How Much of the Natural Gradient's Value Do Cheap Approximations Retain?

Anonymous authors
Paper under double-blind review

Abstract

Cheap approximations of natural gradient descent (NGD), such as K-FAC, are justified by the argument that a direction close to the natural gradient inherits its benefits. We test this argument on small dense and convolutional networks whose exact Fisher is computable and strongly ill-conditioned. Every update is normalised to unit Fisher norm, damping and horizon are shared, and each method tunes its own step size. We score each method by its *retention*: the fraction of exact NGD's improvement over gradient descent that it achieves. Under this protocol, on one-hidden-layer networks with at most 106 parameters, K-FAC, EK-FAC and the exact diagonal retain 41%–71% of the training-loss improvement, and Kronecker approximations beat diagonal ones in every configuration tested. The angle to exact NGD ranks the approximations, but a GD-to-NGD interpolation at the same angle predicts only a third to a tenth of their retention. At a 60° Euclidean angle, error in the lowest-curvature decile keeps all of NGD's benefit, while error in the highest-curvature decile makes the update far worse than gradient descent; at a matched damped-Fisher angle this contrast largely disappears, so it mainly reflects the Fisher-norm step budget. On dense networks, most retention survives at exact NGD's step size. These results point to relative curvature fidelity as a promising design target. In this regime, the exact layer-wise block retains 115%, so much higher fidelity is achievable, at higher cost.

1 Introduction

Many ML models output a probability distribution and are fitted by maximising its likelihood. Examples include variational inference (Salimbeni et al., 2018), policy gradients (Kakade, 2002; Schulman et al., 2015), probabilistic boosting (Duan et al., 2020), and in general, any network with a distributional head (Johnson et al., 2016), including classifiers trained with cross-entropy. For these models, natural gradient descent (NGD) follows the geometry of the model's own distributions, which makes its direction invariant to the parameterisation (Amari, 1998). The natural gradient direction requires a solve with the Fisher information matrix F_w, which has one row and one column per parameter. In practice the (expensive) Fisher is replaced by a diagonal (Kingma & Ba, 2015), block-diagonal (Martens, 2020) or Kronecker-factored (Martens & Grosse, 2015; George et al., 2018) approximation, and the resulting direction is used in place of the natural gradient.

These approximations are motivated in part by the intuition that a good approximation of F_w produces a direction sufficiently close to the natural gradient to inherit its optimization benefits. We are not aware of prior work that directly measures this retention under a common directional protocol. Prior work has focused primarily on evaluating approximate natural-gradient methods as optimizers. Surprisingly, Benzing (2022) show that K-FAC can outperform exact natural-gradient methods, and argue that it is better understood as a first-order method on activations. Such end-to-end comparisons, however, conflate the effects of the update direction with those of its magnitude and other optimizer-specific factors. Thus, they cannot isolate how much of the observed outcome is attributable to the approximate direction itself. Therefore, we pose the following research question:

Under the same optimization protocol, how much of exact NGD's improvement over gradient descent does an approximation retain, and which properties of the approximation predict that retention?

To isolate the direction from confounds that arise at scale, we restrict our experiments to small networks whose exact Fisher is tractable and whose condition number exceeds 10^6. In these settings,

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.

exact NGD improves on gradient descent by 16%–61%, against at most 0.01% at $\mathrm{cond}(F_w) \approx 6 \times 10^2$ (Appendix C). We isolate the directional component of each update rule by using unit Fisher-norm directions, shared damping and horizon, and a step size retuned per method. We quantify performance via retention, defined as the fraction of exact NGD's improvement over gradient descent that a given method achieves. Computational cost per step is deliberately not held fixed, because it is the principal advantage of cheap approximations. Note that retention can exceed 1, because exact NGD is locally steepest in the Fisher metric but does not necessarily follow a geodesic or maximize finite-step objective improvement.

We make the following contributions in this work:

- *A measurement protocol.* A protocol and a score that separate the directional component of an approximation's update rule from its step size (§3). Both of its controls are necessary: without them, the correlation between angle and final loss changes sign ($\rho = -0.67$ to $+0.40$).
- *Four findings.* Under this protocol, Kronecker approximations retain about two-thirds of exact NGD's benefit and diagonal ones less than half (§4); the angle to exact NGD ranks them, but an interpolation at the same angle predicts far less retention (§5); at a fixed Euclidean angle, retention is strongly affected by the spectral location of the error, mostly through the Fisher-norm step budget (§6); and most retention survives at exact NGD's step size (§7).
- *A design target.* Relative curvature fidelity, supported by how it ranks approximations and by the exact layer-wise block's retention (§8).

2 PRELIMINARIES

This section establishes the definitions and assumptions that the rest of the paper relies on. Full context with respect to related work is provided in Appendix A.

Definition 1 (Distributional model)**.** *A* distributional model *is a differentiable map $f\colon \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^k$ that maps an input x and parameters w to the parameters $\eta = f(x, w)$ of a distribution $p(y \mid \eta)$.*

We write m for the number of parameters, N for the number of examples, k for the head dimension, and $n = kN$ for the total output dimension. Given training data $\{(x_j, y_j)\}_{j=1}^N$, the objective is to return w that minimizes the full-batch negative log-likelihood loss:

$$L(w) = \sum_{j=1}^{N} -\log p\big(y_j \mid f(x_j, w)\big).$$

Definition 2 (Fisher Information)**.** *The per-example output Fisher is*

$$F_{\eta_j} = \mathbb{E}_{y \sim p(\cdot \mid \eta_j)}\big[\nabla_\eta \log p(y \mid \eta_j)\, \nabla_\eta \log p(y \mid \eta_j)^\top\big].$$

The full output Fisher is $F_\eta = \bigoplus_{j=1}^N F_{\eta_j} \in \mathbb{R}^{n \times n}$. The stacked Jacobian is $J = \partial\eta/\partial w \in \mathbb{R}^{n \times m}$, the output gradient is g_η, and the parameter gradient is $g_w = J^\top g_\eta$. The parameter Fisher F_w is the Fisher of the composite model $w \mapsto p(y \mid f(x, w))$, summed over training inputs.

F_η has a closed form for the Gaussian, Bernoulli and categorical heads used here, and the Jacobians are computed in full. The resulting F_w is therefore the exact Fisher, not the empirical Fisher of Kunstner et al. (2019). This allows exact NGD to serve as a reliable reference. F_η and J are verified against finite-difference estimates to a relative tolerance 3×10^{-10}. Kronecker factors are also computed exactly. This is a deliberate control: it isolates the structural error of each approximation from the sampling noise that would arise if the factors were estimated from mini-batches.

Definition 3 (Natural gradient and approximate directions)**.** *For damping $\lambda > 0$ and any $P \succeq 0$, the three directions, all carrying the descent sign, are:*

$$d_{\mathrm{NG}} = -(F_w + \lambda I)^{-1} g_w, \qquad d_{\mathrm{GD}} = -g_w, \qquad d_P = -(P + \lambda I)^{-1} g_w.$$

Damping does not merely shrink the update. It *rotates* it toward gradient descent: the Euclidean angle between the two shrinks as λ grows (Lemma 1, Appendix B). Throughout this paper, "exact NGD" means exact *damped* NGD, and the damping must be shared across methods (Convention 1).

Table 1: Methods studied. All directions carry the descent sign.

Method	P or direction	Note
GD	$P = I$	Direction $\propto -g_w$
Exact diagonal	$P = \mathrm{diag}(F_w)$	
Square-root diagonal	$-\big(\mathrm{diag}(F_w) + \lambda I\big)^{-1/2} g_w$	Adam idealisation
Layer-wise block	Exact per-layer blocks of F_w	Reference; not cheap
K-FAC (exact factors)	$A_{\ell-1} \otimes G_\ell$ per layer	KFC on convolutions
EK-FAC	K-FAC eigenbasis, re-fitted eigenvalues	
Output-space	$-J^\top F_\eta^{-1} g_\eta$	Undamped
Exact NGD	$P = F_w$	

Definition 4 (NGD approximation)**.** *An NGD approximation is a rule $w \mapsto P(w) \succeq 0$ that replaces F_w in Definition 3. K-FAC, EK-FAC and the layer-wise block additionally use the network's layer structure. The methods studied are listed in Table 1.*

Throughout, *cheap approximations* refers to K-FAC, EK-FAC, the exact diagonal, and the square-root diagonal, an idealisation of Adam's preconditioner (Table 1).

Convention 1 (Shared damping)**.** *A single absolute damping $\lambda_t = \lambda_{\mathrm{rel}} \operatorname{tr}(F_{w_t})/m$ for m parameters, computed from the true Fisher at each iterate and with $\lambda_{\mathrm{rel}} = 10^{-4}$ unless stated otherwise, is applied to every method except output-space preconditioning, including exact NGD. Two methods incorporate it differently from Definition 3: K-FAC splits λ between its two Kronecker factors, and the square-root diagonal adds it inside the square root. Output-space preconditioning is left undamped.*

Because structured approximations absorb damping differently, a shared λ does not make their damped preconditioners equivalent: on convolutions, trace-matching K-FAC's factors alone moves its angle by 11° (Appendix C). We therefore treat the damping convention as part of the protocol rather than as a property of each approximation. By Lemma 1, unequal damping would change directions and not only step lengths, so sharing it is part of isolating the direction.

3 Measuring Retention

An approximation alters both the direction and the magnitude of a step. Attributing a difference in final loss to the directional component of the update rule therefore requires removing the length, giving every method its own best step from a common grid, and scoring the outcome.

