EVIE-KF: RISK SENSITIVE KRONECKER FACTORED PRECONDITIONING FOR ADAPTIVE OPTIMIZERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adaptive optimizers such as Adam adapt each update to the magnitude of its coordinate, but this diagonal view discards correlations in gradient noise. Those correlations contain information about which directions are more reliable. We propose Evie-KF, an adaptive optimizer derived from a risk sensitive formulation of stochastic optimization. Evie-KF extends Adam's diagonal normalization with a matrix preconditioner obtained by solving the risk sensitive Riccati equation in coordinates whitened by Adam. The resulting operator, $(I + \gamma\Sigma_z)^{-1/2}$, uses the centred covariance of gradient noise so that correlated directions are treated jointly. It recovers AdamW exactly at $\gamma = 0$ and adds one tuned parameter. Across vision, language, and finance tasks, Evie-KF improves over AdamW and other common adaptive optimizers on most tasks tested across multiple seeds. On CIFAR-100, it improves top 1 accuracy over AdamW by 5.6%, winning in all five seeds. On WikiText-103 language modeling, it improves over AdamW by 11.5% and over SGD by 40.7%, again in all five seeds. On a 33.7M-parameter OpenWebText model, the largest we train, Evie-KF improves over AdamW by 13.1% on three of three seeds at $0.98\times$ its wall time. Ablating the Kronecker structure on its own, against a diagonal gate at matched normalization, it gains 6.75% on a finance panel and 6.67% on WikiText-2, on nine and ten of ten seeds, both surviving Holm correction. **Our code is available**, with Evie-KF released as the **`evie-kf`** Python package.

1 INTRODUCTION

At modern scales, optimizer choice is a cost decision: Meta's Llama 3.1 405B used roughly 31 million H100 GPU hours, with family-wide training on up to 16,000 H100s (Meta AI, 2024; Grattafiori et al., 2024). Adam and AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) remain cheap, tunable defaults, but coordinatewise moments cannot represent cross-coordinate correlation or tail weight.

GLM-130B's diary records five restarts in ten days after loss explosions or gradient spikes, plus two 200-step skips (GLM-130B Team, 2022a). Embedding-gradient norms spiked before collapse, and shrinking them restored stability (Zeng et al., 2023); the 60-day, 768-A100 run was valued at $4.9 million (GLM-130B Team, 2022b). At 546B parameters, Molybog et al. (2023) likewise observed gradient explosion one step before perplexity spikes, sometimes requiring batch skipping, and linked it to Adam's update becoming misaligned with descent. Instability therefore forces restarts, skipped data and wasted compute.

The same issue arises with correlated variables: one macro shock can move a whole sector's gradients at once (Figure 1), structure a coordinatewise noise model cannot represent before it affects the update.

K-FAC, Shampoo and SOAP (Martens & Grosse, 2015; Gupta et al., 2018; Vyas et al., 2025) use Kronecker-factored matrices; Muon (Jordan et al., 2024) orthogonalizes momentum; and Sophia (Liu et al., 2024) clips against a diagonal Hessian proxy. They precondition by the *uncentred* second moment or Fisher, $\mathbb{E}[\hat{g}\hat{g}^\top]$, not centred noise or an explicit trust model. Entropic-risk work instead chooses a constant global step offline (Can & Gürbüzbalaban, 2025).

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

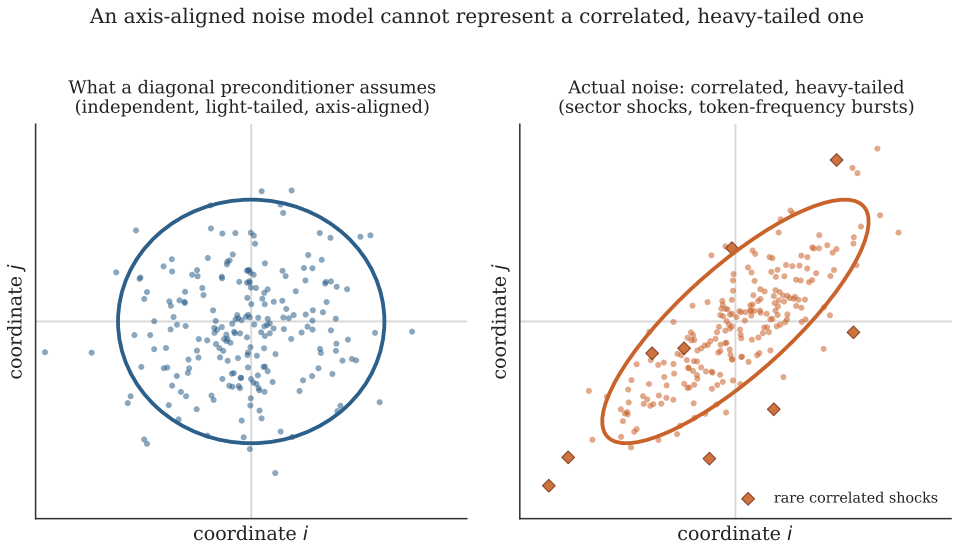


Figure 1: A diagonal preconditioner represents uncorrelated, axis-aligned noise (left), while training noise may be correlated and potentially heavy-tailed with finite second moments (right). Schematic, not measured data.

We close this gap with a closed-form, shrink-only Jacobson–Whittle branch, whose update-noise interpretation and sign convention are derived in Appendix B. A theorem proves that Adam-moment scalar gates only rescale axes and cannot use cross-coordinate covariance, motivating a matrix operator. Adam whitening reduces its Riccati equation to the contraction $(I + \gamma\Sigma_z)^{-1/2}$, recovering AdamW at γ=0. Its centred noise covariance is estimated free by a split-batch sample provably decoupled from the gradient it preconditions.

On a pre-registered 12-task vision, language and finance benchmark, Kronecker structure beats its matched diagonal gate with Holm significance on the two structured-noise hosts. Evie-KF's margin over AdamW grows from matching at the smallest finance model to $+7.6\%$ at the largest (Figure 2); on language it is largest at the smallest, noisiest batch (Figure 3), unlike Muon. Evie-KF is significantly ahead of Sophia overall ($p_{\text{Holm}} = 0.007$). Fixing γ and tuning only learning rate costs 0.9% at the median across four controls, while mean-eigenvalue normalisation cuts the selected γ spread from $2{,}700\times$ to $3.3\times$ on three hosts (Appendix G.3).

Our contributions are:

1. **Risk-sensitive derivation & impossibility theorem.** We derive both Jacobson–Whittle sign branches in closed form. The negative process-noise branch yields a globally defined, shrink-only gate with no reciprocation step, while a complementary update-noise criterion gives its variance-averse interpretation. A structural theorem proves that scalar Adam-moment gates can only rescale axes and cannot exploit covariance eigenvectors, making a matrix operator necessary to use correlated noise. At constant γ, the adaptive schedule reduces exactly to applying this gate directly.

2. **Kronecker-factored risk-sensitive preconditioning.** We propose Evie-KF, which whitens coordinates with Adam's curvature estimate and reduces the governing Riccati equation to $(I + \gamma\Sigma_z)^{-1/2}$, whose eigenvalues lie in $(0, 1]$. Each parameter tensor's noise covariance is factored into row and column Kronecker factors; their joint eigenbasis makes the gate a closed-form shrinkage for each eigenvalue pair.

3. **Extensive empirical validation.** We evaluate Evie-KF on twelve vision, language and finance tasks, comparing eight strong optimizers over five seeds per task. Evie-KF ranks second of eight by mean rank and significantly outperforms Sophia, while remaining statistically on par with AdamW, AdaBelief, Shampoo, SGD, Muon, and Lion. Fixed-γ controls retain the aggregate result under a baseline-sized tuning budget.

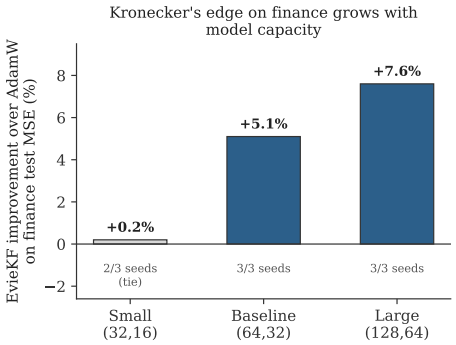


Figure 2: Evie-KF vs. AdamW on finance across model sizes (3 seeds).

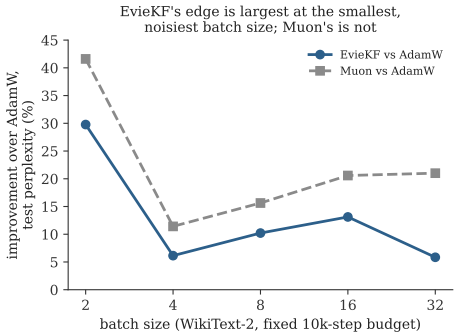


Figure 3: Evie-KF and Muon vs. AdamW on WikiText-2 across batch sizes.

2 Motivation, Related Work, and Setup

Let $\theta \in \mathbb{R}^d$ and $\hat{g} = \nabla L(\theta) + \xi$ be the minibatch gradient with noise covariance $\Sigma = \mathrm{Cov}(\hat{g})$. AdamW keeps bias corrected moments $\hat{m}_1, \hat{v}$; write its denominator $P = \mathrm{diag}(\hat{v}^{\odot 1/2} + \epsilon)$ and whiten by it:

$$z = P^{1/2}\theta, \qquad \nabla_z L = P^{-1/2}\nabla L, \qquad \Sigma_z = P^{-1/2}\Sigma P^{-1/2}, \qquad y := P^{-1/2}\hat{m}_1. \tag{1}$$

The AdamW step is $\Delta\theta_{\mathrm{AdamW}} = -\eta P^{-1/2} y$; every update below has the form $\Delta\theta = -\eta P^{-1/2} z$, with $z = y$ recovering AdamW.

Gradient noise concentrates along rows and columns of a weight matrix (Figure 4a), not coordinate axes, so a diagonal gate, diagonal AdaGrad, RMSProp, Adam, AdamW, Lion, Sophia, vSGD, SDProp, AdaBelief (Duchi et al., 2011; Tieleman & Hinton, 2012; Kingma & Ba, 2015; Loshchilov & Hutter, 2019; Chen et al., 2023; Liu et al., 2024; Schaul et al., 2013; Balles & Hennig, 2018; Ida et al., 2017; Zhuang et al., 2020), can rescale but not rotate toward it; SDProp and AdaBelief's centred variance is exactly this section's diagonal case (Section 3.1). Full-matrix AdaGrad and GGT rotate using uncentred second moments (Duchi et al., 2011; Agarwal et al., 2019); Kronecker methods, K-FAC, EKFAC, Shampoo, SOAP, KL-Shampoo, Muon (Martens & Grosse, 2015; George et al., 2018; Gupta et al., 2018; Anil et al., 2020; Vyas et al., 2025; Lin et al., 2026; Jordan et al., 2024), impose scalable structure on an uncentred second moment or Fisher estimate. Centred gradient-noise covariance has instead been used to choose dense or diagonal preconditioners for approximate Bayesian inference (Mandt et al., 2017); TONGA, natural Newton, and McCandlish's noise scale (Le Roux et al., 2008; Le Roux & Fitzgibbon, 2010; McCandlish et al., 2018) also read that covariance as informative, and Nayak et al. (2026) correct its coupling bias by cross fitted microbatches, which our split batch sample removes structurally at no extra cost. Recent work unifies structured preconditioners and learns covariance-aligned bases (Xie et al., 2025; Nguyen et al., 2026). None combines a centred Kronecker noise estimate with a gate derived from an explicit model of trust; that model is the risk sensitive Riccati equation of Jacobson (1973) and Whittle (1981), so far used only to tune a scalar step size or momentum constant on strongly convex problems (Can & Gürbüzbalaban, 2025; Gürbüzbalaban et al., 2025). In one dimension its optimal step is

$$(1 + \gamma R q)^{-1/2} \in (0, 1], \tag{2}$$

equal to 1 at $\gamma = 0$; a gate of this form from Adam's own moments only reparametrizes AdamW's step length (Section 3.1), so the formulation adds information only once the gain is a matrix, solved exactly in the coordinates of equation 1 as $G = (I + \gamma\Sigma_z)^{-1/2}$, turning the step away from noisy directions (Figure 4) while a rescale keeps AdamW's step length unchanged.

We now derive G from the risk sensitive Riccati equation.

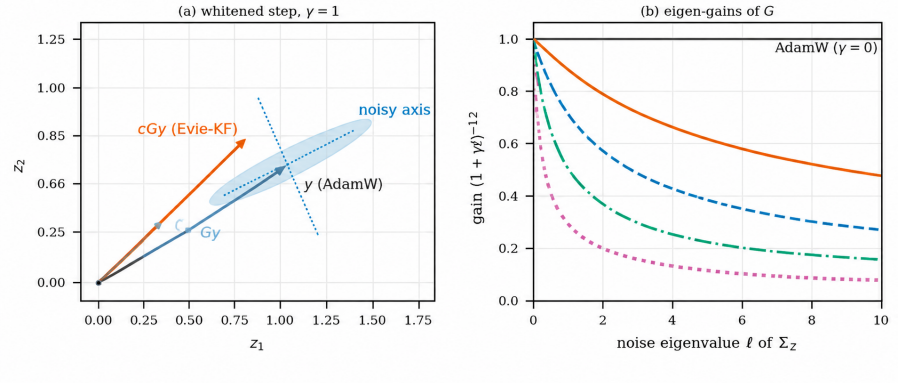


Figure 4: (a) Whitened coordinates. The ellipse is the gradient noise covariance Σ_z around the AdamW direction y. Gy with $G = (I+\gamma\Sigma_z)^{-1/2}$ shrinks the component along the noisy axis. After rescaling, cGy has the length of y and points away from the noise. (b) Eigen gains $(1+\gamma\ell)^{-1/2}$ of G against the noise eigenvalue ℓ for several γ. Each gain lies in $(0, 1]$ and equals 1 at $\ell = 0$.

3 Method

3.1 Risk Sensitive Control Formulation

We use Jacobson's risk sensitive Riccati equation (Jacobson, 1973) with process noise covariance N and risk parameter σ,

$$-\dot{S} = Q + SA + A^\top S - S\big(BR^{-1}B^\top - \sigma N\big)S, \tag{3}$$

under the following assumptions.

(A1) **Local quadratic model.** $L(\theta) \approx L(\theta^*) + \frac{1}{2}x^\top Hx$, $x = \theta - \theta^*$, $H = \nabla^2 L(\theta^*) \succ 0$.

(A2) **Whitened control problem.** In z: $\dot{z} = u_z$ ($A = 0$, $B = I$), state weight I, control weight I. Equivalently, in θ: $\dot{\theta} = u$ with $Q = R = H$, so the risk neutral optimum $u = -x$ is the Newton step.

(A3) **Noise identification.** $\xi \sim \mathcal{N}(0, \Sigma)$ with Σ full. Through $x \approx H^{-1}\nabla L$, gradient noise becomes state noise $H^{-1}\Sigma H^{-1}$, which is Σ_z in z; it is identified with N.

(A4) **Risk sign and dial.** $\sigma = -\gamma \le 0$ with $\gamma \ge 0$ continuous (Whittle, 1981), so the noise term adds to the control penalty. Under the exponential process-noise criterion in Appendix B, this is the optimistic sign branch; the risk-averse interpretation used for Evie-KF is with respect to update noise.

(A5) **State proxy.** $x \approx H^{-1}\nabla L$, with $\hat{m}_1$ in place of ∇L; in z, $x_z \approx y$.

(A6) **Curvature proxy.** $H \approx P$ (not P^2), up to a global scale absorbed by the learning rate.

(A7) **Kronecker noise.** For $W \in \mathbb{R}^{n\times m}$, $\Sigma_z \approx (B \otimes A)/\tau$ with $B \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{m\times m}$, $\tau = \operatorname{tr} A$, under row major $\mathrm{vec_r}$; block diagonal across tensors; a dimension above $n_{\max}$ keeps only its diagonal.

(A8) **Slow variation.** P and Σ_z are constant over one decision, so the algebraic Riccati equation applies.

Theorem 1 (Coordinatewise gates cannot add cross-coordinate information)**.** *Let* $r_i := \hat{m}_{1,i}^2/\hat{v}_i$ *and let each nonnegative gate entry depend only on coordinate* i*'s Adam moments:* $w_i = f_i((\hat{v}_i - \hat{m}_{1,i}^2)_+, \hat{v}_i)$*. Then*

$$\Delta\theta_i = -\eta\ \operatorname{sgn}(\hat{m}_{1,i})\sqrt{r_i}\,F_i(r_i, \hat{v}_i), \qquad F_i(r_i, \hat{v}_i) = f_i\big(\hat{v}_i(1 - r_i)_+, \hat{v}_i\big).$$

Thus distributions with the same coordinatewise moments but different off-diagonal covariance receive the same update. The gate can rescale axes but cannot mix coordinates or align with covariance eigenvectors. Appendix Result R2 proves the statement and records its support and sign consequences.

