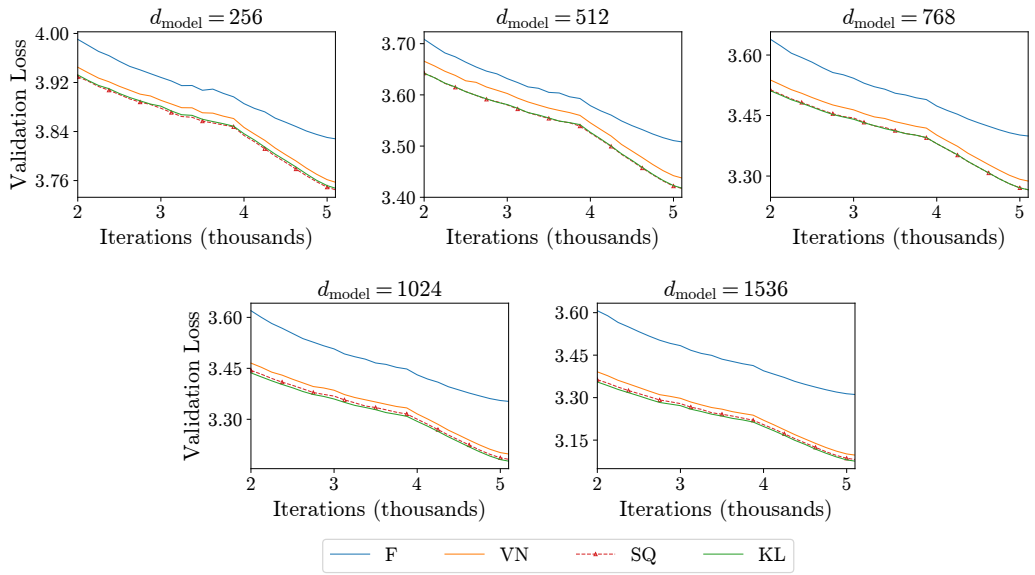


Figure 16: Validation loss from step $2,000$ at all five widths.