Definition 5 (Normalised run)**.** *Given a direction rule $d(\cdot)$ evaluated at every iterate, a norm $\|\cdot\|_\bullet$, a step size α and a horizon T, a* normalised run *is*

$$w_{t+1} = w_t + \alpha \, \frac{d(w_t)}{\|d(w_t)\|_\bullet}, \qquad t = 0, \dots, T-1,$$

from the pair's shared initialisation w_0.

Convention 2 (Fisher-norm steps)**.** *All steps are normalised to unit norm under the undamped Fisher, $\|d\|_{F_w} = \sqrt{d^\top F_w \, d}$. Damping λ appears only in the preconditioner of Definition 3, not in the norm. Table 6 repeats the comparison under ℓ_2 normalisation.*

To second order, a unit Fisher-norm step fixes the KL divergence per step at $\alpha^2/2$ (Remark 1, Appendix B).

Convention 3 (Full-batch evaluation and common horizon)**.** *All directions are formed from the full training set, and every method runs for the same horizon T ($T = 300$ on dense networks unless stated otherwise).*

Convention 4 (Per-method step selection on a common grid)**.** *A single step size is held for all T steps of a run. Writing $w_T(\alpha, d)$ for the final iterate of the normalised run of Definition 5 with step α and direction rule d, each method selects*

$$\alpha^\star(d) = \arg\min_{\alpha \in \mathcal{A}} L\big(w_T(\alpha, d)\big),$$

separately for every dataset–seed pair. *The grid $\mathcal{A}$ holds* 16 *log-spaced values in* $[10^{-4}, 10^{0.5}]$*, so adjacent values differ by* $10^{0.3} \approx 2$*. A run whose final loss is not finite is marked diverged and excluded from the* $\arg\min$.

No $\alpha^\star$ falls on a boundary of $\mathcal{A}$ in any pair. Every optimum lies at least two grid points inside the grid, so no method's result is truncated. Appendix G reports the effect of a finer, 31-point grid.

Definition 6 (Retention)**.** *Write $L^\star(d) = L\big(w_T(\alpha^\star(d),\, d)\big)$ for the training loss at the final iterate of the run driven by d at its own tuned step (Convention 4), d_{GD} and d_{NG} included. Where exact NGD improves on GD, $L^\star(d_{\mathrm{GD}}) > L^\star(d_{\mathrm{NG}})$, the* retention *of d is*

$$R(d) = \frac{L^\star(d_{\mathrm{GD}}) - L^\star(d)}{L^\star(d_{\mathrm{GD}}) - L^\star(d_{\mathrm{NG}})}.$$

$R = 0$ is GD and $R = 1$ is exact NGD; $R < 0$ is worse than GD. $R > 1$ means d ends lower than exact NGD, which the condition above permits. Values above 1 do occur: exact NGD is locally steepest per unit Fisher norm, which makes it neither optimal over a finite step nor over a finite horizon, so it is a reference point rather than a ceiling: for products of location–scale families it follows a path strictly longer in the Fisher metric than the Fisher–Rao geodesic (Anonymous, 2026). Retention is reported as the median over pairs, with a percentile bootstrap 95% interval from 5000 resamples of the nine pairs. With nine pairs, these intervals describe resampling variability only. Retention is a property of an approximation under this protocol, not an intrinsic one: it depends on the damping, horizon, step grid and data (§4, Appendix G).

Retention is well posed only where exact NGD consistently outperforms GD. In every pair, the gap is 16%–61% of GD's loss (Convention 5, Appendix C).

Together, Convention 2 removes the length confound and Convention 4 provides each method its own step, both anchored to the loss at iterate T because $\alpha^\star$ is tuned to minimise $L(w_T)$. Because each direction also shapes the trajectory on which it is evaluated, retention measures the update rule as an optimiser, not a direction at fixed iterates.

Both controls are necessary for the angle to become informative. We blend a base direction toward exact NGD and correlate each run's angle with its final loss (Appendix D). With neither control, the correlation is negative ($\rho = -0.67$ and -0.29 from the two base directions), as if closing the angle hurt. With both, it turns positive ($\rho = +0.40$ and $+0.36$; weaker, $+0.08$ and $+0.06$, under Fisher normalisation). With one control only, it is null or negative.

Experimental setting. The default architecture is a dense network with one hidden layer of width 8, $\tanh$ activations, and a heteroscedastic Gaussian head ($k = 2$) with a scale floor, $\sigma = 0.05 + e^s$, below every dataset's noise level; without it, exact NGD gains training loss by collapsing σ on fitted points (Appendix M). Three regression datasets are used: the scikit-learn diabetes dataset (Efron et al., 2004) ($N = 442$, $m = 106$) and two synthetic datasets with homoscedastic and heteroscedastic noise respectively ($N = 500$, $m = 66$) (Appendix N). Each dataset is split $60/40$ into training and held-out sets; we score training loss, and held-out loss is reported in Appendix M. Because the ill-conditioned trajectories are numerically sensitive, reruns are not bit-reproducible: in one pair, individual angles move by up to 8° between runs, while medians move by at most 0.8°. The solves themselves are accurate; the sensitivity lies in the trajectory (Appendix C). We therefore report medians and orderings across pairs rather than individual runs. Convolutional networks are described in Appendix J.

4 How Much Cheap Approximations Retain

Under the experimental protocol of Section 3, we attempt to answer the question: *how much of exact NGD's improvement does each approximation retain?*

Finding 1. *Under the protocol of Section 3, K-FAC, EK-FAC and the exact diagonal retain* 41%–71% *of exact NGD's improvement over GD. Kronecker approximations outperform both diagonal approximations in every configuration tested.*

Retention levels. Table 2 reports the median retention over nine pairs. Output-space preconditioning scores -12.0% retention, performing worse than plain GD at the median.

The layer-wise block drops the same inter-layer blocks as K-FAC yet retains 114.7% against K-FAC's 66.9%, and sits at $9.2°$ from exact NGD against K-FAC's $58.4°$ (§5). This is consistent with a substantial contribution from the Kronecker factorisation within layers, rather than from the block-diagonal assumption alone.

Retention levels depend on the protocol; the ordering does not. Retention shifts by up to 22 percentage points across two decades of damping, and by up to a third of its value when the step grid is refined (Appendix G). Kronecker approximations stay ahead of both diagonals under every variation, although the order of EK-FAC and K-FAC is not stable. The gap even widens with scale. With a matrix-free reference up to $m = 6{,}362$, the exact diagonal's retention falls from 36.6% to 2.8%, while K-FAC and EK-FAC stay above 60% (Table 9).

Table 2: Retention of each approximation under the protocol of Section 3. Median over nine dataset–seed pairs (three datasets $\times$ three seeds), with bootstrap 95% confidence intervals.

Method	Retention (%)	95% CI
Layer-wise block	114.7	[91, 151]
EK-FAC	70.5	[61, 88]
K-FAC (exact factors)	66.9	[61, 84]
Exact diagonal	40.7	[25, 46]
Square-root diagonal	17.0	[9, 30]
Output-space	−12.0	[−69, −3]

Takeaway. If cheap approximations help because they approximate the natural gradient (Martens & Grosse, 2015; George et al., 2018), retention should rise as the angle to exact NGD falls.

5 The Angle to NGD Ranks Approximations

To test whether retention follows the angle between an approximation's direction and exact NGD, we define angles in the Euclidean, Fisher, damped Fisher and inverse-Fisher metrics (Definition 7). We first verify that reducing the angle to exact NGD improves retention along a controlled interpolation path, and then compare real approximations with that path's prediction. We then ask which metric first-order theory recommends (Proposition 1), and whether that metric ranks the approximations reliably.

Finding 2. *Measured along their own trajectories, the Euclidean angle ranks the approximations by retention (correct sign in all nine pairs; $\rho = -0.94$ pooled), but approximations retain three to ten times what an interpolation between gradient descent and exact NGD retains at the same angle. The Fisher-metric angles, singled out by a first-order identity, rank them almost as well (eight of nine pairs).*

Definition 7 (Angle between directions)**.** *For $M \succ 0$, the angle between directions u and v in the M-metric is*

$$\angle_M(u, v) = \arccos \frac{u^\top M v}{\|u\|_M \, \|v\|_M}.$$

Four metrics are used: Euclidean ($\angle_I$), Fisher ($\angle_{F_w}$), damped Fisher ($\angle_{F_w+\lambda I}$), and inverse damped Fisher ($\angle_{(F_w+\lambda I)^{-1}}$).

Each angle compares a method's direction with exact NGD's direction computed at the same parameters w (and, for the Fisher metrics, the same F_w). We take w from two trajectories. Along exact NGD's trajectory, every method is evaluated at the same iterates, giving a common reference. Along each method's own trajectory, w runs over that method's iterates, and we report the median angle over the run.

Measured angles. Table 3 reports the angles. Along exact NGD's trajectory, the cheap approximations are $47°$–$61°$ from exact NGD in the Euclidean metric. The layer-wise block, which also leads in retention (Table 2), is much closer, at $9.2°$. Along each method's own trajectory, the cheap approximations' Euclidean angles widen to $78°$–$85°$. Fisher angles are systematically smaller, and the two metrics drift in opposite directions during training (Appendix E).