This is why the gate below is built from the full noise covariance Σ_z instead of a coordinatewise function of Adam's own diagonal statistics: Section 3.2 shows the whitening step that keeps this matrix-valued gate in closed form.

Gaussianity in (A3) only defines the criterion; G and its estimator use second moments. No approximation is made in solving the Riccati equation; the model rests on (A1) through (A8).

Interpretation. Assumption (A4) is the optimistic process-noise branch; the positive entropic branch instead gives $I - \gamma\Sigma_z$ and a finite validity range. Under update noise, positive γ penalizes post-update cost variance and locally decreases the optimal step. Appendix B derives both conventions and prevents this sign distinction from being hidden.

3.2 The Exact Whitened Solution

Theorem 2 (Exact whitened solution). ***(i)*** *With $A = 0$, $B = R = I$, $Q = H \succ 0$, $N = \Sigma \succeq 0$ and (A4), (A8), the steady state of equation 3 is $SMS = H$ with $M = I + \gamma\Sigma \succeq I$, whose unique positive semidefinite solution*

$$S = M^{-1/2}\big(M^{1/2}HM^{1/2}\big)^{1/2}M^{-1/2} \tag{4}$$

exists for every $\gamma \geq 0$. ***(ii)*** *Under (A2), (A3), (A6), (A8), $H_z = P^{-1/2}HP^{-1/2} = I$ and the whitened equation $S_z(I + \gamma\Sigma_z)S_z = I$ has the solution*

$$S_z = G := (I + \gamma\Sigma_z)^{-1/2}. \tag{5}$$

(iii) *With (A5) and an explicit Euler step of length η,*

$$\Delta\theta = -\eta\, P^{-1/2}\big(I + \gamma\Sigma_z\big)^{-1/2}P^{-1/2}\,\hat{m}_1 = -\eta\, P^{-1/2}Gy. \tag{6}$$

The same control solves the θ-frame problem with $R = H = P$ and state noise $P^{-1}\Sigma P^{-1}$: $S = P^{1/2}GP^{1/2}$ satisfies $S(R^{-1} + \gamma P^{-1}\Sigma P^{-1})S = H$. ***(iv)*** *At $\gamma = 0$, equation 6 is AdamW exactly. For diagonal Σ it is the per coordinate gate $(1 + \gamma\sigma_i^2/\sqrt{\hat{v}_i})^{-1/2}$, i.e. equation 2 with $R = H$. If $H = \kappa_0 P$, G is unchanged and the step is multiplied by $\kappa_0^{-1/2}$.*

Part (i) is dense: its inner square root of a product of noncommuting matrices costs $O(d^3)$. Whitening by Adam's own denominator makes the curvature the identity, and the solution collapses to a single inverse square root of $I + \gamma\Sigma_z$, with no approximation. Only the shape of the curvature proxy matters, since a global constant is absorbed by η.

3.3 Kronecker Form and Noise Estimate

For a linear layer the per example gradient is $\delta a^\top$, so $\mathrm{vec_r}(\delta a^\top) = \delta \otimes a$, whose second moment factorizes as $\mathbb{E}[\delta\delta^\top] \otimes \mathbb{E}[aa^\top]$ when δ and a are independent (Martens & Grosse, 2015); this motivates (A7). With row major $\mathrm{vec_r}$, $(X \otimes W)\,\mathrm{vec_r}(Y) = \mathrm{vec_r}(XYW^\top)$, so B acts on rows and A on columns. Let $D \in \mathbb{R}^{n\times m}$ be the whitened noise sample of a layer and $A = \mathbb{E}[D^\top D]$, $B = \mathbb{E}[DD^\top]$ (EMAs with rate β_σ in practice).

Theorem 3 (Kronecker factorization). ***(i)*** *If $\mathrm{Cov}(\mathrm{vec_r}\, D) = B' \otimes A'$ and $\mathbb{E}D = 0$, then $A = \mathrm{tr}(B')A'$, $B = \mathrm{tr}(A')B'$, $\mathrm{tr}\, A = \mathrm{tr}\, B = \mathrm{tr}\,\Sigma_z$, and*

$$\frac{B \otimes A}{\mathrm{tr}\, A} = B' \otimes A' = \Sigma_z. \tag{7}$$

For any Σ_z, $\mathrm{tr}\left[(B \otimes A)/\,\mathrm{tr}\, A\right] = \mathbb{E}\|D\|_F^2 = \mathrm{tr}\,\Sigma_z$. ***(ii)*** *With $A = U_A\,\mathrm{diag}(\lambda)U_A^\top$, $B = U_B\,\mathrm{diag}(\mu)U_B^\top$ and $\tau = \mathrm{tr}\, A$,*

$$\begin{aligned}\Big(I + \gamma\tfrac{B\otimes A}{\tau}\Big)^{-1/2}\mathrm{vec_r}(Y) &= \mathrm{vec_r}(Z),\\ Z &= U_B\big[(U_B^\top Y U_A) \odot G\big]U_A^\top,\\ G_{ij} &= \Big(1 + \gamma\frac{\mu_i\lambda_j}{\tau}\Big)^{-1/2}.\end{aligned} \tag{8}$$

(iii) *Applying equation 8 costs $O(nm(n+m))$, the eigendecompositions $O(n^3 + m^3)$, and every one of the nm directions of the layer receives a gain.*

Algorithm 1 Evie-KF (one step; per parameter tensor unless marked global)

Require: $\eta, \lambda_{\text{wd}}, \gamma$; β_1=0.9, β_2=0.999, ϵ=10^{-8}, β_σ=0.95, warmup 20, refresh 20, $n_{\max}$=1024

1: $m \leftarrow \beta_1 m + (1-\beta_1) g$; $v \leftarrow \beta_2 v + (1-\beta_2) g^{\odot 2}$
2: $\hat{m}_1 \leftarrow m/(1-\beta_1^t)$; $\hat{v} \leftarrow v/(1-\beta_2^t)$
3: $p \leftarrow (\hat{v}^{\odot 1/2} + \epsilon)^{\odot -1/2}$; $y \leftarrow p \odot \hat{m}_1$; $d \leftarrow p \odot \delta$ ▷ $d = p \odot g$ for Evie-KFu
4: Two dimensional, $D = d$ as $n \times m$: $A \leftarrow \beta_\sigma A + (1-\beta_\sigma) D^\top D$
$B \leftarrow \beta_\sigma B + (1-\beta_\sigma) D D^\top$ ▷ diagonal if dimension $> n_{\max}$
5: One dimensional or `diag_only`: $s \leftarrow \beta_\sigma s + (1-\beta_\sigma) d^{\odot 2}$
$z \leftarrow y \odot (1 + \gamma s/\bar{s})^{\odot -1/2}$
6: Every 20 steps: eigendecompose bias corrected A, B in float64; $\tau \leftarrow \operatorname{tr} A$
7: $G_{ij} \leftarrow (1 + \gamma \mu_i \lambda_j/\tau)^{-1/2}$
8: $Z \leftarrow U_B[(U_B^\top Y U_A) \odot G] U_A^\top$ ▷ $Y = y$ reshaped
9: **global:** $c \leftarrow \|y\|/\|z\|$; $z \leftarrow c\, z$ ▷ after warmup only
10: $\theta \leftarrow \theta(1 - \eta\lambda_{\text{wd}}) - \eta\, p \odot z$

The eigenvalues $\mu_i \lambda_j/\tau$ sum to $\operatorname{tr} \Sigma_z$, an absolute noise energy with units of gradient. In the Kronecker branch, γ therefore controls the absolute coupling between noise and curvature (units $1/[g]$) and is tuned per task on a logarithmic grid. The vector branch normalizes by the layer mean and is independent of scale.

Proposition 4 (Split batch noise sample)**.** *Let per example gradients be i.i.d. with mean ∇L and covariance Σ_1, and split a minibatch of size b into two independent halves with mean gradients g_A, g_B. With $g = \frac{1}{2}(g_A + g_B)$ and $\delta = \frac{1}{2}(g_A - g_B)$: **(i)** g is the ordinary minibatch gradient; **(ii)** $\mathbb{E}\delta = 0$ and $\mathbb{E}[\delta\delta^\top] = \Sigma_1/b = \operatorname{Cov}(g)$; **(iii)** $\mathbb{E}[gg^\top] = \operatorname{Cov}(g) + \nabla L \nabla L^\top$, whereas $\delta\delta^\top$ needs no centring; **(iv)** $\operatorname{Cov}(g, \delta) = 0$; **(v)** the two half batch passes cost the same FLOPs as one full pass.*

So D is built from $P^{-1/2}\delta$: the centred noise of the gradient actually used, uncorrelated with it, with no held out data. K-FAC, Shampoo and SOAP (Martens & Grosse, 2015; Gupta et al., 2018; Vyas et al., 2025) build their factors from uncentred statistics as in (iii). Under clipping, δ is scaled by the same coefficient as g.

3.4 Algorithm

Algorithm 1 applies equation 6 with $\hat{\Sigma}_z = \bigoplus_\ell (B_\ell \otimes A_\ell)/\tau_\ell$ and one global rescale:

$$\Delta\theta_t = -\eta\, P_t^{-1/2}\, c_t \big(I + \gamma \hat{\Sigma}_{z,t}\big)^{-1/2} P_t^{-1/2} \hat{m}_{1,t}, \qquad c_t = \frac{\|y_t\|}{\|(I + \gamma\hat{\Sigma}_{z,t})^{-1/2} y_t\|}. \tag{9}$$

Lines 1 through 3 are AdamW's moments and whitening; lines 4 through 8 build and apply G; line 9 restores the AdamW step length; and line 10 applies the update. Warmup, refresh, the vector gate and the $n_{\max}$ fallback are implementation choices, and only γ is tuned, jointly with η.

3.5 Properties

Proposition 5 (Properties of the step)**.** *Let $\Sigma_z = V \operatorname{diag}(\ell_k) V^\top$ and $g_{\min} = (1 + \gamma\ell_{\max})^{-1/2}$. **(i) Contraction.** $G = V \operatorname{diag}\big((1 + \gamma\ell_k)^{-1/2}\big) V^\top$ has eigenvalues in $[g_{\min}, 1]$, equal to 1 exactly on zero noise directions, decreasing in γ, and $\approx (\gamma\ell_k)^{-1/2}$ for $\gamma\ell_k \gg 1$. **(ii) Direction change.** With $a = g_{\min}$ and $b = (1 + \gamma\ell_{\min})^{-1/2}$, $\cos\angle(y, Gy) \geq 2\sqrt{ab}/(a+b)$ for all $y \neq 0$, with equality at $y = \sqrt{b}\, v_{\max} + \sqrt{a}\, v_{\min}$. **(iii) Norm preservation.** $c \in [1, 1/g_{\min}]$ and $\|cGy\| = \|y\|$: the whitened step length equals AdamW's. The effective gains $c(1 + \gamma\ell_k)^{-1/2}$ can exceed 1, so after the rescale the step redistributes AdamW's step instead of shrinking it. **(iv) Nesting.** $\gamma = 0$, or the first 20 steps, gives AdamW exactly. `diag_only` gives the gate $(1 + \gamma s_i/\bar{s})^{-1/2}$, a member of the centred variance diagonal family that includes AdaBelief and SDProp (Zhuang et al., 2020; Ida et al., 2017). `centered=False` replaces Σ by $\mathbb{E}[gg^\top]$ (Evie-KFu). The Kronecker, centred case is Evie-KF.*

By (ii) and (iii) the step differs from AdamW's by a bounded rotation at equal whitened length; without c, the derived radial contraction also remains. Appendix Result R8 gives the corresponding

help/hurt heuristic. By (iv), AdamW $\to$ diagonal $\to$ Kronecker is one path along the shape allowed for Σ.

Theorem 6 (Convergence of the idealized update)**.** *Let L be L_s smooth and bounded below by $L_{\inf}$, and let $\theta_{t+1} = \theta_t - \eta\, T_t \tilde{g}_t$ with $T_t = P_t^{-1/2} G_t P_t^{-1/2}$, where P_t, G_t are measurable with respect to the history $\mathcal{F}_t$ before step t, $\mathbb{E}[\tilde{g}_t \mid \mathcal{F}_t] = \nabla L(\theta_t)$, $\mathbb{E}[\|\tilde{g}_t - \nabla L(\theta_t)\|^2 \mid \mathcal{F}_t] \le \sigma^2$, $\hat{v}_t \le G_\infty^2$ in every coordinate and $\ell_{\max}(\hat{\Sigma}_{z,t}) \le \bar{\Lambda}$. With $c_{\mathrm{lo}} = (1+\gamma\bar{\Lambda})^{-1/2}(G_\infty+\epsilon)^{-1}$, $c_{\mathrm{hi}} = \epsilon^{-1}$ and $\eta \le c_{\mathrm{lo}}/(L_s c_{\mathrm{hi}}^2)$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla L(\theta_t)\|^2 \;\le\; \frac{2\,(L(\theta_0) - L_{\inf})}{\eta\, c_{\mathrm{lo}}\, T} + \frac{\eta\, L_s\, c_{\mathrm{hi}}^2}{c_{\mathrm{lo}}}\,\sigma^2. \tag{10}$$

The deterministic term decays as $O(1/T)$ and the noise term sets the stationary neighbourhood. The constants come from the spectrum of T_t, bounded by Proposition 5(i) and $P_t \in [\epsilon, G_\infty + \epsilon]$. If the half batch gradients satisfy $\|g_A\|, \|g_B\| \le G_h$, then $\bar{\Lambda} = G_h^2/\epsilon$ is admissible, since $\ell_{\max} \le \operatorname{tr} \hat{\Sigma}_z = \operatorname{tr} B \le G_h^2/\epsilon$. The theorem guarantees the idealized core update, with statistics from earlier steps and an unbiased unclipped oracle. An $\mathcal{F}_t$-measurable rescale $c_t \in [1, 1/g_{\min}]$ is covered by setting $c_{\mathrm{hi}} = (\epsilon g_{\min})^{-1}$; Algorithm 1 additionally uses momentum, clipping and a same step rescale.

Cost. Evie-KF uses structured state per $n \times m$ tensor: A, B and their eigenbases require about $2(n^2+m^2)$ floats, compared with $2nm$ for AdamW. Each step is $O(nm(n+m))$, each $O(n^3+m^3)$ refresh is amortized over 20 steps, and the split backward pass processes the same examples as one full pass.

4 Experimental Setup

Tasks and models. We evaluate on twelve tasks in three domains. Vision uses CIFAR-10, CIFAR-100, SVHN and STL-10 on one 11.2M-parameter CIFAR-stem ResNet-18 at batch 128, scoring test top-1 accuracy; we resize STL-10 to 32×32 so the architecture stays fixed while its 5,000-image low-data setting is retained. Language uses WikiText-2, WikiText-103, Penn Treebank and character-level enwik8 on one 8.9M-parameter 6-layer decoder-only Transformer at batch 32, scoring test perplexity; whitespace tokenisation with a 16k-word cap makes these perplexities comparisons between optimizers on the same task. Finance uses four committed sector panels of 49 to 50 tickers, one 2,497-parameter MLP per ticker at batch 32, scoring test MSE on the next-day log return. Appendix H gives preprocessing and splits.

Baselines. We compare Evie-KF against AdamW, AdaBelief, Nesterov SGD, Sophia, Muon, Lion and Shampoo, each with its own tuned learning rate.

Protocol. Every cell runs 10,000 steps with cosine decay to one tenth of the base rate, decoupled weight decay 10^{-4}, gradient clipping at 5.0, float32, five seeds (42 to 46), and best-validation checkpoint selection before a single evaluation on the untouched test split. The budget is fixed by steps, not by epochs or wall-clock time. On models with BatchNorm every arm uses the same split-batch forward and backward pass that Evie-KF requires, so the normalisation statistics are identical across arms and only the preconditioner differs.

Hyperparameter tuning. We search a 7-point half-decade learning-rate grid for 2,500 steps on seed 42 and score the best validation checkpoint, not the final step; an edge winner extends that axis twice more, at most three times, and the choice is then frozen for the five-seed run. Evie-KF's broad γ surfaces (Section 6) support a baseline-budget alternative: fixing γ to one default and searching learning rate alone costs 0.9% at the median across four controls. The main runs use the optional joint (lr, γ) search and Shampoo searches (lr, β), giving 49 to 99 points per task against 7 for single-axis arms; Appendix G.3 reports every fixed-γ outcome.