Reducing the angle improves retention. If no other property of the direction varies, one might think that a smaller angle to exact NGD should improve retention. To test this, we rotate gradient descent's direction toward exact NGD's. With unit vectors $\bar{d}_{\mathrm{GD}}$ and $\bar{d}_{\mathrm{NG}}$ at angle Ω in a chosen

metric,

$$d_\beta = \frac{\sin\big((1-\beta)\Omega\big)}{\sin\Omega}\,\bar{d}_{\mathrm{GD}} + \frac{\sin(\beta\Omega)}{\sin\Omega}\,\bar{d}_{\mathrm{NG}},$$

so the angle to exact NGD is exactly $(1-\beta)\Omega$. We recompute d_β at every iterate, take unit Fisher-norm steps (Convention 2), and retune the step at each β. Under Euclidean interpolation, a smaller angle improves retention on average: the mean final training loss falls monotonically from 0.807 at $\beta = 0$ to 0.562 at $\beta = 1$, exact NGD is the best β in seven of nine pairs, and the rank correlation between angle and retention is $\rho = -0.906$. No individual run is monotone, however. Under Fisher interpolation, the relationship is weaker (Appendix E).

Approximations exceed the interpolation. Along the Euclidean interpolation path, retention stays above 70% out to $18°$, falls to about 50% near $40°$, and reaches zero at GD's $87°$. Evaluating the interpolation curve at each method's own-trajectory angle yields a predicted retention for that method. Every approximation exceeds it, by a factor of 3.5 to 9.8 (Table 4). The interpolation's error is GD's, which weights high-curvature directions, so the factor partly reflects GD's error profile.

Along exact NGD's trajectory, where angles are smaller, the factor narrows to 1.4–2.3 for all but the square-root diagonal (Appendix E). The angle is therefore not a sufficient statistic for comparing NGD approximations.

Table 3: Angles to exact NGD. "NGD traj." is the Euclidean angle along exact NGD's trajectory, median over all checkpoint measurements from the nine dataset–seed pairs. "Own traj." is measured along each method's own trajectory, in both the Euclidean and Fisher metrics: the mean over each run, then the median over the nine pairs. Layer block (layer-wise block).

	NGD traj.	Own trajectory	
Method	$\angle_I$	$\angle_I$	$\angle_{F_w}$
Layer block	$9.2°$	$51.9°$	$38.2°$
EK-FAC	$47.2°$	$79.1°$	$53.9°$
K-FAC (exact)	$58.4°$	$78.2°$	$53.0°$
Exact diagonal	$57.0°$	$83.1°$	$60.2°$
Square-root diagonal	$60.9°$	$84.7°$	$59.8°$
Output-space	$89.8°$	$85.3°$	$69.8°$
GD	$85.2°$	$86.8°$	$72.3°$

Which angle does first-order theory recommend? Under a unit Fisher-norm budget, first-order theory singles out the Fisher angle. The fraction of exact NGD's first-order decrease that a direction achieves is the cosine of its Fisher angle to the natural gradient (Proposition 1, Appendix B). Our reference, however, is damped. For a damped reference, the same identity holds in the damped Fisher metric (Corollary 1). The difference matters here. Exact NGD concentrates 51%–92% of its squared norm in the bottom decile of the spectrum, where the damping dominates. First-order theory therefore favours the damped Fisher angle.

Three angles rank the approximations. Within each dataset–seed pair, we rank six approximations by angle and by retention (Table 11, Appendix H). The Euclidean angle orders them consistently with retention in all nine pairs ($p = 0.004$, sign test). The damped Fisher angle, which Corollary 1 favours, does so in eight ($p = 0.04$), as does the undamped one; the inverse-Fisher angle does so in only six ($p = 0.51$). Pooled over approximations, the Euclidean and damped Fisher angles both correlate strongly with retention ($\rho = -0.94$), the undamped Fisher angle less so ($\rho = -0.71$, $p = 0.11$).

Table 4: Retention predicted by the Euclidean interpolation curve at each method's own-trajectory angle, against measured retention. Layer block (layer-wise block), Exact d. (exact diagonal), Sq.-root d. (square-root diagonal).

Method	Angle	Predicted	Measured
Layer block	$51.9°$	32.9%	114.7%
K-FAC	$78.2°$	9.6%	66.9%
EK-FAC	$79.1°$	8.7%	70.5%
Exact d.	$83.1°$	4.2%	40.7%
Sq.-root d.	$84.7°$	2.4%	17.0%

Takeaway. The angle to exact NGD predicts order but not magnitude. The Euclidean angle largely ranks the approximations by retention, yet every approximation retains three to ten times what the interpolation path predicts at its angle. The damped Fisher angle, which first-order theory favours, ranks them almost as well. An angle compresses an m-dimensional error into a single scalar, discarding where in the spectrum the error lies. Section 6 decomposes the error by spectral band to find which part of it determines retention.

6 WHERE THE ERROR LIES IN THE SPECTRUM

Two directions can sit at the same angle to exact NGD yet deviate from it along different eigendirections of F_w: for example, one along high-curvature eigendirections (large eigenvalues) and the other along low-curvature ones. We call this deviation a direction's *error*, and ask whether its location in the spectrum affects retention at a fixed angle, and whether it accounts for what the angle misses. To answer this, we first partition the spectrum into bands and show which spectral region each angle metric is most sensitive to (Definition 8, Lemma 2). We then construct directions whose Euclidean angle to exact NGD is held fixed while the spectral location of their error varies (Definition 10, Lemma 3). Finally, we ask which band discriminates between the approximations of Table 1.

Finding 3. *In our matched-angle construction, at a fixed Euclidean angle to exact NGD, error in the Fisher's lowest-curvature eigendirections has no measurable cost, whereas error in its highest-curvature eigendirections severely degrades retention. At* $60°$*, bottom-decile error yields retention* 106%*; top-decile error yields* -354% *(medians over nine pairs and five draws). At a matched damped-Fisher angle the contrast largely disappears (Appendix L).*

Definition 8 (Spectral bands)**.** *With* $\mu_1 \geq \cdots \geq \mu_m$ *the eigenvalues of* F_w *and* $u_1, \ldots, u_m$ *the corresponding eigenvectors, partition the index set* $\{1, \ldots, m\}$ *into three bands:* $\mathcal{B}_{\text{top}}$*, the indices of the* $\lfloor m/10 \rfloor$ *largest eigenvalues;* $\mathcal{B}_{\text{bot}}$*, the indices of the* $\lfloor m/10 \rfloor$ *smallest; and* $\mathcal{B}_{\text{mid}}$*, the remaining indices (*6 *top,* 54 *middle and* 6 *bottom at* $m = 66$*;* 10*,* 86 *and* 10 *at* $m = 106$*). The projector onto band* $\mathcal{B}$ *is* $\Pi_{\mathcal{B}} = \sum_{i \in \mathcal{B}} u_i u_i^\top$.

The nominal $10/80/10$ split is fixed in advance. It matches the band edges of the tangent-kernel analysis (Appendix K) and is held fixed across methods, pairs, dampings, widths and scales (Appendix I). The matched-angle result below does not depend on it (Appendix I).

Definition 9 (Band alignment and band energy)**.** *The* band alignment *of a direction* d *in band* $\mathcal{B}$ *is* $\cos_{\mathcal{B}}(d) = \cos \angle_I \big(\Pi_{\mathcal{B}} d,\ \Pi_{\mathcal{B}} d_{\text{NG}}\big)$*. The* band energy *of* d_{NG} *in* $\mathcal{B}$ *is* $\|\Pi_{\mathcal{B}} d_{\text{NG}}\|^2 / \|d_{\text{NG}}\|^2$.

All four angle metrics are diagonal in the eigenbasis of F_w, with weights 1, μ_i, $\mu_i + \lambda$ and $1/(\mu_i + \lambda)$. Each angle's inner product therefore splits into per-band sums (Lemma 2, Appendix B). The Fisher angle weights each eigendirection by μ_i. Exact NGD does the opposite: it scales each gradient component by $1/(\mu_i + \lambda)$. For every method except the layer-wise block (5.4%), the top decile therefore accounts for 75%–97% of the Fisher inner product, although exact NGD carries about 10^{-4} of its squared norm there (Table 12). In that band, every method except output-space preconditioning already aligns with exact NGD (cosines 0.96–0.98), so the Fisher angle is dominated by a band that barely separates the methods. Because the approximations differ from exact NGD in several correlated ways, we isolate spectral location by constructing directions at a fixed Euclidean angle to exact NGD whose error lies in a single band.

Definition 10 (Band-confined direction at a matched angle)**.** *Given a band* $\mathcal{B}$ *and a target angle* $\theta \in (0°, 90°)$*, let* $\hat{d}_{\text{NG}} = d_{\text{NG}} / \|d_{\text{NG}}\|$ *and let* e *be a unit vector in* $\operatorname{range} \Pi_{\mathcal{B}}$ *orthogonal to* $\Pi_{\mathcal{B}} \hat{d}_{\text{NG}}$*. The* band-confined direction *is* $d_{\mathcal{B},\theta} = \cos\theta\ \hat{d}_{\text{NG}} + \sin\theta\ e$.

We take e to be the normalised component of $\Pi_{\mathcal{B}} r$ orthogonal to $\Pi_{\mathcal{B}} \hat{d}_{\text{NG}}$, where r is a standard Gaussian vector in the eigenbasis of F_w, drawn once per run. Since eigenvector signs are arbitrary, e's direction within the band varies between steps.

By construction, $d_{\mathcal{B},\theta}$ lies exactly θ from exact NGD in the Euclidean metric. Outside $\mathcal{B}$, it agrees with exact NGD's direction (Lemma 3, Appendix B). The Fisher angle is not held fixed.