Implementation and compute. Optimizer state is float32; symmetrised Kronecker factors are eigendecomposed in float64 every 20 steps, with a diagonal fallback above dimension 1024. Overhead falls from $1.5\times$ to $1.7\times$ AdamW at 9M to 11M parameters to $0.98\times$ at 33.7M; peak memory is below AdamW's on three of four language tasks. Runs use one NVIDIA Tesla T4 and PyTorch 2.10.0+cu128. Appendix H gives grids, fixed settings, divergence rules and provenance.

Table 1: Primary metric per task, median over five seeds; best per row in bold. Evie-KF contains AdamW at $\gamma = 0$ and, through its `diag_only` branch, the centred-variance diagonal family that AdaBelief belongs to. The two scale-generalization tasks ran three seeds and the arms shown; Muon's OpenWebText median is over the two seeds that finished.

	Evie-KF	AdamW	AdaBelief	SGD	Sophia	Muon	Lion	Shampoo
Vision — test top-1 accuracy (%), higher is better								
CIFAR-10	*81.94*	81.55	81.00	81.71	81.52	**85.86**	81.28	81.90
CIFAR-100	*52.28*	49.51	49.45	**54.27**	47.39	54.24	49.59	50.80
SVHN	*94.63*	94.66	94.75	94.73	94.89	**95.52**	94.55	94.83
STL-10	*48.59*	**51.29**	50.84	47.75	47.48	48.85	49.96	51.25
Language — test perplexity, lower is better								
WikiText-2	*100.46*	105.70	97.15	103.80	123.75	**83.62**	106.65	123.91
WikiText-103	*35.41*	39.99	41.17	59.75	39.51	**33.49**	47.31	36.25
Penn Treebank	*192.17*	179.35	179.77	180.35	201.49	187.01	201.94	**172.94**
enwik8	*3.04*	3.18	3.15	3.60	3.34	**2.97**	3.11	**2.97**
Finance — test MSE $\times 10^{-4}$, lower is better								
Tech	*4.1497*	4.3336	4.3745	4.1537	4.5594	**4.1475**	4.1502	4.2170
Finance	*2.2669*	2.2643	2.2583	2.2278	2.6356	3.7243	**2.2269**	3.3812
Healthcare	*2.9150*	3.0969	3.1291	**2.9077**	3.1868	3.2588	3.1281	3.1132
Consumer/Ind.	*1.9687*	2.0169	2.1041	**1.8491**	2.0623	1.9124	2.1161	2.0574
Scale generalization — outside the twelve tasks								
Tiny-ImageNet, top-1 (%) ↑	*37.54*	36.08	36.22	—	—	**41.88**	—	—
OpenWebText, perplexity ↓	*72.79*	83.74	83.30	—	—	**55.73**	—	—

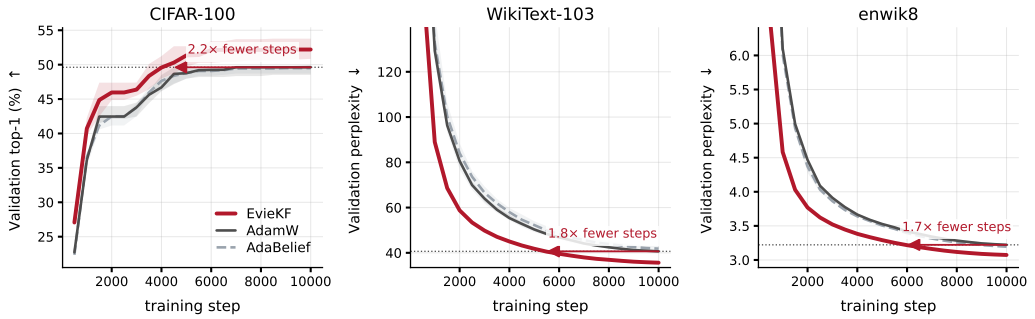


Figure 5: Running-best validation metric over five seeds, with median and range. The dotted line is AdamW's final value; arrows mark Evie-KF's first match.

5 Results

Overall standing. Evie-KF ranks second of eight optimizers by mean rank (3.67), behind only Muon (3.00), and ties Muon for the most top-three finishes (8/12 tasks).

Twelve tasks. Table 1 reports all eight optimizers. Evie-KF, with Shampoo, has a positive median against AdamW in every domain. Its largest margins, $+11.5\%$ on WikiText-103 and $+5.6\%$ on CIFAR-100, hold on 5/5 seeds; on WikiText-103 it places second to Muon and beats SGD by 40.7% and AdaBelief by 14.0%, again on 5/5. It completes 99.8% of 1,025 cells; its 0.20% divergence rate is the best among arms with any divergence.

Training dynamics. Evie-KF separates from both diagonal arms within 2,000 steps and first reaches AdamW's final validation score at 45%, 55% and 60% of the budget on CIFAR-100, WikiText-103 and enwik8, or $1.7\times$ to $2.2\times$ fewer steps (Figure 5).

Confirmatory testing. Paired two-sided Wilcoxon tests across 12 tasks, with Holm correction over seven comparisons, find Evie-KF over Sophia significant ($p = 0.0068$); Lion (10/12, raw $p = 0.021$) and AdamW (8/12, median $+3.4\%$) remain statistically on par with Evie-KF after correction. With AdamW as reference, the corrected tests likewise place all arms statistically on par within this benchmark's resolution.

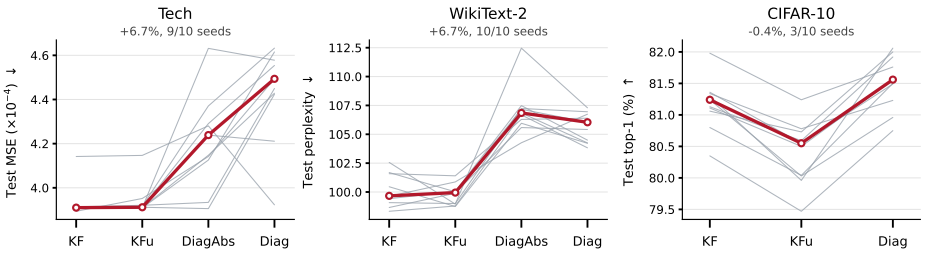


Figure 6: Ten-seed ablations; CIFAR-10's selected operator stays close to AdamW.

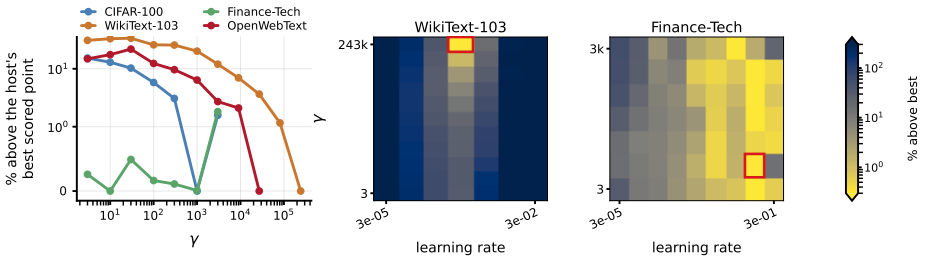


Figure 7: Risk sensitivity and joint surfaces; outlines mark frozen choices (Appendix G).

Scale generalization. On 33.7M-parameter OpenWebText, Evie-KF reaches 72.79 against AdamW's 83.74 perplexity (+13.08%, 3/3 seeds) at 0.98× its wall time and 1.01× its peak memory, the largest margin over AdamW anywhere in this work and on its largest model. Muon, which searched its own learning rate on the same grid and ran at 1.02× AdamW's wall clock, reaches 55.73 on the two seeds that completed and leads the task. On Tiny-ImageNet-200 the ordering repeats: Evie-KF is ahead of both Adam-family arms on three of three seeds, with Muon ahead of all three. Both are three-seed checks outside the twelve-task test.

6 Ablation Studies

Each ablation independently tunes (lr, γ) over 10 seeds (42 to 51).

Kronecker structure versus a diagonal gate. Relative to Evie-DiagAbs, Evie-KF gains 6.75% on finance-tech (9/10, Holm $p = 0.0195$) and 6.67% on WikiText-2 (10/10, Holm $p = 0.0156$), agreeing within 0.1 points and clearing correction across eight contrasts (Table 8). Normalisation improves finance-tech by +6.23% while preserving WikiText-2 within 0.98%, identifying it as a task-dependent design choice. Diagnostic CIFAR-10 confirms that $\gamma = 1$ preserves AdamW closely ($\cos(\Delta_{\text{Evie-KF}}, \Delta_{\text{AdamW}}) = 0.991$).

Centred versus uncentred noise. Against Evie-KFu, Evie-KF changes finance-tech by 0.03% (6/10, $p = 0.232$) and WikiText-2 by 0.27% (5/10, $p = 0.62$), with comparable CIFAR-10 performance. Split-batch centring therefore preserves performance at no extra FLOPs while the Kronecker structure supplies the measured gains.

Sensitivity to the risk parameter. Across 774 points, fixing $\gamma = 3000$ at the baseline budget costs 0.9% at the median across four controls. Mean-eigenvalue normalisation cuts the selected spread from 2,700× to 3.3× on three hosts, with $\gamma = 30$ within 0.33% of each optimum (Appendix G). These are baseline-budget controls, not a twelve-task claim of one universal γ.

7 Conclusion

Evie-KF applies $(I+\gamma\Sigma_z)^{-1/2}$ in Adam-whitened coordinates, nesting AdamW and diagonal gates. Against matched diagonal gates it gains 6.75% on finance-tech and 6.67% on WikiText-2, both surviving correction. Across twelve tasks it ranks second of eight, reaches AdamW's final score in 45 to 60% of the budget, and improves on AdamW by 13.1% at 0.98× its wall time on 33.7M-parameter OpenWebText.

AI USE DISCLOSURE

In this work, we used generative AI tools for software coding, debugging, code refinement, manuscript refining, and creating and refining visualizations and graphs. We also used generative AI tools to check mathematical derivations and proofs developed by the authors. We have reviewed all AI assisted work manually and verified the code, mathematical checks, manuscript content, and visualizations. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work evaluates an optimization method and does not involve human-subject experiments, interventions, or the collection of personally identifiable information. The vision and language benchmarks and historical market series are used only to compare optimization behaviour. The finance experiments are methodological and should not be interpreted as investment advice or evidence of deployable trading performance. More efficient optimization may reduce the compute and energy required to reach a target score, but it may also lower the cost of training models with downstream risks; we therefore report wall time, memory use, divergences, and negative results, and do not present accuracy alone as an efficiency benefit. We are not aware of additional ethical concerns specific to this work beyond the general risks associated with large-scale machine learning.

REPRODUCIBILITY STATEMENT

We release the full harness, every task file and the statistical pipeline at `https://github.com/rpextra2026-afk/evie-kf-optim`, and the optimizer as a package at `https://pypi.org/project/evie-kf/`. Both are published from accounts created for this submission and carry no author information. Appendix H describes what is in the release, the test suite that gates it, and every frozen hyperparameter.

REFERENCES

Naman Agarwal, Brian Bullins, Xinyi Chen, Elad Hazan, Karan Singh, Cyril Zhang, and Yi Zhang. Efficient full-matrix adaptive regularization. In *International Conference on Machine Learning*, pp. 102 to 110, 2019.

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

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

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *International Conference on Machine Learning*, 2018.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Bugra Can and Mert Gürbüzbalaban. Entropic risk-averse generalized momentum methods. *Optimization Methods and Software*, 40(6), 2025.

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

Dami Choi, Christopher J. Shallue, Zachary Nado, Jaehoon Lee, Chris J. Maddison, and George E. Dahl. On empirical comparisons of optimizers for deep learning. *arXiv preprint arXiv:1910.05446*, 2019.

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.

Aram Davtyan, Sepehr Sameni, Llukman Cerkezi, Givi Meishvili, Adam Bielski, and Paolo Favaro. KOALA: A Kalman optimization algorithm with loss adaptivity. In *AAAI Conference on Artificial Intelligence*, pp. 6471 to 6479, 2022.

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

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 *Advances in Neural Information Processing Systems*, 2025.

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

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

GLM-130B Team. GLM-130b training log. `https://github.com/zai-org/GLM-130B/blob/main/logs/main-log-en.md`, 2022a. Accessed September 26, 2026.

GLM-130B Team. GLM-130b project page. `https://keg.cs.tsinghua.edu.cn/glm-130b/posts/glm-130b/`, 2022b. Accessed September 26, 2026.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Mert Gürbüzbalaban, Yasa Syed, and Necdet Serhat Aybat. Accelerated gradient methods with biased gradient estimates: Risk sensitivity, high-probability guarantees, and large deviation bounds. *arXiv preprint arXiv:2509.13628*, 2025.

Yasutoshi Ida, Yasuhiro Fujiwara, and Sotetsu Iwamura. Adaptive learning rate via covariance matrix based preconditioning for deep neural networks. In *International Joint Conference on Artificial Intelligence*, 2017.

David H. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic Control*, 18 (2):124 to 131, 1973. doi: 10.1109/TAC.1973.1100265.

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

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 *International Conference on Learning Representations*, 2025.

Mikail Khona, Aditya Vavre, Boxiang Wang, Deyu Fu, Hao Wu, Mike Chrzanowski, Bryan Catanzaro, Dheevatsa Mudigere, Jeff Pool, Michael Lightstone, Mohammad Shoeybi, Mostofa Patwary, Nima Tajbakhsh, and Tijmen Blankevoort. SOAP, Muon, and beyond: Pushing LLM pretraining scales. *arXiv preprint arXiv:2607.20548*, 2026.

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

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

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on transformers, but sign descent might be. In *International Conference on Learning Representations*, 2023.

Nicolas Le Roux and Andrew W. Fitzgibbon. A fast natural Newton method. In *International Conference on Machine Learning*, pp. 623 to 630, 2010.

Nicolas Le Roux, Pierre-Antoine Manzagol, and Yoshua Bengio. Topmoumoute online natural gradient algorithm. In *Advances in Neural Information Processing Systems*, volume 20, pp. 849 to 856, 2008.

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

Wu Lin, Scott C. Lowe, Felix Dangel, Runa Eschenhagen, Zikun Xu, and Roger B. Grosse. Understanding and improving Shampoo and SOAP via Kullback-Leibler minimization. In *International Conference on Learning Representations*, 2026.

Bing Liu, Wenjie Zhou, and Chengcheng Zhao. Rethinking Bregman divergences in Kronecker-factored optimizers. *arXiv preprint arXiv:2606.00542*, 2026a.

Bingbin Liu, Rachit Bansal, Depen Morwani, Nikhil Vyas, David Alvarez-Melis, and Sham M. Kakade. Adam or Gauss-Newton? A comparative study in terms of basis alignment and SGD noise. In *International Conference on Machine Learning*, 2026b.

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

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

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

Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(134):1 to 35, 2017.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1 to 76, 2020.

James Martens and Roger Grosse. Optimizing neural networks with Kronecker-factored approximate curvature. In *International Conference on Machine Learning*, 2015.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018.

Meta AI. Llama 3.1 model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md`, 2024.

Igor Molybog, Peter Albert, Moya Chen, Zachary DeVito, David Esiobu, Naman Goyal, Punit Singh Koura, Sharan Narang, Andrew Poulton, Ruan Silva, Binh Tang, Diana Liskovich, Puxin Xu, Yuchen Zhang, Melanie Kambadur, Stephen Roller, and Susan Zhang. A theory on Adam instability in large-scale machine learning. *arXiv preprint arXiv:2304.09871*, 2023.

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

Nikhil Nayak, Julia White, Urchade Zaratiana, Kelton Zhang, Henrijs Princis, Dhruv Atreja, Henry Fawcett, Matthew Thomas, George Hurn-Maloney, and Ash Lewis. Correcting stochastic update bias in preconditioned language model optimizers. *arXiv preprint arXiv:2605.20756*, 2026.

Son Nguyen, Bo Liu, Lizhang Chen, and Qiang Liu. Improving adaptive moment optimization via preconditioner diagonalization. In *International Conference on Artificial Intelligence and Statistics*, pp. 4312 to 4320, 2026.

Thanh Huy Nguyen, Umut Şimşekli, Mert Gürbüzbalaban, and Gaël Richard. First exit time analysis of stochastic gradient descent under heavy-tailed gradient noise. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yann Ollivier. Online natural gradient as a Kalman filter. *Electronic Journal of Statistics*, 12(2): 2930 to 2961, 2018.

Tom Schaul, Sixin Zhang, and Yann LeCun. No more pesky learning rates. In *International Conference on Machine Learning*, 2013.

Robin M. Schmidt, Frank Schneider, and Philipp Hennig. Descending through a crowded valley: Benchmarking deep learning optimizers. In *International Conference on Machine Learning*, 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.

Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban. A tail-index analysis of stochastic gradient noise in deep neural networks. In *International Conference on Machine Learning*, pp. 5827 to 5837, 2019.

Valentin Thomas, Fabian Pedregosa, Bart van 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*, 2020.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5: RMSProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham M. Kakade. SOAP: Improving and stabilizing Shampoo using Adam for language modeling. In *International Conference on Learning Representations*, 2025.