At a fixed angle, the band strongly affects retention. Table 5 reports retention at four matched angles, with error confined to each band. Bottom-decile error has no measurable cost (1.04–1.16, with no decline as θ grows); middle-band error degrades retention steadily; top-decile error is severe, reaching -3.06 at $60°$, far worse than GD. The bottom decile yields higher retention than the top in all nine pairs at every angle, and higher than the middle in all nine at $\theta \geq 45°$ ($p = 0.004$, Wilcoxon signed-rank), eight at $30°$ and seven at $15°$. The result replicates across five draws of e at $60°$: the bottom decile yields higher retention than the top in all 45 runs (pooled medians 1.06 and -3.54). Because the Fisher angle moves with the band, the effect is consistent with the Fisher-norm step budget: the same Euclidean error costs $3°$–$7°$ of Fisher angle in the bottom decile but $81°$–$88°$ in the top, and retention tracks the Fisher angle across all 108 runs ($\rho = -0.86$). Matching

the damped Fisher angle instead, which equalises first-order decrease per unit damped-Fisher norm (Corollary 1), removes most of the effect: up to 45° every band retains 0.93–1.19, and at 60° bottom-decile error is the costlier (0.69 against 0.96 at the top over five draws; Appendix L). The Euclidean-matched contrast therefore mostly reflects how the Fisher-norm budget prices high-curvature error. Across the six ranked approximations, band alignment spreads by 0.49 in the bottom band, 0.50 in the middle and 0.02 at the top: the approximations differ from exact NGD, and from each other, below the top decile. The Fisher angle therefore concentrates on a band that no longer separates the approximations (at $m \sim 10^2$; not at $m = 6{,}362$, Appendix I). Middle-band alignment ranks the six approximations consistently with retention in all nine pairs ($p = 0.004$, sign test), including where K-FAC overtakes EK-FAC (Appendix I). The Euclidean angle does equally well, so the band decomposition explains the ranking rather than improving on it.

Takeaway. At a fixed Euclidean angle, where the error lies matters as much as how large it is, but mostly because the Fisher-norm budget prices high-curvature error; at a fixed damped-Fisher angle, spectral location matters much less. The approximations already align with exact NGD in the top band and differ below it, where middle-band alignment ranks them as well as the Euclidean angle does.

Table 5: Retention under band-confined error at a matched Euclidean angle θ, with the realised Fisher angle below each band. Median over nine dataset–seed pairs.

Band		15°	30°	45°	60°
Bottom decile	Retention	1.16	1.04	1.04	1.10
	Fisher angle	3°	4°	4°	7°
Middle 80%	Retention	0.84	0.65	−0.54	−1.60
	Fisher angle	64°	71°	84°	82°
Top decile	Retention	0.04	−1.24	−2.28	−3.06
	Fisher angle	81°	85°	86°	88°

7 Retention Survives at Exact NGD's Step Size

Because each method tunes its own step size (Convention 4), we next ask: *is an approximation valuable because of where it points, or because it permits a step size that exact NGD cannot use?*

Finding 4. *Run at exact NGD's best step size, K-FAC and EK-FAC preserve between* 71% *and* 94% *of their retention at every dense horizon from* $T = 25$ *to* 300*. For these two methods, the step size decides the outcome only on convolutional networks at horizons up to* $T = 50$*, where exact NGD's loss diverges at step sizes they tolerate: they keep almost none of their retention at* $T = 20$ *and about a third at* $T = 50$*.*

We therefore also measure $R_{\mathrm{NG}}(d)$: the retention of d when it runs at exact NGD's best step instead of its own (Definition 12, Appendix B).

On dense networks, most retention is directional. On dense networks, exact NGD remains stable at all 16 grid points, and approximations typically select a step size approximately four times as large (two grid points higher on the grid). At exact NGD's step size, K-FAC and EK-FAC retain 76% and 71% of their retention at $T = 300$, positive across all nine dataset–seed pairs, with higher values at shorter horizons (Table 10, Appendix G); the exact diagonal retains 66%. The square-root diagonal depends most on its larger step size, declining from 0.17 to 0.05.

On convolutional networks, the step size decides short horizons. On convolutional networks (Appendix J), K-FAC and EK-FAC outperform exact NGD at short horizons: at $T = 20$ their retention reaches 1.26–2.29 (Table 13). The advantage comes from the step size. Exact NGD diverges at steps they tolerate, and its best step is $25\times$ smaller than K-FAC's. Held to exact NGD's step, their retention at $T = 20$ falls to between -0.11 and 0.15. By $T = 100$, they keep 69% and 73% of it, close to the dense values. The asymmetry replicates on a second architecture and under minibatches (Appendix J).

Takeaway. Cheap approximations lose most retention through their direction, not their step size, except on convolutional networks at short horizons. What property, then, should an approximation preserve?

8 A Design Target: Relative Curvature Fidelity

We propose relative curvature fidelity as a promising design target for NGD approximations, after first asking whether high fidelity is achievable in this regime at all.

Ill-conditioning does not preclude high fidelity. The anisotropy that makes exact NGD valuable persists across nineteen architectures and a critical-gain sweep, even at dynamical isometry (Saxe et al., 2014; Pennington et al., 2017) (Appendix K). Yet the exact layer-wise block lies at only 9.2° from exact NGD, aligns at 0.68 in the middle band (against 0.18–0.29 for the cheap approximations), and retains 114.7%. Substantially higher fidelity is therefore achievable without changing the geometry, albeit at higher computational cost. No Kronecker-cost method we tested reaches this level, but the layer-wise block can serve as the reference against which cheaper approximations are scored.

Definition 11 (Relative curvature fidelity)**.** *For $P \succeq 0$ and the shared damping λ (Convention 1), let P_λ be P damped as the method applies it ($P + \lambda I$ in Definition 3) and $F_\lambda = F_w + \lambda I$. The* relative curvature fidelity *of P with respect to F_w is*

$$\kappa_P = \operatorname{cond}\big(P_\lambda^{-1/2} F_\lambda \, P_\lambda^{-1/2}\big) = \frac{\lambda_{\max}(P_\lambda^{-1} F_\lambda)}{\lambda_{\min}(P_\lambda^{-1} F_\lambda)}.$$

$\kappa_P = 1$ if and only if $F_\lambda = cP_\lambda$ for some scalar $c > 0$: P captures every damped curvature ratio *correctly, and exact NGD scores $\kappa_P = 1$. Like the Fisher-normalised protocol, κ_P ignores the global scale of P_λ.*

Three properties support κ_P as a target (Appendix B). A relative guarantee on P bounds it (Proposition 2). It bounds the angle between d_P and exact NGD in P's metric, though only loosely here (Remark 2). An additive guarantee controls it only weakly, because Frobenius error hides errors in the smallest-eigenvalue directions (Remark 3). On our approximations, however, relative Frobenius error ranks the approximations about as well as κ_P.

Takeaway. Along exact NGD's trajectory, κ_P correlates negatively with retention in every dataset–seed pair ($9/9$, $p = 0.004$, sign test) and ranks the layer-wise block first and both Kronecker approximations ahead of both diagonals, as retention does; finer orderings are fragile ($7/9$ exact; one miss swaps K-FAC and EK-FAC, the other puts K-FAC above the layer-wise block). κ_P and spectral location answer different questions: GD, for instance, aligns well in the bottom band yet has the largest κ_P of any method. As a ratio of extreme eigenvalues, κ_P is set by the spectrum's ends, not its middle band. We suggest targeting relative curvature fidelity and using spectral location to identify the errors that matter.

9 Conclusion

In a controlled mechanistic study, we measured how much of exact NGD's improvement cheap approximations retain, under a protocol that separates the directional component of each update rule from its step size. K-FAC, EK-FAC and the exact diagonal retain two-fifths to seven-tenths of it, and Kronecker approximations stay ahead of diagonal ones. The angle to exact NGD ranks the approximations, but they retain three to ten times what a GD-to-NGD interpolation retains at the same angle. At a fixed Euclidean angle, where the error lies in the spectrum strongly affects retention, mostly through the Fisher-norm step budget. On dense networks, most of what the Kronecker approximations retain comes from their direction. These results suggest relative curvature fidelity as a promising design target. The exact layer-wise block retains 114.7%, so much higher fidelity is achievable, at higher cost.

Limitations. Our networks are small ($m \leq 1{,}666$), trained full batch without momentum, and scored on training loss. At $T = 300$ exact NGD is worse than GD on held-out loss in every pair, though early stopping brings all methods within 6%, so our results concern optimisation of the training objective; levels also depend on the head's variance floor (Appendix M). Retention at $m/n > 0.2$, where angles behave differently, is untested (Appendix F). Normalisation uses the exact Fisher, and our damping is below practical K-FAC settings. All statistics rest on nine pairs from three datasets, and the convolutional arm is a case study, not a test of external validity.

AI USE STATEMENT

We used a large language model (Claude, Anthropic) as a coding, analysis and writing assistant: to review experiment scripts, to check the paper's quantitative claims against the stored results, and to polish the text. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text, claims and artifacts.

ETHICS STATEMENT

This work uses only public datasets (the scikit-learn diabetes dataset, whose records are de-identified, MNIST and CIFAR-10) and synthetic data. It involves no human subjects and studies optimisation methods with no direct societal application. We are not aware of ethical concerns raised by this work.

REPRODUCIBILITY STATEMENT

The measurement protocol is specified in Section 3 (Conventions 1–4 and Definitions 5–6), the methods in Table 1, and all hyperparameters above. Every table is produced by a numbered script in the supplementary code that writes its data to CSV. Exact Fisher matrices, Jacobians and Kronecker factors are checked against finite differences (Section 2), and formal statements are proved in full (Appendix B). Reruns of the most ill-conditioned trajectories are not bit-reproducible across numerical-library configurations (Section 3).