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. *arXiv preprint arXiv:2509.02046*, 2025.

Peter Whittle. Risk-sensitive linear/quadratic/Gaussian control. *Advances in Applied Probability*, 13(4):764 to 777, 1981. doi: 10.1017/S0001867800036508.

Zixuan Xia, Aram Davtyan, and Paolo Favaro. KOALA++: Efficient Kalman-based optimization of neural networks with gradient-covariance products. *arXiv preprint arXiv:2506.04432*, 2025.

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

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. GLM-130b: An open bilingual pre-trained model. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=-Aw0rrrPUF`.

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean P. Foster, and Sham M. Kakade. How does critical batch size scale in pre-training? In *International Conference on Learning Representations*, 2025.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham M. Kakade. Deconstructing what makes a good optimizer for autoregressive language models. In *International Conference on Learning Representations*, 2025.

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*, 2019.

Juntang Zhuang, Tommy Tang, Yifan Ding, Sekhar C. Tatikonda, Nicha Dvornek, Xenophon Papademetris, and James Duncan. AdaBelief optimizer: Adapting stepsizes by the belief in observed gradients. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18795 to 18806, 2020.

A Notation, Assumptions and Units

$\odot$ and $(\cdot)^{\odot a}$ are elementwise. For symmetric $X = U\,\mathrm{diag}(x)U^\top$, $f(X) = U\,\mathrm{diag}(f(x))U^\top$; $X^{1/2}$ is the unique positive semidefinite square root. $\mathrm{vec_r}$ flattens an $n \times m$ matrix row by row, entry (i,j) going to position $m(i-1)+j$, so $(X \otimes W)\,\mathrm{vec_r}(Y) = \mathrm{vec_r}(XYW^\top)$ and $(B' \otimes A')_{(i,j),(k,l)} = B'_{ik}A'_{jl}$. $P = \mathrm{diag}(\hat{v}^{\odot 1/2} + \epsilon)$, $z = P^{1/2}\theta$, $\Sigma = \mathrm{Cov}(\hat{g})$, $\Sigma_z = P^{-1/2}\Sigma P^{-1/2}$, $y = P^{-1/2}\hat{m}_1$, $G = (I + \gamma\Sigma_z)^{-1/2}$. Jacobson's Riccati equation, with process noise covariance N and risk parameter σ:

$$-\dot{S} = Q + SA + A^\top S - S\big(BR^{-1}B^\top - \sigma N\big)S. \tag{S1}$$

Assumptions (as in the main paper). (A1) $L(\theta) \approx L(\theta^*) + \frac{1}{2}x^\top Hx$, $x = \theta - \theta^*$, $H \succ 0$. (A2) In z: $\dot{z} = u_z$, $A = 0$, $B = I$, state and control weights I; equivalently $Q = R = H$ in θ. (A3) $\xi \sim \mathcal{N}(0, \Sigma)$, full Σ; state noise $H^{-1}\Sigma H^{-1}$, i.e. Σ_z in z, identified with N. (A4) $\sigma = -\gamma$, $\gamma \geq 0$. (A5) $x \approx H^{-1}\nabla L$ with $\hat{m}_1$ for ∇L; $x_z \approx y$. (A6) $H \approx P$ up to a global scale. (A7) $\Sigma_z \approx (B \otimes A)/\tau$ per tensor, $\tau = \mathrm{tr}\,A$, block diagonal across tensors, diagonal fallback above $n_{\max}$. (A8) P, Σ_z constant over one decision (algebraic Riccati equation).

Units. With $[\theta]$ and $[g]$ the units of parameters and gradients:

H	$[g]/[\theta]$	Σ	$[g]^2$
$\hat{m}_1,\ \hat{v}$	$[g],\ [g]^2$	$\Sigma_z,\ A,\ B,\ \tau$	$[g]$
P	$[g]$	γ (Kronecker branch)	$1/[g]$
$G,\ c$	dimensionless	γ (vector branch, normalized by the mean)	dimensionless

Because $[P] = [g] \neq [H]$, (A6) can hold only up to a global constant with units $1/[\theta]$; Result R4 shows that such a constant is absorbed by the learning rate.

B The Sign of γ and Risk Aversion

Reproducible convention. Sign conventions differ across the risk-sensitive-control literature, so we state the mapping rather than referring to a "corrected sign." Whittle's criterion (5) describes $-\theta$ as the measure of risk aversion (Whittle, 1981); below we use $\vartheta = -\theta$, so $\vartheta > 0$ is risk averse. In the Jacobson-form Riccati equation equation S1, direct substitution of the exponential criterion gives $\sigma = \vartheta$. Consequently, the Evie-KF choice $\sigma = -\gamma$ with $\gamma \geq 0$ means $\sigma = -\gamma \leq 0$ and $\vartheta = -\gamma \leq 0$. It is therefore the optimistic branch of this *process-noise* model, not its risk-averse branch. The risk-averse interpretation of positive γ used in the paper comes from the update-noise model in Section B.2. Thus the reproducible references are Whittle's criterion (5), the parameter mapping above and the HJB calculation below. We do not attribute an algebraic error to Jacobson or Whittle; the correction is to keeping their sign conventions distinct when interpreting the Evie-KF branch.

Model	Parameter sign	Algebraic equation	Effect
Process noise, pessimistic	$\vartheta = \sigma > 0$	$S(I - \vartheta\Sigma_z)S = I$	gain rises; finite validity range
Process noise, Evie branch	$\sigma = -\gamma \leq 0$	$S(I + \gamma\Sigma_z)S = I$	globally defined contraction
Update noise	$\gamma > 0$	Eq. equation S4	cost variance penalized; step shrinks locally

B.1 Process Noise Model

Take whitened dynamics $dx = u\,ds + \Sigma_z^{1/2} dW_s$, running cost $c = \frac{1}{2}(\|x\|^2 + \|u\|^2)$ and $J_\vartheta = \vartheta^{-1} \log \mathbb{E} \exp(\vartheta \int c\,ds)$. For $\vartheta > 0$, Jensen gives $J_\vartheta \geq \mathbb{E} \int c$, and $J_\vartheta = \mathbb{E}C + \frac{\vartheta}{2}\mathrm{Var}C + O(\vartheta^2)$, so $\vartheta > 0$ is the pessimistic (variance penalizing) branch. With $W = \mathbb{E}e^{\vartheta C}$ and $V = \vartheta^{-1} \log W$, $\nabla W = \vartheta W \nabla V$ and $\nabla^2 W = \vartheta W(\nabla^2 V + \vartheta \nabla V \nabla V^\top)$, so

$$-\partial_s V = \min_u \Big\{ c + \nabla V^\top u + \tfrac{1}{2} \operatorname{tr}(\Sigma_z \nabla^2 V) + \tfrac{\vartheta}{2} \nabla V^\top \Sigma_z \nabla V \Big\}. \tag{S2}$$

With $V = \frac{1}{2} x^\top S x + \kappa(s)$, $u^* = -Sx$ and, at stationarity, $S(I - \vartheta\Sigma_z)S = I$, so $S = (I - \vartheta\Sigma_z)^{-1/2}$, which exists only for $\vartheta \lambda_{\max}(\Sigma_z) < 1$ when $\vartheta > 0$ (the breakdown point). Equation equation S1 has the same quadratic term when $\sigma = \vartheta$. The branch (A4) instead sets $\vartheta = \sigma = -\gamma$, giving $G = (I + \gamma\Sigma_z)^{-1/2}$. Numerical check of the sign: iterating the discrete time exponential quadratic recursion ($h = 2$, $\sigma^2 = 0.3$, $\vartheta = 1$, $\Delta s = 10^{-4}$) converges to 1.6904, against $\sqrt{h/(1 - \vartheta\sigma^2)} = 1.6903$ and $\sqrt{h/(1 + \vartheta\sigma^2)} = 1.2403$.

B.2 Update Noise Model

Per coordinate and for one step, let $L(x) = \frac{1}{2} h x^2$, $g = hx$, $\hat{g} = g + \xi$ with $\xi \sim \mathcal{N}(0, s)$, and $x^+ = x - \lambda \hat{g} = a - \lambda\xi$ with $a = x(1 - \lambda h)$. The objective is

$$J(\lambda) = \tfrac{1}{\gamma} \log \mathbb{E} \exp\big(\gamma \cdot \tfrac{1}{2} h (x^+)^2\big) = \mathbb{E}C + \tfrac{\gamma}{2}\mathrm{Var}C + O(\gamma^2), \tag{S3}$$

so $\gamma > 0$ penalizes cost variance and $\gamma \to 0$ is risk neutral.

Closed form. The exponent is $-\frac{D}{2s}\xi^2 - \gamma h a \lambda \xi + \frac{\gamma h a^2}{2}$ with $D := 1 - \gamma h \lambda^2 s$. For $D > 0$ the Gaussian integral gives $\mathbb{E}[\cdot] = D^{-1/2} \exp(\gamma h a^2/(2D))$, using $D + \gamma h \lambda^2 s = 1$, hence

$$J(\lambda) = \frac{h a^2}{2D} - \frac{1}{2\gamma} \log D, \tag{S4}$$

which agrees with a Monte Carlo estimate based on 4×10^6 samples. The constraint $D > 0$ appears by itself: a step so large that noise amplification makes the exponential moment infinite.

Stationarity is a cubic. With $a = g(1 - \lambda h)/h$ and $D' = -2\gamma h \lambda s$, $J'(\lambda) = 0$ becomes

$$\gamma h^2 s^2 \lambda^3 + g^2 \gamma h s\, \lambda^2 - \big(g^2 \gamma s + g^2 h + h s\big)\lambda + g^2 = 0, \tag{S5}$$

whose root in $(0, 1/\sqrt{\gamma h s})$ equals the numerical argmin of J.

Properties. (i) $\gamma \to 0$: $\lambda_0 = g^2/(h(g^2 + s)) = \rho/h$, $\rho = g^2/(g^2 + s)$, the step of Schaul et al. (2013); this limit is not new. (ii) $\frac{d\lambda^*}{d\gamma}\big|_0 = -g^4 s^3/(h^2(g^2 + s)^4) < 0$. (iii) $\lambda^*(\gamma)$ decreases in γ; we verified this numerically over six decades of γ but do not prove it globally. (iv) $\lambda^* < 1/\sqrt{\gamma h s}$ and $\lambda^* \to 1/(\gamma s)$ as $\gamma \to \infty$. (v) With $\lambda = \mu\lambda_0$ and $\kappa = \gamma(g^2 + s)/h$, $\kappa(1 - \rho)^2 \rho^2 \mu^3 + \kappa\rho^2(1 - \rho)\mu^2 - \kappa\rho(1 - \rho)\mu - \mu + 1 = 0$, so the gate μ depends only on (ρ, κ). This one step model is not the derivation of G; it fixes the meaning of γ (variance aversion toward update noise) and agrees with G on monotone shrinkage.

C Proofs of the Main Results

The derivation is organized as ten results, R1 through R10. The main text states them in condensed form:

Here	Content	Main text
R1	Scalar risk sensitive gain	Intuition, eq. for $(1+\gamma Rq)^{-1/2}$
R2	Gates from Adam's own moments	Theorem 1
R3	Matrix Riccati closed form	Theorem 2(i)
R4	Whitened solution	Theorem 2(ii) through (iv)
R5	Spectrum and angle bound	Proposition 5(i) and (ii)
R6	Kronecker factorization	Theorem 3
R7	Split batch noise sample	Proposition 4
R8	Norm preserving rescale	Proposition 5(iii)
R9	Nesting	Proposition 5(iv)
R10	Convergence of the idealized update	Theorem 6

C.1 R1: Scalar Gain (Main Text: Intuition)

R1. *With $A = 0$, $B = 1$, $Q = H$, $N = q$ and (A4), (A8), equation S1 has the positive steady state $S = \sqrt{HR/(1+\gamma Rq)}$; with (A5), $u = -g/(\sqrt{HR}\sqrt{1+\gamma Rq})$ and $|u|/|u_0| = (1+\gamma Rq)^{-1/2} \in (0,1]$ for every $\gamma \geq 0$.*

Proof. Substitution gives $-\dot{S} = H - S^2(R^{-1} + \gamma q)$. At $\dot{S} = 0$, $S^2 = H/(R^{-1} + \gamma q) = HR/(1+\gamma Rq)$; the positive root is taken because $V = \frac{1}{2}Sx^2$ is a cost to go. Since $R^{-1} + \gamma q > 0$ for all $\gamma \geq 0$, no existence condition arises. The control is $u = -R^{-1}Sx$ and, with $x \approx g/H$, $u = -R^{-1}\sqrt{HR/(1+\gamma Rq)}\, g/H = -g/(\sqrt{HR}\sqrt{1+\gamma Rq})$; dividing by $u_0 = -g/\sqrt{HR}$ gives the gate. □

C.2 R2: Coordinatewise Gates (Main Text: Theorem 1)

R2. *With $\epsilon = 0$, let each nonnegative gate entry $w_i = f_i((\hat{v}_i - \hat{m}_{1,i}^2)_+, \hat{v}_i)$ depend only on coordinate i's Adam moments. Then $\Delta\theta_i = -\eta\,\mathrm{sgn}(\hat{m}_{1,i})\sqrt{r_i}\,F_i(r_i, \hat{v}_i)$, $r_i = \hat{m}_{1,i}^2/\hat{v}_i$, and $F_i(r_i, \hat{v}_i) = f_i(\hat{v}_i(1-r_i)_+, \hat{v}_i)$. The update is unchanged by any change to off-diagonal gradient covariance that preserves the marginal first and second moments.*

Proof. For every i, $|\Delta\theta_i^{\mathrm{AdamW}}|/\eta = |\hat{m}_{1,i}|/\sqrt{\hat{v}_i} = \sqrt{r_i}$ and $\hat{\sigma}_i^2 = (\hat{v}_i - \hat{m}_{1,i}^2)_+ = \hat{v}_i(1-r_i)_+$; substitute into f_i. Because each resulting entry is a function only of $(\hat{m}_{1,i}, \hat{v}_i)$, changing $\mathrm{Cov}(g_i, g_j)$ for $i \neq j$ while preserving these marginals cannot affect the update. Nonnegativity preserves the sign pattern and the diagonal form cannot mix coordinates. It can rotate the full vector only in the limited sense of unequal axis-aligned rescaling, not by aligning with an off-axis covariance eigenvector. □

With $\beta_1 \neq \beta_2$, r can exceed 1 after a gradient spike (we observed $r = 1.84$), hence the clamp. *Units.* A γ multiplying a raw variance carries units $[g]^{-2}$: rescaling the loss by c maps $\hat{\sigma}^2 \mapsto c^2\hat{\sigma}^2$ with r fixed, and at $\gamma = 1$, $\hat{\sigma}^2 = 0.5$ the gate $(1+\gamma\hat{\sigma}^2)^{-1/2}$ takes the values $0.998/0.816/0.140$ for $c = 0.1/1/10$. A gate that depends on r alone is independent of scale but, by R2, only reshapes Adam's curve.