REFERENCES

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

Anonymous. Suboptimality of natural gradient descent under a Fisher-chord budget. Companion paper under review; author details withheld for anonymous review, 2026.

Qinxun Bai, Steven Rosenberg, and Wei Xu. Generalized tangent kernel: A unified geometric foundation for natural gradient and standard gradient. *Transactions on Machine Learning Research*, 2025.

Frederik Benzing. Gradient descent on neurons and its link to approximate second-order optimization. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1817–1853, 2022.

Aleksandar Botev, Hippolyt Ritter, and David Barber. Practical Gauss-Newton optimisation for deep learning. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 557–565, 2017.

Felix Dangel, Frederik Kunstner, and Philipp Hennig. BackPACK: Packing more into backprop. In *International Conference on Learning Representations*, 2020.

Tony Duan, Avati Anand, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. NGBoost: Natural gradient boosting for probabilistic prediction. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2690–2700, 2020.

Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

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

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.

Roger Grosse and James Martens. A Kronecker-factored approximate Fisher matrix for convolution layers. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 573–582, 2016.

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

Karl E. Gustafson and Duggirala K. M. Rao. *Numerical Range: The Field of Values of Linear Operators and Matrices*. Springer, 1997.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Anas Jnini and Flavio Vella. Dual natural gradient descent for scalable training of physics-informed neural networks. *Transactions on Machine Learning Research*, 2025.

Anas Jnini, Flavio Vella, and Marius Zeinhofer. Gauss-Newton natural gradient descent for physics-informed computational fluid dynamics. *Computers & Fluids*, 307:106955, 2026. doi: 10.1016/j.compfluid.2025.106955.

Matthew J. Johnson, David Duvenaud, Alexander B. Wiltschko, Ryan P. Adams, and Sandeep R. Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems*, 2016.

Sham Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, 2002.

Ryo Karakida and Kazuki Osawa. Understanding approximate Fisher information for fast convergence of natural gradient descent in wide neural networks. In *Advances in Neural Information Processing Systems*, 2020.

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

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

Wu Lin, Felix Dangel, Runa Eschenhagen, Juhan Bae, Richard E. Turner, and Alireza Makhzani. Can we remove the square-root in adaptive gradient methods? A second-order perspective. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024.

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

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

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

Johannes Müller and Guido Montúfar. Geometry and convergence of natural policy gradient methods. *Information Geometry*, 7:485–523, 2024. doi: 10.1007/s41884-023-00106-z.

Yann Ollivier. Riemannian metrics for neural networks I: feedforward networks. *Information and Inference*, 4(2):108–153, 2015.

Kazuki Osawa, Yohei Tsuji, Yuichiro Ueno, Akira Naruse, Rio Yokota, and Satoshi Matsuoka. Large-scale distributed second-order optimization using Kronecker-factored approximate curvature for deep convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in Neural Information Processing Systems*, 2017.

Garvesh Raskutti and Sayan Mukherjee. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

Yousef Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, 2 edition, 2003.

Hugh Salimbeni, Stefanos Eleftheriadis, and James Hensman. Natural gradients in practice: Non-conjugate variational inference in Gaussian process models. In *International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 689–697, 2018.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

Jesse van Oostrum, Johannes Müller, and Nihat Ay. Invariance properties of the natural gradient in overparametrised systems. *Information Geometry*, 6(1):51–67, 2023. doi: 10.1007/s41884-022-00067-9.

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

Guodong Zhang, James Martens, and Roger Grosse. Fast convergence of natural gradient descent for over-parameterized neural networks. In *Advances in Neural Information Processing Systems*, 2019.

A Related Work

Natural gradient and its geometry. Amari (1998) introduced natural gradient descent and its efficiency argument, and Ollivier (2015) derives invariant training algorithms for neural networks from the same geometry. Raskutti & Mukherjee (2015) relate natural gradient to mirror descent, and van Oostrum et al. (2023) study which of its invariance properties survive in overparametrised models, the setting of our m versus n comparisons. Bai et al. (2025) place the natural and standard gradients in a common tangent-kernel framework. Martens (2020) surveys the method, including its relation to Gauss–Newton and the role of damping. Beyond network training, natural gradients are used in variational inference (Salimbeni et al., 2018), policy optimisation (Kakade, 2002; Schulman et al., 2015; Müller & Montúfar, 2024), probabilistic boosting (Duan et al., 2020) and physics-informed networks (Jnini & Vella, 2025; Jnini et al., 2026). A companion paper (Anonymous, 2026), included anonymised in the supplementary material, shows that idealised NGD need not follow the shortest path the Fisher metric admits, which bounds the regime in which exact NGD is the right reference (Section 3).

Structured approximations. Diagonal preconditioners are the cheapest stand-in, and Adam's second-moment accumulator (Kingma & Ba, 2015) is often read as the square root of a diagonal empirical Fisher, the reading our square-root diagonal idealises; Lin et al. (2024) argue that removing the square root strengthens this second-order reading. K-FAC (Martens & Grosse, 2015) approximates each layer block of the Fisher by a Kronecker product of activation and gradient statistics. KFC (Grosse & Martens, 2016) extends it to convolutions under assumptions that include spatially uncorrelated derivatives, and Eschenhagen et al. (2023) generalise the construction to linear layers with weight sharing, as in transformers, convolutional and graph networks. EK-FAC (George et al., 2018) keeps K-FAC's eigenbasis and re-estimates the eigenvalues in it. Distributed implementations made K-FAC practical for ImageNet-scale training (Osawa et al., 2019). Botev et al. (2017) give a practical block-diagonal Gauss–Newton method, and BackPACK (Dangel et al., 2020) computes exact curvature quantities of the kind we use. Shampoo (Gupta et al., 2018) and SOAP (Vyas et al., 2025) are not derived from the Fisher, although Shampoo's preconditioner can be read as a Kronecker approximation of the Gauss–Newton matrix (Morwani et al., 2025); we leave them to future work.

Evaluating approximations. Most evidence about these approximations compares optimisers end to end. Benzing (2022) finds that K-FAC outperforms exact natural-gradient methods, including where the block-diagonal structure K-FAC assumes is exact, and argues that K-FAC is better understood as a first-order method on activations. Section 7 reaches a compatible reading by measuring alignment and step size directly. Kunstner et al. (2019) show that the empirical Fisher is a poor substitute for the Fisher and attribute its behaviour to rescaling rather than curvature. On the theoretical side, Zhang et al. (2019) prove that NGD converges to a global minimum at a linear rate for over-parameterised networks, and that K-FAC does too under a stronger over-parameterisation requirement, and Karakida & Osawa (2020) identify conditions under which layer-wise approximations, including block-diagonal and K-FAC, retain exact NGD's fast convergence in wide networks. These analyses characterise when approximations converge as fast as exact NGD; none measures, under a fixed per-step budget, how much of exact NGD's improvement an approximation retains in a regime where it does not, which is the quantity this paper isolates.

Spectral structure and conditioning. Our anisotropy diagnostics are stated on the tangent kernel (Jacot et al., 2018). Dynamical isometry (Saxe et al., 2014; Pennington et al., 2017) controls the conditioning of the input–output Jacobian $\partial f/\partial x$. Our architecture sweep asks whether it transfers to the parameter Jacobian $\partial f/\partial w$ that determines F_w, and finds that it does not (Appendix K). The angle consequence of the relative guarantee in Section 8 is the Kantorovich inequality in its operator-trigonometric form (Gustafson & Rao, 1997). The matrix $P^{-1/2}F_wP^{-1/2}$ that defines κ_P is the standard measure of spectral equivalence in preconditioning theory (Saad, 2003). We apply it to approximate NGD and connect it to finite-horizon retention.

B Formal Statements and Proofs

Damping (Section 2).

Lemma 1 (Damping rotates the natural gradient toward the gradient)**.** *The Euclidean angle $\angle_I(d_{\mathrm{NG}}(\lambda), d_{\mathrm{GD}})$ is non-increasing in λ (strictly so unless g_w lies in a single eigenspace of F_w) and tends to 0 as $\lambda \to \infty$.*

Proof. Let (μ_i, u_i) be the eigenpairs of F_w and $g_i = u_i^\top g_w$, with $g_w \neq 0$. Then $u_i^\top d_{\mathrm{NG}} = -g_i/(\mu_i + \lambda)$ and $u_i^\top d_{\mathrm{GD}} = -g_i$. Let $p_i = g_i^2/\|g_w\|^2$ and $W = 1/(\mu_i + \lambda)$ with $i \sim p$, so

$$c(\lambda) = \cos \angle_I(d_{\mathrm{NG}}, d_{\mathrm{GD}}) = \frac{\mathbb{E}[W]}{\sqrt{\mathbb{E}[W^2]}}.$$

Since $dW/d\lambda = -W^2$, $\frac{d}{d\lambda}\mathbb{E}[W] = -\mathbb{E}[W^2]$ and $\frac{d}{d\lambda}\mathbb{E}[W^2] = -2\mathbb{E}[W^3]$, hence

$$c'(\lambda) = \mathbb{E}[W^2]^{-3/2}\big(\mathbb{E}[W]\,\mathbb{E}[W^3] - \mathbb{E}[W^2]^2\big) \geq 0$$

by Cauchy–Schwarz, $\mathbb{E}[W^2] = \mathbb{E}[W^{1/2}W^{3/2}] \leq \sqrt{\mathbb{E}[W]\,\mathbb{E}[W^3]}$. Equality requires W to be constant on the support of p, that is, every u_i with $g_i \neq 0$ to share one eigenvalue; otherwise $c' > 0$ and the angle is strictly decreasing. Finally, $\lambda W \to 1$ uniformly in i as $\lambda \to \infty$, so $c(\lambda) = \mathbb{E}[\lambda W]/\sqrt{\mathbb{E}[(\lambda W)^2]} \to 1$ and the angle tends to 0. □

The Fisher-norm budget (Section 3).

Remark 1 (A Fisher-norm budget)**.** *For a unit direction $\hat{d}$ with $\|\hat{d}\|_{F_w} = 1$,*

$$\mathrm{KL}\big(p_w \,\|\, p_{w+\alpha\hat{d}}\big) = \tfrac{1}{2}\alpha^2 + O(\alpha^3),$$

with the KL summed over training inputs. The budget the protocol equalises is therefore the Fisher norm itself, which is exact by construction. Its reading as a KL budget is accurate only to second order, and for a nonlinear model the true KL departs from $\frac{1}{2}\alpha^2$ at finite α.

The first-order identity (Section 5).

Proposition 1 (First-order identity)**.** *Let $F_w \succ 0$ and $d^\star = -F_w^{-1}g_w$. Then (1) over all d with $d^\top F_w d = 1$, the first-order decrease $-g_w^\top d$ is maximised by $d^\star/\|d^\star\|_{F_w}$, with maximum $\|d^\star\|_{F_w}$. (2) every such d achieves the fraction $-g_w^\top d \,/\, \|d^\star\|_{F_w} = \cos \angle_{F_w}(d, d^\star)$.*

Proof. Assume $g_w \neq 0$, so $d^\star \neq 0$. Since $F_w d^\star = -g_w$, every d satisfies $-g_w^\top d = (F_w d^\star)^\top d = \langle d, d^\star\rangle_{F_w}$. (i) By Cauchy–Schwarz in the F_w inner product, $\langle d, d^\star\rangle_{F_w} \leq \|d\|_{F_w}\|d^\star\|_{F_w} = \|d^\star\|_{F_w}$ whenever $\|d\|_{F_w} = 1$, with equality if and only if d is a positive multiple of $d^\star$, that is, $d = d^\star/\|d^\star\|_{F_w}$. (ii) For any such d, $-g_w^\top d/\|d^\star\|_{F_w} = \langle d, d^\star\rangle_{F_w}/(\|d\|_{F_w}\|d^\star\|_{F_w}) = \cos \angle_{F_w}(d, d^\star)$. □

Corollary 1 (Damped reference)**.** *Since $g_w = -(F_w + \lambda I)d_{\mathrm{NG}}$:*

$$-g_w^\top d = \langle d, d_{\mathrm{NG}}\rangle_{F_w} + \lambda\, d^\top d_{\mathrm{NG}} = \langle d, d_{\mathrm{NG}}\rangle_{F_w+\lambda I}.$$

Under $d^\top(F_w + \lambda I)d = 1$, d_{NG} is the maximiser and $\cos \angle_{F_w+\lambda I}(d, d_{\mathrm{NG}})$ is the achieved fraction.

Proof. Apply Proposition 1 with $F_w + \lambda I \succ 0$ in place of F_w; then $d^\star = -(F_w + \lambda I)^{-1}g_w = d_{\mathrm{NG}}$. □

Band decomposition and the matched construction (Section 6).

Lemma 2 (Band decomposition of the inner product)**.** *For any $M = \sum_{i=1}^m m_i\, u_i u_i^\top$ with $m_i > 0$, the M-inner product between a direction d and d_{NG} decomposes as*

$$\langle d, d_{\mathrm{NG}}\rangle_M = \sum_{\mathcal{B}} \sum_{i \in \mathcal{B}} m_i\, (u_i^\top d)(u_i^\top d_{\mathrm{NG}}).$$

The four angle metrics of Definition 7 correspond to: (1) Euclidean ($\angle_I$): $m_i = 1$; (2) Fisher ($\angle_{F_w}$): $m_i = \mu_i$; (3) damped Fisher ($\angle_{F_w+\lambda I}$): $m_i = \mu_i + \lambda$; (4) inverse Fisher: $m_i = 1/(\mu_i + \lambda)$.

Proof. Since $M = \sum_i m_i u_i u_i^\top$, $\langle d, d_{\mathrm{NG}}\rangle_M = d^\top M\, d_{\mathrm{NG}} = \sum_{i=1}^m m_i\, (u_i^\top d)(u_i^\top d_{\mathrm{NG}})$, and the bands partition $\{1, \ldots, m\}$. □

Lemma 3 (Angle and confinement of $d_{\mathcal{B},\theta}$)**.** *(1) $\angle_I(d_{\mathcal{B},\theta},\, d_{\mathrm{NG}}) = \theta$. (2) For every band $\mathcal{B}' \neq \mathcal{B}$, $\Pi_{\mathcal{B}'} d_{\mathcal{B},\theta} = \cos\theta\, \Pi_{\mathcal{B}'} \hat{d}_{\mathrm{NG}}$. The $\cos\theta$ factor cancels in the cosine ratio, so $\cos_{\mathcal{B}'}(d_{\mathcal{B},\theta}) = 1$.*

Proof. By Definition 10, $e = \Pi_{\mathcal{B}} e$ and $e^\top \hat{d}_{\mathrm{NG}} = e^\top \Pi_{\mathcal{B}} \hat{d}_{\mathrm{NG}} = 0$. (i) Hence $\|d_{\mathcal{B},\theta}\|^2 = \cos^2\theta + \sin^2\theta = 1$ and $d_{\mathcal{B},\theta}^\top \hat{d}_{\mathrm{NG}} = \cos\theta$, so $\cos\angle_I(d_{\mathcal{B},\theta}, d_{\mathrm{NG}}) = \cos\theta$; since $\theta \in (0°, 90°)$, the angle is θ. (ii) For $\mathcal{B}' \neq \mathcal{B}$, the projectors are orthogonal, so $\Pi_{\mathcal{B}'} e = \Pi_{\mathcal{B}'} \Pi_{\mathcal{B}} e = 0$ and $\Pi_{\mathcal{B}'} d_{\mathcal{B},\theta} = \cos\theta\, \Pi_{\mathcal{B}'} \hat{d}_{\mathrm{NG}}$. With $\cos\theta > 0$ and $\Pi_{\mathcal{B}'} d_{\mathrm{NG}} \neq 0$, this is a positive multiple of $\Pi_{\mathcal{B}'} d_{\mathrm{NG}}$, so $\cos_{\mathcal{B}'}(d_{\mathcal{B},\theta}) = 1$. □

Numerically, the constructed angle matches the target θ to within 10^{-13}, and band alignment in each unperturbed band equals 1 to machine precision.

Retention at exact NGD's step (Section 7).

Definition 12 (Retention at exact NGD's step)**.** *For the same dataset–seed pair and horizon, against the same GD and exact NGD references, the* retention at exact NGD's step *is*

$$R_{\mathrm{NG}}(d) = \frac{L^\star(d_{\mathrm{GD}}) - L\big(w_T(\alpha^\star(d_{\mathrm{NG}}),\, d)\big)}{L^\star(d_{\mathrm{GD}}) - L^\star(d_{\mathrm{NG}})}.$$

Relative curvature fidelity (Section 8). Proposition 2 and Remark 2 are stated for $\lambda = 0$, with damping they apply to P_λ and F_λ.

Proposition 2 (A relative guarantee bounds κ_P)**.** *If $\|P^{-1/2} F_w\, P^{-1/2} - I\|_2 \le \epsilon < 1$, then $\kappa_P \le (1+\epsilon)/(1-\epsilon)$.*

Proof. $A = P^{-1/2} F_w P^{-1/2}$ is symmetric, so $\|A - I\|_2 = \max_i |\nu_i - 1|$ over its eigenvalues ν_i. Hence every $\nu_i \in [1-\epsilon,\, 1+\epsilon]$ with $1 - \epsilon > 0$, and $\kappa_P = \nu_{\max}/\nu_{\min} \le (1+\epsilon)/(1-\epsilon)$. □

Remark 2 (κ_P bounds the angle in P's metric)**.** *For $P, F_w \succ 0$, let $d_P = -P^{-1} g_w$ and $d^\star = -F_w^{-1} g_w$. Then*

$$\cos\angle_P(d_P,\, d^\star) \;\ge\; \frac{2\sqrt{\kappa_P}}{1+\kappa_P}.$$

With damping, the same bound applies to $P + \lambda I$ and $F_w + \lambda I$. At the measured damped κ_P (1.7×10^3 for the layer-wise block to 4.5×10^4 for the square-root diagonal), the bound guarantees only angles below $87.2°$–$89.5°$.