C.3 R3: Matrix Riccati Equation (Main Text: Theorem 2(i))

R3. *With $M = I + \gamma\Sigma$, the equation $SMS = H$ ($H \succ 0$) has the unique positive semidefinite solution $S = M^{-1/2}(M^{1/2}HM^{1/2})^{1/2}M^{-1/2}$ for every $\gamma \geq 0$.*

Proof. Substituting $A = 0$, $B = R = I$, $Q = H$, $N = \Sigma$, $\sigma = -\gamma$ in equation S1 and setting $\dot{S} = 0$ gives $SMS = H$. *Existence.* Let $K = M^{1/2}HM^{1/2} \succ 0$ and $S = M^{-1/2}K^{1/2}M^{-1/2}$; then $SMS = M^{-1/2}K^{1/2}K^{1/2}M^{-1/2} = M^{-1/2}KM^{-1/2} = H$. *Uniqueness.* For any solution $S \succeq 0$, $X = M^{1/2}SM^{1/2} \succeq 0$ satisfies $X^2 = M^{1/2}SMSM^{1/2} = K$, so $X = K^{1/2}$. Since $M \succeq I$, the construction is valid for all $\gamma \geq 0$. □

C.4 R4: Whitened Solution (Main Text: Theorem 2(ii) through (iv))

R4. *Under (A2), (A3), (A5), (A6), (A8): $H_z = I$, $S_z(I + \gamma\Sigma_z)S_z = I$, $S_z = G$, and $\Delta\theta = -\eta P^{-1/2} G P^{-1/2}\hat{m}_1$. The same control solves the problem in θ coordinates with $R = H = P$ and state noise $P^{-1}\Sigma P^{-1}$. At $\gamma = 0$ the step is AdamW; for diagonal Σ the gate is $(1+\gamma\sigma_i^2/\sqrt{\hat{v}_i})^{-1/2}$.*

Proof. (a) With $z = P^{1/2}\theta$ and $x_z = P^{1/2}x$, the chain rule gives $\nabla_z L = P^{-1/2}\nabla_\theta L$, $H_z = P^{-1/2}HP^{-1/2}$, $\xi_z = P^{-1/2}\xi$, $\Sigma_z = P^{-1/2}\Sigma P^{-1/2}$, and $\frac{1}{2}x_z^\top H_z x_z = \frac{1}{2}x^\top Hx$. (b) With $H = P$, $H_z = P^{-1/2}PP^{-1/2} = I$. (With $H = P^2$ one would get $H_z = P$.) (c) By (A2) and (A3), $-\dot{S}_z = I - S_z(I + \gamma\Sigma_z)S_z$; at steady state $S_z M_z S_z = I$ with $M_z = I + \gamma\Sigma_z$. For $X = M_z^{1/2} S_z M_z^{1/2} \succeq 0$, $X^2 = M_z$, so $X = M_z^{1/2}$ and $S_z = M_z^{-1/2}$. (d) $u_z = -S_z x_z$, and by (A5) $x_z \approx H_z^{-1}\nabla_z L \approx P^{-1/2}\hat{m}_1 = y$, so $u_z = -Gy$. (e) $\Delta\theta = P^{-1/2}\Delta z$ with $\Delta z = \eta u_z$. (f) Put $S = P^{1/2}GP^{1/2}$ and $R = P$. Then

$$S\big(R^{-1} + \gamma P^{-1}\Sigma P^{-1}\big)S = P^{1/2}G\,(I + \gamma\Sigma_z)\,GP^{1/2} = P = H, \tag{S6}$$

using $P^{1/2}P^{-1}P^{1/2} = I$ and $P^{1/2}P^{-1}\Sigma P^{-1}P^{1/2} = \Sigma_z$. The control $u = -R^{-1}Sx$ with $x = P^{-1}\hat{m}_1$ is $-P^{-1/2}GP^{-1/2}\hat{m}_1$, i.e. (e). (g) $\gamma = 0$ gives $G = I$ and $\Delta\theta = -\eta P^{-1}\hat{m}_1$. Diagonal Σ gives $\Sigma_z = \operatorname{diag}(\sigma_i^2/\sqrt{\hat{v}_i})$ (up to ϵ) and a diagonal G. □

Global scale. If $H = \kappa_0 P$, then $H_z = \kappa_0 I$, $S_z M_z S_z = \kappa_0 I$ gives $S_z = \sqrt{\kappa_0}\,G$, and $x_z \approx \kappa_0^{-1}y$, so $u_z = -\kappa_0^{-1/2}Gy$: the same operator, with the constant absorbed by the learning rate.

C.5 R5: Spectrum and Angle Bound (Main Text: Proposition 5(i) and (ii))

R5. *With $\Sigma_z = V\operatorname{diag}(\ell_k)V^\top$, $G = V\operatorname{diag}((1+\gamma\ell_k)^{-1/2})V^\top$ has eigenvalues in $(0, 1]$, equal to 1 on zero noise directions, decreasing in γ, $\approx (\gamma\ell_k)^{-1/2}$ for $\gamma\ell_k \gg 1$; and $\cos\angle(y, Gy) \geq 2\sqrt{ab}/(a+b)$ with $a = (1+\gamma\ell_{\max})^{-1/2}$, $b = (1+\gamma\ell_{\min})^{-1/2}$, with equality at $y = \sqrt{b}\,v_{\max} + \sqrt{a}\,v_{\min}$.*

Proof. (i) through (iii) follow from $G = V\operatorname{diag}((1+\gamma\ell_k)^{-1/2})V^\top$ with $\ell_k \geq 0$. (iv) The spectrum of G lies in $[a, b]$, so $(G - aI)(G - bI) \preceq 0$, i.e. $G^2 + abI \preceq (a+b)G$, hence $\|Gy\|^2 + ab\|y\|^2 \leq (a+b)\langle Gy, y\rangle$. By arithmetic geometric mean the left side is at least $2\sqrt{ab}\|Gy\|\|y\|$, which gives $\cos\angle(y, Gy) \geq 2\sqrt{ab}/(a+b)$. For equality, take $y = \alpha v_{\max} + \beta v_{\min}$: the cosine is $(a\alpha^2 + b\beta^2)/(\sqrt{\alpha^2+\beta^2}\sqrt{a^2\alpha^2 + b^2\beta^2})$, and $\alpha^2 = b$, $\beta^2 = a$ give $2ab/((a+b)\sqrt{ab}) = 2\sqrt{ab}/(a+b)$. □

C.6 R6: Kronecker Factorization (Main Text: Theorem 3)

R6. *(i) If $\operatorname{Cov}(\operatorname{vec_r} D) = B' \otimes A'$ and $\mathbb{E}D = 0$, then $\mathbb{E}[D^\top D] = \operatorname{tr}(B')A'$, $\mathbb{E}[DD^\top] = \operatorname{tr}(A')B'$, and $(B \otimes A)/\operatorname{tr} A = \Sigma_z$; for any Σ_z, $\operatorname{tr}[(B \otimes A)/\operatorname{tr} A] = \operatorname{tr}\Sigma_z$. (ii) $(I + \gamma(B \otimes A)/\tau)^{-1/2}\operatorname{vec_r}(Y) = \operatorname{vec_r}(U_B[(U_B^\top Y U_A) \odot G]U_A^\top)$ with $G_{ij} = (1 + \gamma\mu_i\lambda_j/\tau)^{-1/2}$.*

Proof. (i) $\mathbb{E}[D_{ij}D_{kl}] = B'_{ik}A'_{jl}$, so

$$\mathbb{E}[DD^\top]_{ik} = \textstyle\sum_j \mathbb{E}[D_{ij}D_{kj}] = \operatorname{tr}(A')B'_{ik}, \qquad \mathbb{E}[D^\top D]_{jl} = \textstyle\sum_i \mathbb{E}[D_{ij}D_{il}] = \operatorname{tr}(B')A'_{jl}. \tag{S7}$$

Then $B \otimes A = \operatorname{tr}(A')\operatorname{tr}(B')\,B' \otimes A'$ and $\operatorname{tr} A = \operatorname{tr}(A')\operatorname{tr}(B')$. For general Σ_z, $\operatorname{tr}[(B \otimes A)/\operatorname{tr} A] = \operatorname{tr} B = \mathbb{E}\|D\|_F^2 = \operatorname{tr}\Sigma_z$. No Gaussianity is needed. (ii) With orthogonal $U_B \otimes U_A$, the mixed product rule gives $B \otimes A = (U_B \otimes U_A)(\operatorname{diag}(\mu) \otimes \operatorname{diag}(\lambda))(U_B \otimes U_A)^\top$, whose diagonal at (i, j) is $\mu_i\lambda_j$; a function of a symmetric matrix acts on its eigenvalues. Finally $(U_B \otimes U_A)^\top \operatorname{vec_r}(Y) = \operatorname{vec_r}(U_B^\top Y U_A)$, the diagonal factor acts as $\odot G$, and $(U_B \otimes U_A)\operatorname{vec_r}(X) = \operatorname{vec_r}(U_B X U_A^\top)$. □

The four products cost $2n^2m + 2nm^2$ flops. The same argument holds with a diagonal factor ($U = I$ on that side). If the D_t are stationary, the bias corrected EMAs are unbiased for $\mathbb{E}[D^\top D], \mathbb{E}[DD^\top]$ and the product estimator is consistent. For a centred covariance the subtracted mean term is not Kronecker in general, which is why (A7) is an assumption; when it fails, the estimator returns the trace matched Kronecker summary of the two marginals.

C.7 R7: Split Batch (Main Text: Proposition 4)

R7. *For i.i.d. per example gradients with covariance Σ_1 and two independent halves of a batch of size b: $g = \frac{1}{2}(g_A + g_B)$ is the minibatch gradient, $\mathbb{E}\delta = 0$, $\mathbb{E}[\delta\delta^\top] = \Sigma_1/b = \mathrm{Cov}(g)$, $\mathbb{E}[gg^\top] = \mathrm{Cov}(g) + \nabla L \nabla L^\top$ and $\mathrm{Cov}(g, \delta) = 0$.*

Proof. Each half has covariance $2\Sigma_1/b$. (i) For a loss that is a mean over examples, $\frac{1}{2}(g_A + g_B)$ is the minibatch mean gradient. (ii) $\mathbb{E}\delta = \frac{1}{2}(\nabla L - \nabla L) = 0$ and $\mathrm{Cov}(\delta) = \frac{1}{4}(2\Sigma_1/b + 2\Sigma_1/b) = \Sigma_1/b = \mathrm{Cov}(g)$. (iii) $\mathbb{E}[gg^\top] = \mathrm{Cov}(g) + \mathbb{E}g\,\mathbb{E}g^\top$. (iv) $\mathrm{Cov}(g, \delta) = \frac{1}{4}[\mathrm{Var}(g_A) - \mathrm{Cov}(g_A, g_B) + \mathrm{Cov}(g_B, g_A) - \mathrm{Var}(g_B)] = \frac{1}{4}[\mathrm{Var}(g_A) - \mathrm{Var}(g_B)] = 0$. The two are uncorrelated, and independent when the noise is Gaussian. □

On batch normalized models the halves use their own batch statistics; in our experiments every optimizer uses the same two half pass there, so this is identical across methods. Under clipping with coefficient c_t, both g and δ are multiplied by c_t, so the signal to noise ratio seen by G is unchanged.

What split batching does not correct. The shipped split-batch construction already removes same-batch gradient–preconditioner coupling using the same examples and total backward cost. We did not additionally apply cross-fitted factor groups or a delta-method correction to the eigenvalues of A and B. Even in a fixed eigenbasis, the nonlinear gate has finite-sample inversion bias. For $f(\ell) = (1 + \gamma\ell)^{-1/2}$,

$$f''(\ell) = \frac{3\gamma^2}{4}(1 + \gamma\ell)^{-5/2}, \qquad \mathbb{E}f(\hat{\ell}) \approx f(\ell) + \frac{3\gamma^2}{8}(1 + \gamma\ell)^{-5/2}\mathrm{Var}(\hat{\ell}).$$

Thus a noisy eigenvalue estimate has, to second order, an upward-biased gain (too little contraction). Subtracting the displayed second-order term is a candidate correction, not an evaluated component of Evie-KF: estimating its variance adds state, and the scalar expansion does not cover eigenvector error or correlated errors between the two Kronecker factors. We therefore cannot claim a measured stability benefit or overhead for this optional refinement; the current method already removes the coupling bias at no additional data cost, and evaluating the remaining inversion correction is future work.

C.8 R8: Norm Preserving Rescale (Main Text: Proposition 5(iii))

R8. *With $c = \|y\|/\|Gy\|$: $\|cGy\| = \|y\|$, $1 \le c \le 1/g_{\min}$, and the effective gains $c(1 + \gamma\ell_k)^{-1/2}$ can exceed 1.*

Proof. (i) $\|cGy\| = \frac{\|y\|}{\|Gy\|}\|Gy\| = \|y\|$. (ii) $G \preceq I$ gives $\|Gy\| \le \|y\|$, so $c \ge 1$; $G \succeq g_{\min} I$ gives $\|Gy\| \ge g_{\min}\|y\|$, so $c \le 1/g_{\min}$. (iii) cG has eigenvalues $c(1 + \gamma\ell_k)^{-1/2}$, which exceed 1 whenever $c > (1 + \gamma\ell_k)^{1/2}$. □

With c, the radial component of the contraction is removed and the operator tests whether covariance-aware redistribution improves the direction at a learning-rate-controlled norm. Without c, both rotation and contraction remain. This gives a concrete heuristic: preserve the norm when radial scale is already well tuned and the expected benefit is directional; omit it when attenuating the whole update in noisy regimes is itself desirable. The latter is plausible at small batches or after severe clipping, but our data do not identify those regimes: the only direct comparison is one WikiText-2 point, and the batch-size sweep kept c enabled throughout.

C.9 R9: Nesting (Main Text: Proposition 5(iv))

R9. *$\gamma = 0$ (or warmup) gives AdamW; `diag_only` gives $(1 + \gamma s_i/\bar{s})^{-1/2}$; `centered=False` gives Evie-KFu; the Kronecker centred case is Evie-KF.*

Proof. (i) $G = I$ gives $c = 1$ and $\Delta\theta = -\eta P^{-1}\hat{m}_1 = -\eta\hat{m}_1/(\hat{v}^{\odot 1/2} + \epsilon)$; our implementation matches `torch.optim.AdamW` to 3×10^{-16}. (ii) A diagonal $\hat{\Sigma}_z$ commutes with P and G acts on each coordinate separately. (iii) and (iv) change only the sample that fills D. □

C.10 R10: Convergence of the Idealized Update (Main Text: Theorem 6)

R10. *Let L be L_s smooth and bounded below by $L_{\inf}$, and let $\theta_{t+1} = \theta_t - \eta T_t \tilde{g}_t$ with $T_t = P_t^{-1/2} G_t P_t^{-1/2}$, where P_t, G_t are measurable with respect to $\mathcal{F}_t$, $\mathbb{E}[\tilde{g}_t \mid \mathcal{F}_t] = \nabla L(\theta_t)$, $\mathbb{E}[\|\tilde{g}_t - \nabla L(\theta_t)\|^2 \mid \mathcal{F}_t] \le \sigma^2$, $\hat{v}_t \le G_\infty^2$ in every coordinate and $\ell_{\max}(\hat{\Sigma}_{z,t}) \le \bar{\Lambda}$. With $c_{\mathrm{lo}} = (1 + \gamma\bar{\Lambda})^{-1/2}(G_\infty + \epsilon)^{-1}$, $c_{\mathrm{hi}} = \epsilon^{-1}$ and $\eta \le c_{\mathrm{lo}}/(L_s c_{\mathrm{hi}}^2)$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla L(\theta_t)\|^2 \le \frac{2(L(\theta_0) - L_{\inf})}{\eta c_{\mathrm{lo}} T} + \frac{\eta L_s c_{\mathrm{hi}}^2}{c_{\mathrm{lo}}}\sigma^2. \tag{S8}$$

Proof. T_t is symmetric and $v^\top T_t v = (P^{-1/2}v)^\top G (P^{-1/2}v) \ge g_{\min}\|P^{-1/2}v\|^2 \ge g_{\min}(G_\infty + \epsilon)^{-1}\|v\|^2$, so $\lambda_{\min}(T_t) \ge c_{\mathrm{lo}}$; also $\lambda_{\max}(T_t) \le \epsilon^{-1}$. By smoothness and measurability, $\mathbb{E}_t L(\theta_{t+1}) \le L(\theta_t) - \eta c_{\mathrm{lo}}\|\nabla_t\|^2 + \frac{L_s \eta^2 c_{\mathrm{hi}}^2}{2}(\|\nabla_t\|^2 + \sigma^2) \le L(\theta_t) - \frac{\eta c_{\mathrm{lo}}}{2}\|\nabla_t\|^2 + \frac{L_s \eta^2 c_{\mathrm{hi}}^2}{2}\sigma^2$; take expectations, sum and use $L \ge L_{\inf}$. □

If half batch gradients satisfy $\|g_A\|, \|g_B\| \le G_h$, then $\|D\|_F^2 \le G_h^2/\epsilon$ and $\bar{\Lambda} = G_h^2/\epsilon$ is admissible for matrix blocks ($\ell_{\max} \le \operatorname{tr}\hat{\Sigma}_z = \operatorname{tr} B$); for a vector block of length n_w, $s_i/\bar{s} \le n_w$. Scope: no momentum, statistics built from earlier steps, an unbiased unclipped oracle and no rescaling; with a rescale $c_t \in [1, 1/g_{\min}]$ measurable with respect to $\mathcal{F}_t$, the bound holds with $c_{\mathrm{hi}} = (\epsilon g_{\min})^{-1}$. It is not a guarantee for Algorithm 1, which uses momentum, clipping and a same step rescale.

D Implementation Details

- $P = \hat{v}^{\odot 1/2} + \epsilon$, not $\max(\hat{v}^{\odot 1/2}, \epsilon)$; only the former makes $\gamma = 0$ coincide with PyTorch's AdamW.
- Buffers in float32 regardless of parameter dtype; `eigh` in float64 on symmetrized factors, eigenvalues clamped at 0.
- δ is clipped with the same coefficient as g; one parameter group, since the rescale c is global.
- At startup the factored operator is checked against the dense $(I + \gamma(B \otimes A)/\tau)^{-1/2}$ in float64; the run does not start if the check fails, which rules out a vectorization ordering error.
- γ is tuned jointly with the learning rate on $\{3, 10, 30, 100, 300, 1000, 3000\}$, extended by a boundary rule; this is a larger search than for single parameter baselines.