Proof. Write $u = P^{-1/2} g_w$ and $A = P^{-1/2} F_w\, P^{-1/2} \succ 0$, so that $F_w^{-1} = P^{-1/2} A^{-1} P^{-1/2}$. Then

$$\langle d_P, d^\star\rangle_P = g_w^\top F_w^{-1} g_w = u^\top A^{-1} u, \qquad \|d_P\|_P^2 = g_w^\top P^{-1} g_w = \|u\|^2,$$
$$\|d^\star\|_P^2 = g_w^\top F_w^{-1} P F_w^{-1} g_w = \|A^{-1} u\|^2,$$

so $\cos\angle_P(d_P, d^\star) = u^\top B u / (\|u\|\, \|Bu\|)$ with $B = A^{-1}$. Let a and b be the smallest and largest eigenvalues of B, so that $b/a = \mathrm{cond}(A) = \kappa_P$. Since $(B - aI)(bI - B) \succeq 0$, $\|Bu\|^2 + ab\|u\|^2 \le (a+b)\, u^\top B u$, and by the AM–GM inequality $2\sqrt{ab}\, \|u\|\, \|Bu\| \le \|Bu\|^2 + ab\|u\|^2$. Combining the two, $u^\top B u / (\|u\|\, \|Bu\|) \ge 2\sqrt{ab}/(a+b) = 2\sqrt{\kappa_P}/(1+\kappa_P)$. The damped case is identical with P_λ and F_λ in place of P and F_w. □

Remark 3 (An additive guarantee controls κ_P only weakly)**.** *Consider $F_w = \mathrm{diag}(1, \delta)$ and $P = \mathrm{diag}(1, \delta + \epsilon)$. Then $\|P - F_w\|_{\mathrm{F}} = \epsilon$, but without damping $\kappa_P = (\delta+\epsilon)/\delta \to \infty$ as $\delta \to 0$. With damping, $\kappa_P = (\delta + \epsilon + \lambda)/(\delta + \lambda) \le 1 + \epsilon/\lambda$, which is large whenever $\epsilon \gg \lambda$: at $\lambda = 10^{-4}\, \mathrm{tr}(F_w)/m$, an error of 1% of the mean eigenvalue already gives $\kappa_P \approx 100$. In general, $\|P - F_w\|_2 \le \epsilon < \lambda$ gives only $\kappa_P \le (1+\epsilon/\lambda)/(1-\epsilon/\lambda)$. Frobenius error is insensitive to the smallest-eigenvalue directions, so an approximation can have small absolute error while being wrong in those directions by a factor of up to $1 + \epsilon/\lambda$ with damping, and without bound otherwise. On the approximations studied here, however, relative Frobenius error and κ_P rank the methods similarly (correct sign in $9/9$ pairs for both).*

C Protocol Details

Exact versus Monte Carlo Kronecker factors. Exact Kronecker factors isolate the *structural* approximation error from factor-estimation noise. They do not bound a deployed K-FAC's angle, however: with Monte Carlo factors, K-FAC's median angle along its own trajectory is larger in the Euclidean metric ($79.7°$ against $78.2°$) but smaller in the Fisher metric ($48.1°$ against $53.0°$). K-FAC with Monte Carlo factors enters only the within-pair rankings of Section 5. Its retention is lower: 58.6% [53, 78] against 66.9% [61, 84] with exact factors, with overlapping intervals. (Without the variance floor of Section 3 the order was reversed, 45.5% against 35.4%.)

Convention 5 (Regime of validity)**.** *Retention divides by $L^\star(d_{\mathrm{GD}}) - L^\star(d_{\mathrm{NG}})$, so it is well posed only where that gap is large relative to the losses it separates. In every pair the gap is at least a sixth of $L^\star(d_{\mathrm{GD}})$, the nine reported pairs spanning 16% to 61% of it. Where the gap is instead near zero, two nearly equal losses set the ratio and their last significant figure decides its value; on the linear models of Appendix C the gap is at most 0.01% of $L^\star(d_{\mathrm{GD}})$, and we report no retention there.*

The low-condition comparison. The linear-model figure in Convention 5 comes from a separate arm holding the design matrix fixed and varying only the model: η linear in w against the same $\tanh$ network, under a Bernoulli head and a heteroscedastic Gaussian head, on the diabetes design with three seeds, $T = 300$ and the same 16-point step grid. The linear models reach $\mathrm{cond}(F_w) = 5.7 \times 10^2$ and 6.2×10^2 respectively, and exact NGD improves on GD by 0.00% and 0.01%; the same heads behind the network reach $\mathrm{cond}(F_w)$ of 2.4×10^6 and 6.5×10^5 and improvements of 55% and 60% (medians over seeds).

Solve accuracy. Damping caps $\mathrm{cond}(F_w + \lambda I)$ at $1 + 10^4 m$ (at most 4.6×10^5 measured, against $\mathrm{cond}(F_w)$ up to 4.5×10^{10}), and along exact NGD's trajectory the float64 solve agrees with a 50-digit solve to a relative 10^{-12}. Non-reproducibility comes from the trajectory instead: perturbing w_0 by a relative 10^{-12} moves exact NGD's direction at $T = 300$ by more than $70°$ in two of nine pairs.

Trace matching on convolutions. On convolutions we additionally trace-match the Kronecker product, because the shared absolute damping is not scale-invariant and an uncorrected product receives an effectively larger λ. That correction changes K-FAC's angle by $11.0°$, larger than the entire K-FAC/EK-FAC gap.

D Protocol Ablation

Construction. We blend a base direction $d_0 \in \{d_{\mathrm{GD}},\ -J^\top F_\eta^{-1} g_\eta\}$ with exact NGD,

$$d_\beta = (1 - \beta)\, d_0 + \beta\, d_{\mathrm{NG}}, \qquad \beta \in \{0, 0.1, \ldots, 1\},$$

and vary two design choices independently: the *mixing rule* (blend the raw directions, or blend unit vectors) and the *step-size rule* (inherit the step tuned at $\beta = 0$, or retune on the grid at every β). The 2×2 gives *raw* (neither control), *normalised* and *retuned* (one each), and *both*, which applies the two controls of §3. For each arm we report Spearman's ρ between the angle to exact NGD and the final training loss, with a percentile bootstrap 95% interval from 5000 resamples clustered on the nine dataset–seed pairs.

Why the raw arms are censored. Table 6's first row rests on a censored sample, and the reason is the confound the ablation is designed to isolate. The step inherited from $\beta = 0$ is tuned for d_{GD}, whose norm is far smaller than exact NGD's: at initialisation $\|d_{\mathrm{NG}}\|/\|d_{\mathrm{GD}}\| = 28$, 90 and 286 on the three datasets, against $\|d_{\mathrm{NG}}\|/\|d_0\| \in [0.2, 1.7]$ for output-space preconditioning. Raw mixing therefore inflates the step by up to two orders of magnitude as β grows, and the run diverges: 43 of 99 runs are lost, and two pairs keep only four values of β. What survives is a censored sample in which the largest β are missing precisely where the step is worst. From output-space preconditioning, whose norm is already comparable to exact NGD's, 89 runs complete and the negative association is measured on a nearly complete sample. The divergence is thus not an artifact but the underlying mechanism: without normalisation, β moves the length and the direction simultaneously.

Table 6: Protocol ablation. Spearman ρ between the angle to exact NGD and final training loss, with cluster-bootstrap 95% intervals over nine pairs. The expected positive association appears only when both controls are active. n is the number of surviving runs out of 99; the raw arms lose 43 (from GD) and 10 (from output-space) to divergence (see text).

Mixing	Step size	From GD	From output-space
raw	inherited	-0.67 $[-0.81, -0.50]$ †	-0.29 $[-0.51, -0.10]$ ‡
unit	inherited	-0.12 $[-0.21, +0.00]$	-0.26 $[-0.45, -0.05]$
raw	retuned	$+0.02$ $[-0.20, +0.36]$	$+0.03$ $[-0.08, +0.33]$
unit	retuned	$\mathbf{+0.40}$ $[\mathbf{+0.32}, \mathbf{+0.75}]$	$\mathbf{+0.36}$ $[\mathbf{+0.31}, \mathbf{+0.63}]$

†$n = 56$ of 99; ‡$n = 89$ of 99; the remaining runs diverged.

Sensitivity to the choice of norm. This ablation mixes and normalises in the Euclidean norm and uses damping $10^{-6}\,\mathrm{tr}(F_w)/m$, whereas the rest of the paper uses the Fisher norm and $\lambda_{\mathrm{rel}} = 10^{-4}$ (Conventions 1 and 2). The choice of norm influences the effect size. When the *both* arm is repeated with mixing directions normalised in the Fisher norm instead of the Euclidean one, the association remains positive but weak, and its interval from GD includes zero: $\rho = +0.08$ $[-0.04, +0.29]$ from GD and $+0.06$ $[+0.01, +0.25]$ from output-space preconditioning, compared with $+0.40$ and $+0.36$ above. The qualitative conclusion this ablation supports, that the sign of the angle–loss association depends on which controls are active and therefore the protocol must be fixed in advance, is robust to this change. The effect sizes in Table 6 are specific to the Euclidean-norm setting of this ablation. The angle–loss relationship under the main protocol is the subject of §5.

E Angles Along the Exact-NGD Trajectory

The Euclidean angle depends on how the model is parameterised, whereas a first-order identity singles out a Fisher-type angle (Appendix H); we therefore report both. Table 7 measures every method against the same reference along the same trajectory, with exact Kronecker factors and shared damping. In the Fisher metric the angles grow during training, by $+30.8°$ for K-FAC against $-5.7°$ for the exact block: a measurement taken only at initialisation, as is common, sees K-FAC at $36.2°$ and misses almost half of its eventual error. On the Euclidean axis the drift runs the other way (K-FAC $76.1° \to 56.7°$, the exact diagonal $88.7° \to 53.2°$, EK-FAC $73.1° \to 42.2°$), every cheap approximation moving *closer* to exact NGD as training proceeds. The two metrics disagree about the *sign* of the drift, which is why both are carried side by side rather than one being quoted as the angle.