E Numerical Verification

All claims were checked in float64 on CPU (`verify_evie_kf_derivation.py`, numpy and scipy, under one minute). MC denotes Monte Carlo.

F Extended Related Work

This appendix expands Section 2. Table 2 places Evie-KF next to the methods it is closest to. The columns are the ones that decide whether two preconditioners are the same method: which statistic is accumulated, what structure it is given, and what operator is applied to the gradient.

What is and isn't new. Several ingredients of Evie-KF exist separately in prior work, and we don't claim them. Preconditioning by the centred gradient variance with exponent $-1/2$ is SDProp (Ida et al., 2017) and, with a modern centring, AdaBelief (Zhuang et al., 2020). Dense and diagonal preconditioners derived from the centred gradient-noise covariance also appear in the approximate Bayesian analysis of Mandt et al. (2017). Inverse root preconditioning with a Kronecker factored covariance is Shampoo (Gupta et al., 2018). Splitting the minibatch to estimate preconditioner statistics is used by Nayak et al. (2026). Choosing optimizer parameters by an entropic risk criterion

Claim	Result	Error
Matrix Riccati closed form	R3	8×10^{-15}
$H_z = I$ under $H = P$; failure under $H = P^2$	R4(b)	5×10^{-16}; $\|H_z - I\|$ up to 3.8
$S_z = G$ solves $S_z M_z S_z = I$	R4(c)	3×10^{-14}
Problem in θ coordinates gives the same control	R4(f)	2×10^{-14}
$\gamma = 0 \Rightarrow$ AdamW; diagonal $\Sigma \Rightarrow$ scalar gate	R4(g)	3×10^{-16}; 4×10^{-16}
$H = \kappa_0 P$ rescales the step by $\kappa_0^{-1/2}$	R4	3×10^{-14}
Spectrum in $(0, 1]$, decreasing in γ; cosine bound and equality	R5	exact; 4×10^{-16}
Factored operator vs. dense; row major identity	R6(ii)	3×10^{-15}; 7×10^{-15}
Kronecker estimator exact under separability	R6(i)	1×10^{-16} (exact), 0.3% (MC)
Gains with absolute τ move with loss scale	R6	mean gain $0.442 \rightarrow 0.155$ (0.516 if normalized)
$\mathbb{E}[\delta\delta^\top] = \mathrm{Cov}(g)$, $\mathrm{Cov}(g, \delta) = 0$, $\mathbb{E}\delta = 0$	R7	$\leq 0.5\%$ (MC, 2×10^5)
$1 \leq c \leq 1/g_{\min}$, $\|cGy\| = \|y\|$	R8	9×10^{-16}
Scalar steady state; loss rescale moves the raw gate	R1 and R2	4×10^{-16}
$r > 1$ possible for $\beta_1 \neq \beta_2$	R2	max $r = 1.84$
Update noise model: closed form, cubic, limits, collapse	§B.2	5×10^{-4} (MC) to 4×10^{-16}

Table 2: Where Evie-KF sits among adaptive and structured preconditioners. $P = \mathrm{diag}(\sqrt{\hat{v}} + \epsilon)$ is Adam's denominator and Σ_z the Kronecker factored noise covariance in coordinates whitened by Adam.

Method	Statistic	Structure	Update direction	Nests AdamW
Adam / AdamW	uncentred $\mathbb{E}[g^2]$	diagonal	$\hat{m}/\sqrt{\hat{v}}$	(is AdamW)
SDProp / AdaBelief	centred gradient variance	diagonal	$\hat{m}/\sqrt{\hat{s}}$	no
Sophia	diagonal Hessian estimate	diagonal	clipped $\hat{m}/\hat{h}$	no
TONGA	gradient covariance	low rank + block diagonal	approximate natural gradient	no
K-FAC / EKFAC	Fisher (activation and backpropagation factors)	Kronecker	$F^{-1}g$	no
Shampoo	uncentred $\mathbb{E}[GG^\top]$, $\mathbb{E}[G^\top G]$	Kronecker	$L^{-1/4}GR^{-1/4}$	no
SOAP	as Shampoo	Kronecker eigenbasis	Adam in Shampoo's eigenbasis	no
KL-Shampoo	uncentred, KL fitted factors	Kronecker	inverse root, no grafting	no
Muon	none (momentum only)	full matrix	orthogonalized momentum	no
Evie-KF	centred noise covariance, coordinates whitened by Adam	Kronecker	$P^{-1/2}(I + \gamma\Sigma_z)^{-1/2}P^{-1/2}\hat{m}$	yes ($\gamma = 0$)

appears in Can & Gürbüzbalaban (2025). What Evie-KF adds is the combination: a Kronecker factored preconditioner built from the centred noise covariance, with the $I+$ term and the $-1/2$ exponent derived from an explicit Riccati sign branch rather than chosen, paired with an update-noise variance criterion, estimated from a split batch sample that is uncorrelated with the gradient it preconditions, and applied on top of Adam's diagonal whitening so that $\gamma = 0$ recovers AdamW exactly.

F.1 Diagonal and Noise Aware Methods

The diagonal literature is large and we only cover the parts that bear on Evie-KF. vSGD (Schaul et al., 2013) and Balles & Hennig (2018) treat Adam style statistics as noise estimates and derive per coordinate step sizes from a signal to noise ratio; Evie-KF inherits that reading and extends it across coordinates. SDProp (Ida et al., 2017) and AdaBelief (Zhuang et al., 2020) are the diagonal limit

of our operator, which is why EVIE-DIAG is the ablation that isolates the contribution of Kronecker structure. Liu et al. (2026b) compare Adam with diagonal Gauss Newton preconditioners through basis choice and minibatch noise, and find that in the stochastic regime Adam behaves like $\mathrm{GN}^{-1/2}$ under a Gaussian data assumption; this is consistent with our use of Adam's denominator as the curvature proxy that defines the whitened coordinates. Kunstner et al. (2023) show that the gap between SGD and Adam on transformers persists at full batch, so noise handling is one of several things adaptive methods do. Our claim is about the noise dependent part, and we test it directly through the batch size sweep.

Kalman filter formulations model training as recursive estimation under noisy observations. Ollivier (2018) shows that online natural gradient is an extended Kalman filter, KOALA (Davtyan et al., 2022) treats the minibatch loss as a noisy measurement with an isotropic covariance, and KOALA++ (Xia et al., 2025) propagates a structured parameter covariance through gradient covariance products. These methods track the posterior covariance of the parameters under a risk neutral criterion; Evie-KF tracks the covariance of the gradient noise and applies a variance-sensitive operator to it.

F.2 KRONECKER FACTORED AND MATRIX PRECONDITIONERS

The original AdaGrad family includes a full-matrix inverse-root preconditioner (Duchi et al., 2011); GGT makes this practical by applying the inverse square root to a low-rank gradient-history matrix (Agarwal et al., 2019). Xie et al. (2025) give a unified analysis spanning diagonal, full-matrix and Kronecker preconditioners, while Nguyen et al. (2026) learn a basis in which a structured covariance preconditioner is approximately diagonal. K-FAC (Martens & Grosse, 2015) approximates Fisher blocks as Kronecker products of activation and backpropagated gradient covariances, and EKFAC (George et al., 2018) rescales in the Kronecker eigenbasis. Shampoo (Gupta et al., 2018) accumulates left and right gradient second moments; its scalable and distributed versions (Anil et al., 2020; Shi et al., 2023) made it practical at scale, and its AlgoPerf result (Kasimbeg et al., 2025) renewed interest in the family. Morwani et al. (2025) show that the square of Shampoo's approximation is one power iteration toward the optimal Kronecker approximation of the gradient covariance or Gauss Newton matrix. SOAP (Vyas et al., 2025) runs Adam in Shampoo's eigenbasis; KL-Shampoo and KL-SOAP (Lin et al., 2026) derive the factor updates from KL minimization and remove the need for Adam grafting; Eschenhagen et al. (2025) explain grafting as a fix for stale, miscaled eigenvalues and propose an adaptive eigenbasis refresh; and Liu et al. (2026a) study how Frobenius, von Neumann and LogDet divergences spread Kronecker approximation error across the covariance spectrum. PSGD (Li, 2018) fits a preconditioner, optionally Kronecker structured, from pairs of parameter and gradient perturbations. Across this family the accumulated statistic is an uncentred second moment or a curvature matrix. Near a stationary point the gradient mean vanishes and the centred and uncentred covariances coincide, so the two objects differ most early in training, during the middle of training, and at small batch. This is the regime our analysis targets.

Muon (Jordan et al., 2024) replaces the momentum matrix by its nearest semiorthogonal matrix via Newton Schulz iterations and scales to large language models (Liu et al., 2025). Bernstein & Newhouse (2024) interpret Shampoo without accumulation as steepest descent under a spectral norm, the view behind Muon. SPlus (Frans et al., 2025) combines a historical eigenbasis with instantaneous sign normalization to stabilize Shampoo style whitening. At scales with billions of parameters, Khona et al. (2026) find that SOAP and Muon outperform AdamW. Muon is our strongest baseline, and it shares one property with Evie-KF: both keep the step magnitude fixed and act only on its direction.

F.3 GRADIENT COVARIANCE, CURVATURE, AND NOISE STRUCTURE

Near a local optimum, Mandt et al. (2017) derive dense and diagonal preconditioners from the centred stochastic-gradient noise covariance to match an approximate Bayesian posterior. Their objective is sampling, not a risk-sensitive training update, but the work is a direct precedent for treating the centred covariance as an operator, not only as a scalar diagnostic. TONGA (Le Roux et al., 2008) derives the natural gradient from a generalization argument and approximates the gradient covariance with low rank and block diagonal structure. The natural Newton method (Le Roux & Fitzgibbon, 2010) argues that the gradient covariance should be used as a source of information in addition to the Hessian, not as an approximation to it. Evie-KF can be read as a structured, online

form of that combination: Adam's diagonal supplies the curvature proxy and a Kronecker factored noise covariance supplies the uncertainty. Kunstner et al. (2019) show that the empirical Fisher can be a poor curvature proxy, and Thomas et al. (2020) study how noise covariance and curvature interact to shape optimization and generalization. This line motivates keeping curvature (Adam's diagonal) and noise (Σ_z) as separate objects inside Evie-KF instead of folding them into one matrix, as the natural gradient view does (Amari, 1998; Martens, 2020).

On noise structure, Zhu et al. (2019) show that SGD noise is anisotropic and aligned with curvature, which is the structure a Kronecker covariance can represent and a diagonal one cannot. Gradient noise can also be strongly non-Gaussian and heavy tailed (Şimşekli et al., 2019; Nguyen et al., 2019); these results motivate looking beyond a Gaussian picture, although Evie-KF itself requires finite second moments. McCandlish et al. (2018) define the gradient noise scale $B_{\text{simple}} = \operatorname{tr}(\Sigma)/\|G\|^2$ and relate it to the largest useful batch size; Zhang et al. (2025) study how that critical batch size scales in pretraining. Because Evie-KF already estimates $\operatorname{tr}(\Sigma)$ every step, we report B_{simple} as a free diagnostic.

For estimation, Nayak et al. (2026) identify two finite sample biases in preconditioned updates, coupling between gradient and preconditioner estimated from the same batch and bias from inverting a noisy preconditioner, and correct them with cross fitted microbatch groups and a delta method term. Our split batch sample addresses the first bias structurally: with equal halves, $\operatorname{Cov}(g, \delta) = \frac{1}{4}\big(\operatorname{Var}(g_A) - \operatorname{Var}(g_B)\big) = 0$, using the same examples and the same total backward cost as one full pass. It does not remove inversion bias in the nonlinear matrix function; Appendix C gives a fixed-eigenbasis delta expansion and states explicitly that we did not evaluate the correction.

F.4 Risk Sensitive Control and Optimization

Exponential quadratic (LEQG) control was introduced by Jacobson (1973) and developed by Whittle (1981). Its two process-noise sign branches are the starting point for Evie-KF, but the distinction matters: the globally defined $(I + \gamma\Sigma)$ contraction is the optimistic process-noise branch, while our risk-aversion interpretation comes from the update-noise criterion in Appendix B.2. In optimization, Can & Gürbüzbalaban (2025) bound the entropic risk and entropic value at risk of generalized momentum methods and choose their constant parameters to trade risk against convergence rate, and Gürbüzbalaban et al. (2025) use the risk sensitive index to analyse robustness of accelerated methods to biased, sub Gaussian gradient errors. Both are analyses of, or parameter choices for, existing methods with scalar parameters on strongly convex objectives. Evie-KF uses risk-sensitive machinery plus a separate update-noise variance criterion to derive a data-dependent matrix operator at each step of nonconvex training.

F.5 Benchmarking Methodology

Optimizer comparisons are sensitive to the tuning protocol. Choi et al. (2019) show that rankings can flip with the search space, Schmidt et al. (2021) benchmark fifteen optimizers and find no method that dominates across tasks, and AlgoPerf (Dahl et al., 2023; Kasimbeg et al., 2025) fixes workloads, hardware and tuning rules to measure time to result. Zhao et al. (2025) find that, once tuned, several diagonal optimizers perform comparably on autoregressive language modelling, and Wen et al. (2025) report that many claimed speedups over AdamW shrink under careful tuning and with scale. These results shaped our protocol. We registered the primary metric for each domain before the run, used a fixed learning rate grid with a frozen boundary rule, a disclosed tuning budget per optimizer, task level Wilcoxon tests with a sign test check, and Holm correction across baseline comparisons.

G Extended Results and Ablations

G.1 Full Multiple Domain Results

Evie-KF and Shampoo were the only arms positive in all three domains, and Muon, the highest-ranked arm overall, was negative on finance.

The representative tasks use a rule independent of Evie-KF; the full per-task values are in Table 1.

Table 3: Median relative improvement over AdamW (%) across each domain's four tasks; positive favours the row. "Wins" counts tasks beaten against AdamW out of 12. Relative differences are used because the metrics within a domain are not on a common scale; this is the quantity the confirmatory test aggregates.

	Vision	Language	Finance	All 12	Wins	Mean rank
Evie-KF	$+0.22$	$+4.68$	$+3.32$	$+3.32$	8/12	3.67
AdamW	N/A	N/A	N/A	N/A	N/A	4.42
AdaBelief	-0.40	$+0.35$	-0.99	-0.45	4/12	5.08
SGD	$+0.14$	-6.88	$+5.13$	$+0.90$	8/12	4.08
Sophia	-2.16	-8.69	-4.06	-4.66	2/12	6.33
Muon	$+3.10$	$+11.43$	-0.47	$+4.74$	8/12	3.00
Lion	-0.22	-6.75	$+0.32$	-0.61	4/12	5.50
Shampoo	$+0.30$	$+5.09$	-1.27	$+0.30$	7/12	3.92

Table 4: One representative task per domain: the largest label space, the largest corpus, and the panel used as the host throughout Sections 5 and 6. Median over five seeds; best per column in bold. Mean rank is over all 12 tasks; the full grid is Table 1.

	CIFAR-100 top-1 (%)	WikiText-103 perplexity	Finance-Tech MSE $\times 10^{-4}$	Mean rank (12 tasks)
Evie-KF	52.28	35.41	4.1497	3.67
AdamW	49.51	39.99	4.3336	4.42
AdaBelief	49.45	41.17	4.3745	5.08
SGD	**54.27**	59.75	4.1537	4.08
Sophia	47.39	39.51	4.5594	6.33
Muon	54.24	**33.49**	**4.1475**	**3.00**
Lion	49.59	47.31	4.1502	5.50
Shampoo	50.80	36.25	4.2170	3.92

Muon and SGD also beat AdamW on 8 of 12 tasks. Evie-KF's WikiText-2 margin was $+5.0\%$ with 5/5 seed wins; WikiText-103 and CIFAR-100 also won 5/5. CIFAR-10 was $+0.48\%$, and the full-factor control above ruled out the diagonal fallback as its cause. STL-10 scored 48.59 against AdamW's 51.29. Full divergence counts were Evie-KF 2/1,025 (0.20%), Muon 34 (3.32%), Lion 30 (2.93%), Shampoo 21 (2.05%), Sophia 18 (1.76%), SGD 13 (1.27%), and zero for AdamW and AdaBelief.

Against Sophia, Evie-KF had median $d = -9.34\%$, Wilcoxon $p = 0.00098$, Holm $p = 0.0068$ and sign-test $p = 0.0063$. Muon led Evie-KF on 10/12 but was not significant because the two finance losses dominated its signed ranks. In the AdamW-referenced analysis, Sophia being worse was the only significant result (Holm $p = 0.034$).

The descriptive sweeps used seeds 42 to 44. WikiText-2's five batch points were 15/15 wins, mostly at the same learning rate; against AdaBelief the gains were 22.6% at batch 2 and 5.9% at batch 4 (3/3 each), followed by losses from batch 8. CIFAR-10 won every seed at batches 4, 8 and 16 by up to 3.2 points, with no gap at 32. With only 4 to 5 points, the batch correlations were patterns, not tests. Muon's largest leads were 13.5 points at CIFAR-10 batch 4 and 20% at WikiText-2 batch 2. Evie-KF cost $7.8\times$ AdamW at batch 2, and the small-batch pattern appeared on two of three hosts. For large WikiText-2, using AdamW's learning rate recovered 5.1 points from the raw gap and produced 104.23, still 11.7% worse than 93.32. The sweeps ran on CIFAR-10, WikiText-2 and finance-tech; SVHN, Penn Treebank and healthcare were planned and were held back by GPU budget.

A harder vision task. Tiny-ImageNet-200 is the largest label space and the highest resolution we train on, and it is the vision task where the gradient noise argument should have the most room. Evie-KF reaches 37.54% against AdamW's 36.08% and AdaBelief's 36.22%, ahead of both on every one of three seeds (Table 1); Muon leads at 41.88%. The ordering repeats CIFAR-100, the other many-class vision task, where Evie-KF gains 5.6% over AdamW.

Table 5: Batch-size sweep. Every batch size runs the same 10,000 steps, so a smaller batch sees proportionally less data and absolute values are comparable only *within* a row. Three seeds (42–44), learning rate re-searched per (task, batch, arm), γ pinned at 3,000 for both Evie arms. The baseline row reuses the main run's own cells, recomputed over these three seeds and these arms, so on finance its complete-case ticker set differs from Table 1.

batch	AdamW	AdaBelief	Sophia	Muon	Evie-KF	Evie-KFu	units
CIFAR-10 — test top-1 accuracy (%), higher is better							
4	51.79	50.70	50.43	**65.94**	52.45	51.31	1
8	67.98	67.61	69.76	**78.91**	70.64	71.05	1
16	75.73	76.18	78.73	**83.76**	77.65	78.25	1
32	79.89	79.19	80.68	**85.09**	79.36	79.89	1
128 (main run)	81.55	81.47	81.45	**85.86**	81.94	—	1
WikiText-2 — test perplexity, lower is better							
2	138.63	125.82	121.35	**80.95**	97.35	94.43	1
4	91.36	91.13	98.57	**80.93**	85.75	85.25	1
8	103.39	88.95	105.65	**87.23**	92.83	90.61	1
16	107.78	92.57	115.36	**85.59**	93.64	93.35	1
32 (main run)	105.86	97.15	119.38	**83.62**	99.67	—	1
Finance-tech — test MSE $\times 10^{-4}$, lower is better							
2	3.7825	3.7906	3.7838	3.7834	3.7850	**3.7823**	39
4	4.2215	4.3953	4.2298	**4.2134**	4.2380	4.5152	19
8	3.5945	3.7588	3.6965	3.6579	3.5977	**3.5784**	36
16	4.1460	4.0394	**4.0285**	4.0674	4.0577	4.1751	44
32 (main run)	4.3734	4.4754	4.7083	**3.9095**	3.9102	—	47

Table 6: Model-size sweep, same fixed step budget and the same three seeds. "Seeds" counts how many of the three Evie-KF wins against AdamW. The baseline row reuses the main run's cells under the same recomputation as Table 5.

size	params	AdamW	AdaBelief	Evie-KF	seeds	units
CIFAR-10 — test top-1 accuracy (%), higher is better						
small	2,797,610	80.53	79.99	**81.17**	3/3	1
baseline (main run)	11,173,962	81.55	81.47	**81.94**	—	1
large	44,662,922	81.91	**82.34**	81.34	0/3	1
WikiText-2 — test perplexity, lower is better						
small	2,675,840	105.42	107.60	**98.35**	3/3	1
baseline (main run)	8,900,608	105.86	**97.15**	99.67	—	1
large	22,213,248	**93.32**	95.63	109.78	0/3	1
Finance-tech — test MSE $\times 10^{-4}$, lower is better						
small	737	3.7886	3.7867	**3.7800**	2/3	43
baseline (main run)	2,497	4.4420	4.4757	**4.0295**	—	48
large	9,089	4.5417	4.5873	**4.1989**	3/3	49

G.2 Batch Size and Model Size Sweeps

What the sweeps hold fixed. Both sweeps keep the 10,000-step budget of the main run at every point, so a smaller batch sees proportionally less data and a larger model costs more per step. Absolute values are therefore comparable within a row and not down a column, which is all the design needs: every arm receives an identical step and data budget at each point. We re-search the learning rate for every (task, point, arm) and pin $\gamma = 3{,}000$ for the Evie arms, using one value instead of the joint search performed in the main run; on finance the main run's own search chose $\gamma = 10$, so that host runs the operator far harder than the protocol selects for it, and we read its numbers with that in mind.

Batch size. On CIFAR-10 Evie-KF is ahead of AdamW and of AdaBelief on every seed at batches 4, 8 and 16, by up to 3.2 points, and the arms converge by batch 32 (Table 5). On WikiText-2 Evie-KF is ahead of AdamW at all five batch sizes on three of three seeds each — mostly at the same learning rate — and by the widest margin, 29.8%, at batch 2; it also overtakes AdaBelief below batch 8, having trailed it above. That crossover is the clearest support in the benchmark for the premise that

Table 7: The four ablation arms. Each arm runs its own joint (lr, γ) search on the host, ten seeds (42–51), on the same task as the main run. Evie-DiagAbs is the derivation's own diagonal case, $(I + \gamma\,\mathrm{diag}(\Sigma_z))^{-1/2}$; Evie-Diag additionally divides the spectrum by its mean. Vision reports test top-1 (%), language test perplexity, finance test MSE $\times 10^{-4}$.

Arm	gate	Finance-tech	WikiText-2	CIFAR-10
Evie-KF	Kronecker, absolute, centred	3.9101	99.65	81.19
Evie-KFu	Kronecker, absolute, uncentred	3.9113	99.92	80.53
Evie-DiagAbs	diagonal, absolute	4.1931	106.77	—
Evie-Diag	diagonal, mean-normalised	4.4718	105.73	81.53

the gain tracks gradient noise. On finance-tech the arms are level at batches 4 to 16, within 0.5% of AdamW. At batch 2 AdamW leads by 6.8% at task level while the per-ticker median gap is 0.09%, which concentrates the difference in a small number of tickers, not across the panel, and we report the task-level figure because it is the one the protocol aggregates.

Model size. Evie-KF is best at the smallest point on all three hosts (Table 6), on three of three seeds on CIFAR-10 and WikiText-2. On finance the margin over AdamW grows with capacity — level on the 737-parameter model, 5.1% at the main-run size and 7.6% on the 9,089-parameter model, three of three seeds — and the operator's activity grows with it, from $\cos(\Delta_{\text{Evie-KF}}, \Delta_{\text{AdamW}}) = 0.98$ on the small model to 0.68 on the large. On CIFAR-10 the largest model favours AdaBelief by 1.0 point, and on WikiText-2 it favours AdamW at a learning rate one grid step below Evie-KF's own, which the sweep does not separate from the capacity effect. The three domains do not share a single capacity trend, and the 33.7M-parameter OpenWebText result in Section 5 is the largest model we train.

Centring across the sweeps. Evie-KFu, the uncentred variant, matches Evie-KF at every batch size and every model size to within the seed spread, including the smallest batch, where the centred estimate would have the most to offer. This is the same reading as the ablations in Appendix G.3: the Kronecker structure carries the effect and centring comes free with the split-batch estimate.

G.3 Full Scale Covariance Ablations

Separating structure from normalisation. The diagonal arm we shipped first, Evie-Diag, divides the spectrum by its mean, while Evie-KF's gate is absolute. A contrast between them therefore reads out two design choices at once, and an independent γ search per arm does not separate them, because the two conventions differ by a mean eigenvalue that varies per layer and over training. Evie-DiagAbs is the diagonal case the derivation itself produces, and it isolates the Kronecker structure on its own. We report that contrast as the mechanism result and keep the mean-normalised one beside it (Table 8).

The structural effect is the same size on both hosts. Kronecker structure alone is worth $+6.75\%$ on finance-tech and $+6.67\%$ on WikiText-2 — two domains, two metrics, two independent searches, landing within 0.1 percentage points of each other, on nine and ten seeds out of ten. Both clear Holm correction across all eight contrasts. The contrasts that also carry the normalisation convention are $+12.56\%$ and $+5.75\%$, a factor of two apart, so the deconfounded measurement is the steadier one as well as the narrower claim. Normalisation on its own moves the two hosts in opposite directions, $+6.23\%$ and -0.98%, and we draw no conclusion from it.

Where the operator is active. On CIFAR-10 the ablation's search selected $\gamma = 1$, at which the gate is nearly inactive: $\cos(\Delta_{\text{Evie-KF}}, \Delta_{\text{AdamW}}) = 0.991$, a few degrees from AdamW. That host therefore exercises little Kronecker structure, and we report it for completeness, not as a third reading. The main run's own search on the same task chose $\gamma = 27{,}000$ and $\cos = 0.47$, the most active operator in the benchmark, and scored 81.94 against this ablation's 81.19 — two searches at opposite ends of the γ range, landing within a point of each other, so on CIFAR-10 the operator's strength has little consequence either way.

For the structural contrasts, the unadjusted two-sided values were $p = 0.0039$ on finance-tech and $p = 0.0020$ on WikiText-2. On finance, normalisation had flattered Evie-KF and the effect fell from 12.56% to 6.75%; on language it had flattered the diagonal arm and the effect rose from 5.75% to 6.67%. Absolute and relative gates selected $\gamma = 3$ versus 10 on finance and 1000 versus 3 on

Table 8: Every ablation contrast, paired over the ten seeds; positive favours Evie-KF. p is the exact two-sided Wilcoxon signed-rank value and the last column applies Holm correction across all eight contrasts in this table.

Host	contrast isolates	effect	seeds	p	Holm p
Finance-tech	Kronecker structure alone	$+6.75\%$	9/10	0.0039	**0.0195**
	structure and normalisation	$+12.56\%$	9/10	0.0039	0.0195
	centring	$+0.03\%$	6/10	0.2324	0.4648
WikiText-2	Kronecker structure alone	$+6.67\%$	10/10	0.0020	**0.0156**
	structure and normalisation	$+5.75\%$	10/10	0.0020	0.0156
	centring	$+0.27\%$	5/10	0.8457	0.8457
CIFAR-10	structure and normalisation	-0.42%	3/10	0.0488	0.1465
	centring	$+0.82\%$	10/10	0.0020	0.0156

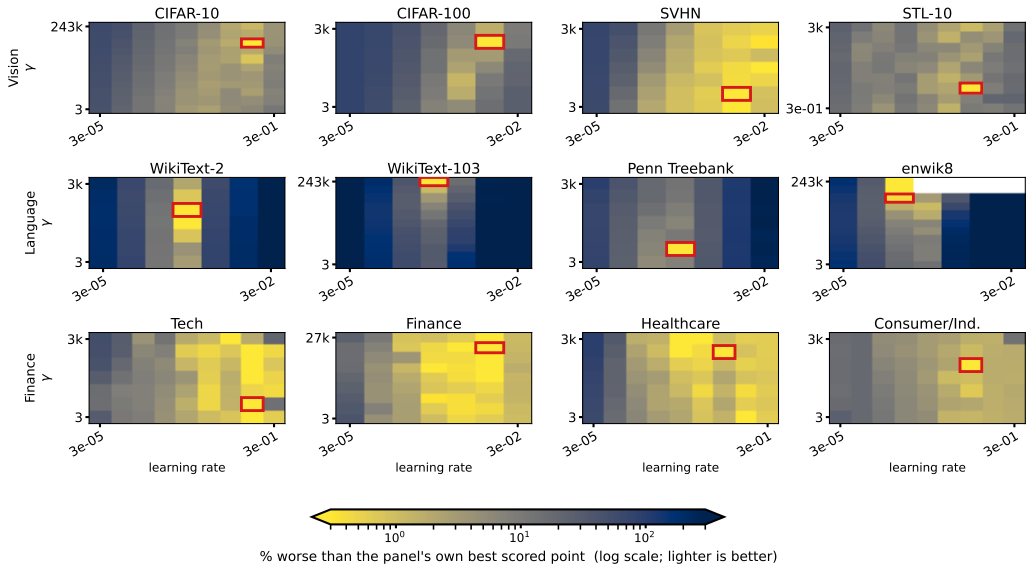


Figure 8: All twelve joint (lr, γ) searches show percent above each panel's best on a logarithmic scale, with lighter cells closer to each panel's best; the blue outline marks the frozen choice. White cells were never scored.

WikiText-2. Evie-DiagAbs had zero divergences over 490 finance cells, while Evie-KF and Evie-KFu lost two tickers; on WikiText-2 its seed spread was 7.7% against 2.7 to 4.2%. CIFAR-10 reached $\gamma = 1$ after two downward extensions.

Centring matched the pre-run prediction that centred and uncentred moments converge as the gradient mean shrinks. On the CIFAR-10 batch sweep, the uncentred variant matched or slightly beat the centred variant from batch 4 through 32. Centring was free because the split pass was already required and was never observed to hurt, but it showed no benefit.

Enwik8 improved monotonically over four orders of magnitude in γ, from 4.3948 to 3.6956 (15.9%). In the fixed-$\gamma = 3000$ control, STL-10 changed from 48.59 to 51.49 ($+5.97\%$, 5/5), SVHN from 94.634 to 94.699 ($+0.07\%$, 4/5), CIFAR-10 from 81.94 to 80.40 (-1.88%, 0/5), and Penn Treebank from 192.17 to 197.53 (-2.79%, 1/5). These four hosts were chosen after their main-run γ values looked suspicious, so the STL-10 result diagnoses the tuning protocol and does not replace Table 4.

Tuning-budget interpretation. The fixed-γ control gives Evie-KF the same one-dimensional search as the single-axis baselines and costs 0.9% at the median across these four hosts. Its individual outcomes move in both directions, which is why we report all four above rather than treating the median as a universal guarantee. Mean-eigenvalue normalisation separately moved WikiText-2 from 300 to 30, Penn Treebank from 10 to 100 and CIFAR-10 from 27,000 to 30; the learning rates

Table 9: Wall-clock per cell relative to AdamW and peak memory, median over each domain's four tasks, all on one T4. On the vision tasks every arm takes the split backward pass, because the models carry BatchNorm; on the language and finance tasks only the Evie arms do.

	Vision		Language		Finance	
Optimizer	time	peak MB	time	peak MB	time	peak MB
AdamW	1.00×	780	1.00×	2560	1.00×	17
AdaBelief	1.04×	780	1.00×	2560	1.29×	17
SGD	0.99×	737	0.93×	2529	1.02×	17
Sophia	4.21×	2349	2.79×	7866	2.42×	18
Muon	1.36×	737	1.05×	2542	2.31×	17
Lion	0.99×	737	0.94×	2529	1.08×	17
Shampoo	1.05×	782	1.36×	2699	2.09×	18
Evie-KF	1.61×	851	1.60×	1981	4.74×	19

Table 10: Evie-KF telemetry at the frozen hyperparameters, median over cells. cos is the angle between Evie-KF's update and AdamW's on the same state, so 1 is an inactive operator; B_{simple} is McCandlish's gradient noise scale, free here because the operator already estimates $\operatorname{tr}\Sigma$; "min gain" is the smallest eigenvalue of G, so a small value is a strong shrink along the noisiest direction.

Task	cells	$\cos(\Delta_{\text{Evie-KF}}, \Delta_{\text{AdamW}})$	B_{simple}	noise eff. rank	min gain
vis/cifar10	5	0.468	15.7	3.5×10^4	0.00553
vis/cifar100	5	0.794	12.3	1.2×10^5	0.00682
vis/stl10	5	0.881	7.76	8.83×10^3	0.761
vis/svhn	5	0.942	15.3	4.5×10^4	0.277
lang/enwik8	5	0.595	19.1	1.9×10^3	0.00143
lang/ptb	5	0.993	9.49	3.46×10^4	0.321
lang/wikitext103	5	0.559	18.7	7.87×10^3	0.000915
lang/wikitext2	5	0.905	10.5	4.18×10^4	0.0504
fin/consumer_industrials	250	0.798	38.7	5.53	0.601
fin/finance	245	0.596	46.9	3.27	0.0958
fin/healthcare	245	0.757	46.3	5.37	0.363
fin/tech	135	0.922	203	4.09	0.98

were inherited from the shipped operator. The resulting $3.3\times$ spread and the fact that $\gamma = 30$ lies within 0.33% of each optimum support a narrow default range on these hosts. Per-domain validation remains appropriate until the normalised operator is run across the full benchmark.

Norm-preservation diagnostic. Disabling norm preservation improved WikiText-2 by 3.41% on 5/5 seeds at one hyperparameter point tuned with the rescale. This shows that c is not load-bearing for Evie-KF's benefit and that the derived shrink-only form is a strong variant rather than a failure mode. R8 gives the concrete rule: retain c when learning-rate tuning already supplies the desired radial scale and directional redistribution is the target; omit it when global attenuation in noisy regimes is useful. The current data do not separate batch-size or clipping regimes because all batch-sweep runs kept c enabled and the on/off comparison inherited (lr, γ); a crossed, retuned study remains future work.

G.4 Efficiency and Resource Usage

Cost falls as the model grows. Evie-KF's overhead is the eigendecomposition of its two factors, refreshed every twenty steps, plus the split backward pass. Neither scales with the forward and backward cost, so the relative overhead shrinks with model size: $4.74\times$ AdamW on the 2,497-parameter finance MLP, $1.6\times$ at 9M to 11M parameters, and $0.98\times$ at 33.7M on OpenWebText, where it is the cheaper of the two. Sophia is the more expensive second-order arm throughout, at $2.4\times$ to $4.2\times$ and up to three times AdamW's peak memory. The split pass costs no extra FLOPs: two half batches process the same examples as one full pass.

Memory. Evie-KF stores A, B and their eigenbases, about $2(n^2 + m^2)$ floats per $n \times m$ tensor against AdamW's $2nm$, which on these models is 9% above AdamW on vision and 12% on finance.

The language column reads lower than AdamW's because only the Evie arms take the split backward there, and two half batches hold less activation memory than one full batch; under a matched split pass on OpenWebText the like-for-like figure is $1.01\times$ AdamW, and that is the number we quote in Section 5.

The operator is doing work, and we can see how much. Table 10 reports the angle between Evie-KF's update and AdamW's at the frozen hyperparameters. It ranges from 0.47 on CIFAR-10, roughly $62°$, to 0.993 on Penn Treebank, where the selected γ leaves the gate nearly off — and Penn Treebank is the task where Evie-KF and AdamW are closest, which is what a gate that reduces to AdamW at $\gamma = 0$ predicts. The two tasks with the largest margins over AdamW, WikiText-103 and CIFAR-100, both run an active operator, at 0.56 and 0.79, with minimum gains of 9×10^{-4} and 7×10^{-3}: the gate is shrinking the noisiest direction by three orders of magnitude relative to the cleanest while preserving the whitened step length. The estimated noise rank separates the domains cleanly — 10^3 to 10^5 on the vision and language tensors against 3 to 6 on the finance MLP — so the Kronecker factors are carrying real structure on the larger models.

H Reproducibility, Code, and PyPI Package

H.1 Implementation and Test Suite

Every task is a single self-contained script, generated from eight shared fragments so that the optimizer block, the training engine and the three data loaders are byte-identical across all 43 generated files; a check in the build refuses to generate if a file has drifted from its fragment. The released package is not a second copy of the optimizer: its source is sliced out of that same fragment, and a test in the release checks the class, the self-check and the defaults line by line against it, and the fragment in turn against a generated task file. The code a reader installs and the code that produced every number here are checked to be identical. The optimizer state lives in float32, the symmetrised Kronecker factors are eigendecomposed in float64 every twenty steps with eigenvalues clamped at zero, and a factor above dimension 1,024 falls back to its diagonal.

Five test modules gate the implementation, all CPU-only and all running in seconds: sixteen `pytest` cases over the shared optimizer block and the Evie-KF family, and fifty-four assertion gates over the derivation, the diagnostic switches and the fairness and aggregation rules. The two that matter most are end-to-end. One builds the update the derivation specifies in fully dense form — forming the whole $nm \times nm$ matrix and taking one eigendecomposition of it, sharing no code path with the optimizer beyond PyTorch — and agrees with the shipped factored operator to 6×10^{-8} across seven configurations. The other checks that $\gamma = 0$ reproduces `torch.optim.AdamW` to 3×10^{-16}, which tests the nesting property of Proposition 4(iv) directly. At startup every run repeats the dense check on its own model and refuses to begin if it fails, which rules out a vectorisation-ordering error in a configuration the tests never saw. Appendix E records the float64 checks that every claim in Section 3 was verified against.

H.2 Anonymized Code Access

We release the complete harness: all 43 generated task files, the fragments they are built from, the test suite, the statistical pipeline that produces every table in this paper, the frozen finance panel snapshots, and the raw per-cell records behind all 8,200 main-run cells. The tables and figures here are regenerated from those records by scripts in the release, with no numbers entered by hand.

The optimizer is also packaged on its own, so that reproducing the method does not require the harness: `pip install evie-kf` gives a `torch.optim.Optimizer` whose defaults are the ones every run in this paper used, and which reduces to `torch.optim.AdamW` at $\gamma = 0$. It is not a drop-in replacement for a single-pass loop, because the operator needs the split-batch noise sample of Proposition 4: the caller supplies it through `set_noise()` before each `step()`, or uses the `split_step()` helper, which runs the two half batches and the clipping exactly as our engine does. A loop that supplies neither trains with $G = I$, which is AdamW, and the optimizer's `n_active_steps` counter reads zero at the end of such a run. Both the package and the repository are published from accounts created for this submission and carry no author information.

Table 11: Every frozen learning rate, and Evie-KF's frozen γ. Each is the best-validation-checkpoint score over that optimizer's own grid at 2,500 steps on seed 42, after any boundary extensions, and is then held fixed for all five seeds.

Task	Evie-KF	AdamW	AdaBelief	SGD	Sophia	Muon	Lion	Shampoo	γ
CIFAR-10	1E-1	1E-2	1E-2	3.0E-1	3.2E-3	6.3E-3	3.2E-3	3.2E-2	27,000
CIFAR-100	1E-2	3.2E-3	3.2E-3	3.0E-1	1E-3	2.0E-2	1E-3	3.2E-3	1,000
STL-10	3.2E-2	1E-2	3.2E-3	3.0E-1	3.2E-3	2.0E-2	3.2E-3	3.2E-3	3
SVHN	1E-2	3.2E-3	3.2E-3	9.5E-2	3.2E-3	6.3E-3	3.2E-3	3.2E-3	10
enwik8	3.2E-4	3.2E-4	3.2E-4	9.5E-2	1E-3	6.3E-3	1E-4	3.2E-3	27,000
Penn Treebank	1E-3	3.2E-4	3.2E-4	9.5E-2	3.2E-4	6.3E-3	1E-4	1E-3	10
WikiText-103	1E-3	1E-3	1E-3	9.5E-2	1E-3	2.0E-3	1E-4	3.2E-3	243,000
WikiText-2	1E-3	1E-3	3.2E-4	9.5E-2	1E-3	6.3E-3	1E-4	3.2E-3	300
Consumer/Ind.	3.2E-2	3.2E-2	3.2E-2	9.5E-1	1E-2	6.3E-3	3.2E-4	1E-1	300
Finance	1E-2	1E-2	3.2E-3	9.5E-1	1E-3	2.0E-4	3.2E-3	3.2E-3	9,000
Healthcare	3.2E-2	3.2E-3	3.2E-3	9.5E-1	3.2E-1	2.0E-3	1E-3	3.2E-2	1,000
Tech	1E-1	1E-2	1E-2	9.5E-1	3.2E-3	6.3E-1	3.2E-3	3.2E-1	10

Table 12: The benchmark. Time and memory are Evie-KF against AdamW, measured over full 10,000-step cells on one T4; memory is the peak for each arm.

Domain	Tasks	Model	Params	Batch	Metric	Time	Memory (MB)
Vision	4	ResNet-18, CIFAR stem	11.2M	128	top-1 (%)	$1.59\times$ to $1.65\times$	851 / 780
Language	4	Transformer, 6 layers, d=256	8.9M	32	perplexity	$1.50\times$ to $1.72\times$	2,377 / 2,856
Finance	4	MLP, $5 \to 64 \to 32 \to 1$	2,497	32	MSE	$4.72\times$ to $4.87\times$	19 / 17

H.3 Hyperparameters and Hardware

The spread in Evie-KF's γ column comes from the shipped normalisation, not the tasks: dividing the Kronecker spectrum by $\mathrm{tr}\, A$ leaves γ carrying each task's absolute noise energy, and normalising by the mean eigenvalue instead collapses the same three hosts from a $2{,}700\times$ spread to $3.3\times$ (Section 6).

The vision tasks used batch size 128 and an 11.2M-parameter CIFAR-stem ResNet-18 with a 3×3 stem and no initial max-pool. STL-10 was resized to 32×32 and retained 5,000 labelled training images. The language tasks used batch size 32 and an 8.9M-parameter 6-layer decoder-only Transformer with $d_{\mathrm{model}} = 256$, 8 heads and context 256. WikiText-2 and Penn Treebank used whitespace tokenisation with a 16k-word cap, so their perplexities are comparable only between optimizers on the same task. The committed finance snapshots contained 49 Tech, 49 Finance, 49 Healthcare and 50 Consumer/Industrials tickers. Each batch-32 MLP predicted next-day log return from standardised OHLCV features; splits were chronological 80/10/10, the scaler used training data only, and a ticker required 300 usable rows.

All arms decayed the learning rate to one tenth, used pre-update decoupled decay $p \leftarrow p(1 - \mathrm{lr}\lambda)$ with $\lambda = 10^{-4}$, Xavier-uniform weights and zero biases, and evaluated the untouched test split once. The 7-point grids were centred at 10^{-3} for Adam-family arms and Shampoo, 3×10^{-2} for SGD, 2×10^{-2} for Muon and 10^{-4} for Lion. Search scored the best validation checkpoint, not the final step, which otherwise selected underfit rates; an edge winner triggered two more half-decade points, for at most three logged extensions. Shampoo searched (lr, β), Evie-KF searched (lr, γ), and the other six arms searched learning rate. Finance tuning used the same 8 tickers per panel, drawn once with `random.Random(42)`.

Finance retained paired complete cases only when every arm finished every seed: 45/50 Consumer/Industrials, 23/49 Finance, 33/49 Healthcare and 37/49 Tech. Wall time, memory, telemetry and secondary metrics were descriptive. A cell diverged when its loss was non-finite, its metric left the task's absolute sane range, or it exceeded the peer-median factor. Seed 42 shared the search batch order and may be optimistic; the seed-excluded rescore and provenance files accompany the full tables.

Runs used PyTorch 2.10.0+cu128 with per-cell checkpoints. Exact Evie-KF time ratios were $1.59\times$ to $1.65\times$ in vision, $1.50\times$ to $1.72\times$ in language and $4.72\times$ to $4.87\times$ in finance. Vision peak mem-

ory was 851 MB against 780 MB; WikiText-2 used 2,377 MB against 2,856 MB because the split batch reduced activation memory. Finance models had 2,497 parameters, making the fixed eigendecomposition dominant; Sophia cost $2.31\times$ to $4.73\times$ on those cells. Eigenbases refreshed every 20 steps, with a diagonal fallback above dimension 1024. Every 512-channel CIFAR-stem convolution has input dimension 4608; raising the limit to 4608 changed CIFAR-10 by $+0.12$ accuracy points ($+0.15\%$) on 4/5 seeds ($p = 0.625$), inside the 1.4-point seed spread, but cost $4.3\times$ wall time and $2.7\times$ peak memory, making Evie-KF $6.9\times$ AdamW versus $1.59\times$.

I Broader Impact, Limitations, and Checklist

I.1 Broader Impact

Evie-KF reduces the number of steps needed to reach a given validation score on the tasks we plot, which lowers the compute and energy a training run costs. The same efficiency lowers the cost of training models with downstream risks, and we make no claim that the method is selective about which. Its cost profile is the ordinary one for a preconditioner: a constant-factor overhead that falls as the model grows, so the benefit accrues mainly to groups already training at scale. We report wall time, peak memory, divergence rates and the tasks where Evie-KF is behind alongside the tasks where it is ahead, so that an efficiency claim can be checked independently of the headline number. The finance experiments are methodological — fixed panels of historical returns used to compare optimizers — and are not evidence of trading performance or a basis for investment decisions.

I.2 Limitations

Mean normalisation supports a narrow default range. Dividing the Kronecker spectrum by $\operatorname{tr} A$ matches the estimator's trace but is not scale-free, so the selected γ absorbs each task's absolute noise energy and ranges from 3 to 243,000 across the twelve. Normalising by the mean eigenvalue makes γ dimensionless and collapses three hosts to a $3.3\times$ spread, with a single $\gamma = 30$ within 0.33% of each host's own choice. This is evidence for a narrow default range, not yet evidence for one universal value: it covers three hosts and inherits their learning rates. We report the shipped operator because it produced every main result; full-benchmark validation of the normalised form is the next step. The present evidence therefore supports a small default range with light validation rather than the broad task-specific search required by the shipped scaling.

Baseline-budget controls retain the aggregate result. Evie-KF scores 49 to 99 points per task against seven to nine for the single-axis arms. Fixing $\gamma = 3{,}000$ and spending the baselines' budget on learning rate alone costs 0.9% at the median across four control hosts, with two improving and two declining. This baseline-budget control shows that the aggregate result does not depend entirely on the extra axis, while the main-run numbers were produced with joint tuning and we report that larger budget in full. Together with the normalised-γ study, this makes a one-dimensional learning-rate search a supported deployment option, while a full twelve-task fixed-γ rerun would provide the strongest final fairness check.

Large factors fall back to a diagonal. A Kronecker factor above dimension 1,024 is treated diagonally, which on the CIFAR-stem ResNet-18 covers every 512-channel convolution, whose input factor has dimension 4,608. Lifting the limit to 4,608 changes CIFAR-10 by $+0.12$ points on four of five seeds ($p = 0.625$, inside the task's own 1.4-point seed spread) at $4.3\times$ the wall time and $2.7\times$ the peak memory, so the fallback is the right default and the vision results are not an artefact of it.

Norm preservation is optional rather than load-bearing. The rescale that restores AdamW's step length turns the operator from a shrink into a rotation. Disabling it on WikiText-2, at hyperparameters tuned with it on, improves perplexity by 3.41% on five of five seeds. This supports the shrink-only form — which is what the derivation itself produces — as a competitive variant. Because the comparison covers one host and one inherited setting, it does not establish when c helps across batch sizes or clipping levels; that requires a crossed, retuned ablation. The favorable shrink-only result nevertheless shows that Evie-KF's covariance operator remains effective without this implementation choice.

A positive scale check, with larger spike diagnostics still open. The largest model tested is the 33.7M-parameter OpenWebText transformer, where Evie-KF improves perplexity over AdamW by

13.1% on three of three seeds at $0.98\times$ wall time and near-matched peak memory. This is encouraging evidence that the structured operator remains useful as model size increases. Muon reaches 55.73 on the two seeds that completed there, so this is evidence that the operator scales within the Adam family rather than that it is the strongest available preconditioner at this size. We have not run 100M–1B-parameter pretraining or a controlled loss-spike and recovery study against AdamW, Shampoo and Muon. Accordingly, the contraction property and the small-model noise-regime sweeps motivate, but do not establish, improved recovery from instability at that scale.

What the benchmark can and cannot separate. Twelve tasks at five seeds resolve Evie-KF against Sophia at $p_{\mathrm{Holm}} = 0.0068$ and leave Lion and AdamW as near misses, so the design bounds what it can separate without implying equivalence. The scale-generalization result rests on two to three seeds at one model size, where Muon leads and Evie-KF leads the Adam family; the sweeps pin γ at a single value instead of searching it, which on finance is far from the value that task's own search selects; and the finance panels are one market period, four sectors, chosen by a stated rule before any results were seen.

I.3 Submission Checklist

Claims. The abstract and introduction state the three results the paper tests: the nesting of AdamW at $\gamma = 0$, the Kronecker-versus-diagonal mechanism, and the twelve-task standing. Each is supported by the section named beside it and no claim is made outside the tasks reported.

Theory. Every assumption is listed as (A1) through (A8) in Section 3 and restated in Appendix A; every result is proved in Appendix C as R1 through R10, with the correspondence to the main text's numbering given in a table there, and each is checked numerically in Appendix E.

Experiments. Tasks, models, batch sizes, step budgets, seeds, the metric per domain and the checkpoint-selection rule are in Section 4; the search grid, boundary rule and every frozen hyperparameter are in Appendix H; preprocessing, splits and the frozen ticker lists are there too. The aggregation and the test are stated before the results and applied to one metric per domain.

Code and data. The harness, the generated task files, the statistical pipeline and the per-cell records are released, with an anonymised copy available for review. The datasets are public; the finance panels are committed as frozen snapshots, so no download happens at run time.

Compute. Every cell ran on one NVIDIA Tesla T4 under PyTorch 2.10.0+cu128. Wall time and peak memory per arm and per domain are in Table 9, and the tuning budget per optimizer is in Table 11.

Ethics. The work involves no human subjects and no personally identifiable information; the statement above Section 6 gives the full position, and generative AI use is disclosed there.