Table 7: Angle to exact damped NGD, degrees, measured **along an exact-NGD trajectory**: initialisation plus five later checkpoints, three datasets × three seeds (46 measurements per method). $\angle_I$ is the median over all checkpoints, comparable with the $\angle_{F_w}$ all column; *init* and *trained* split the Fisher column by checkpoint and *growth* is their difference. IQR in brackets. These are not the angles an approximation shows along its *own* trajectory, which are larger (Table 3).

method	$\angle_I$	$\angle_{F_w}$ init	$\angle_{F_w}$ trained	$\angle_{F_w}$ all	growth
layer-wise block	9.2	24.5	18.8	$\mathbf{19.5}$ $[15.7, 26.7]$	-5.7
EK-FAC	47.2	34.6	47.8	44.9 $[34.8, 52.6]$	$+13.1$
exact diagonal	57.0	42.1	59.1	58.1 $[45.9, 66.4]$	$+17.0$
square-root diagonal	60.9	44.0	80.1	78.8 $[68.3, 81.7]$	$+36.1$
K-FAC (exact factors)	$\mathbf{58.4}$	36.2	67.0	62.6 $[51.0, 71.2]$	$+30.8$
GD	85.2	48.6	84.3	83.6 $[72.1, 85.8]$	$+35.7$
output-space preconditioning	89.8	47.2	90.3	89.8 $[87.3, 91.0]$	$+43.2$

The Euclidean interpolation curve. Figure 1 shows the curve from which Table 4 is read. Each point on the left is one value of $\beta \in \{0, 0.1, \ldots, 1\}$, refined inside $[0.9, 1]$: the retention of the interpolated direction d_β, run as an optimiser at its own tuned step, against its mean Euclidean angle to exact NGD over that run, each a median over the nine pairs. Retention stays above 95% below $2°$, is about 73% at $18°$ and 50% near $40°$, and reaches zero at GD's $87°$. The five approximations'

own-trajectory angles (Table 3) lie at 52°–85°, where the curve predicts 2%–33%; they measure 17%–115%.

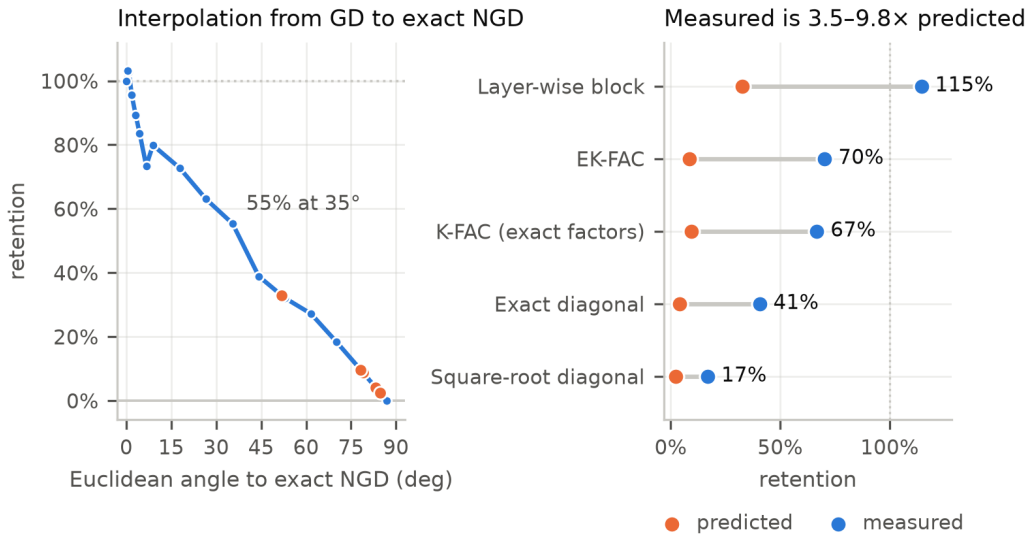

Figure 1: Retention predicted by the angle, against retention measured. *Left:* the Euclidean interpolation from GD ($\beta = 0$) to exact NGD ($\beta = 1$) of Section 5: retention of d_β against its Euclidean angle to exact NGD, median over nine dataset–seed pairs at each β. Orange points mark the five approximations' own-trajectory angles on the curve; their heights are the predicted retentions. *Right:* predicted (orange) against measured (blue) retention for each approximation, as in Table 4; measured values are those of Table 2.

Interpolation in the Fisher metric. Under Fisher interpolation (Section 5), the rank correlation between the Fisher angle and retention is $\rho = -0.270$, and exact NGD is the best β in none of the nine pairs. At $\beta = 0.1$ the Euclidean angle already falls to 3.6° and retention jumps to 88%, while the Fisher angle is still 41°. Figure 2 breaks both arms down by pair. Closing the swept angle raises retention in every pair of the Euclidean arm but in only six of nine pairs of the Fisher arm, whose association is weaker and more variable: its per-pair correlations average -0.31 against the Euclidean arm's -0.95. Dropping the endpoints $\beta = 0$ and $\beta = 1$, which could carry a trend on their own, leaves the Euclidean arm almost unchanged (-0.93) and weakens the Fisher arm to -0.17.

Predictions along exact NGD's trajectory. Reading the Euclidean interpolation curve of Section 5 at the angles of Table 7 predicts higher retention than at own-trajectory angles. The underestimate narrows to 1.4–2.3× for the Kronecker approximations, the layer-wise block and the exact diagonal, while the square-root diagonal falls below its prediction (0.6×). The discrepancy therefore persists on both trajectories for all but the square-root diagonal, though own-trajectory angles amplify it.

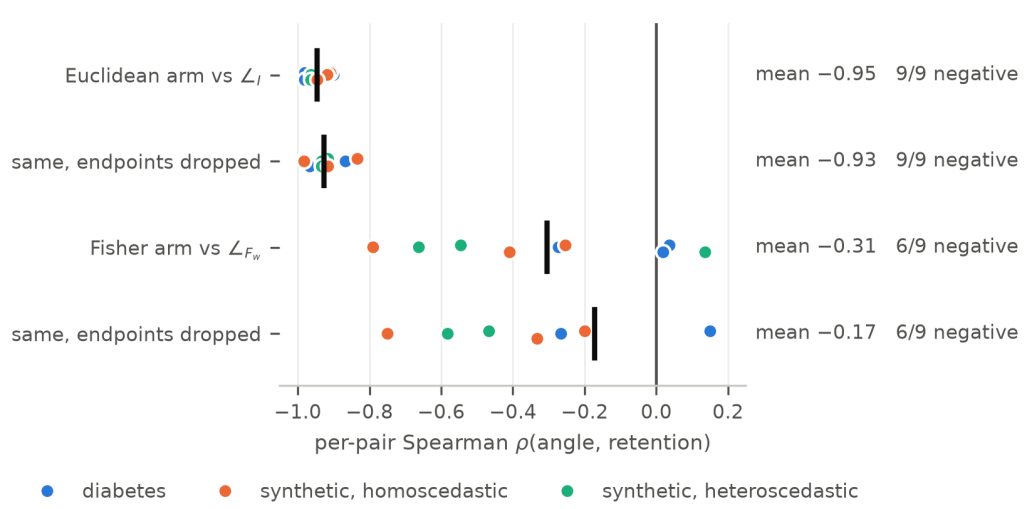

Figure 2: Angle against retention within each dataset–seed pair, for the two interpolation arms. Each point is one pair's Spearman correlation between the angle an arm sweeps (Euclidean $\angle_I$ or Fisher $\angle_{F_w}$) and retention across the eleven values of β, or the nine interior values when the endpoints $\beta = 0, 1$ are dropped; negative values mean that closing the angle raises retention. Vertical bars mark the mean over the nine pairs; the right-hand column gives that mean and the number of pairs with a negative correlation.

F Width Sweep

The experiments with dense networks use one hidden layer of width 8 ($m = 106$). To test whether the angles depend on width, we repeat the measurement of Appendix E at hidden widths 8, 16, 32, 64 and 128 ($m = 106$ to 1,666), on the same three datasets and three seeds, at $\lambda_{\text{rel}} = 10^{-4}$. Each angle is the Fisher-metric angle to exact NGD at the last iterate of the exact-NGD trajectory (iteration 300, or earlier if exact NGD finds no descent step).

Table 8: Fisher-metric angle to exact NGD at the last iterate of the exact-NGD trajectory, median over nine dataset–seed pairs.

Method	$m = 106$	$m = 210$	$m = 418$	$m = 834$	$m = 1{,}666$
K-FAC (exact factors)	73.8°	58.4°	74.2°	81.8°	71.7°
EK-FAC	51.4°	53.5°	68.6°	81.0°	70.1°
Exact diagonal	64.1°	65.4°	76.0°	85.6°	84.9°
Layer-wise block	17.9°	28.6°	31.9°	6.4°	8.8°

No cheap approximation moves closer to exact NGD as the network widens: K-FAC stays between 58° and 82°, and EK-FAC's and the exact diagonal's angles rise, to 70°–86° at widths 64 and 128. Only the layer-wise block closes in, to 6°–9°. Widening the network, however, also raises m/n, the number of parameters per Fisher row ($n = 2N_{\text{train}}$, 530–600 here), from 0.11–0.20 at width 8 to 1.71–3.14 at width 128, and at widths 64 and 128 exact NGD reaches an interpolating fit and stops after 22–69 iterations, so those columns are read at a different stage of training.

Separating width from m/n. We hold each width fixed and vary the number of samples instead. At hidden widths 16, 32, 64 and 128 ($m = 130$ to 1,026), on the two synthetic datasets with three seeds each, N is chosen so that m/n takes the same eleven values from 0.05 to 2 at every width ($N = 54$ to 8,567; the widest network starts at $m/n = 0.1$ to keep $N \leq 10{,}000$). Everything else is as above. Figure 3 shows the result.

The four widths trace similar curves. Over the ten m/n values shared by all four widths, m/n accounts for 83%–94% of the variation in the median angle of K-FAC, EK-FAC and the layer-wise

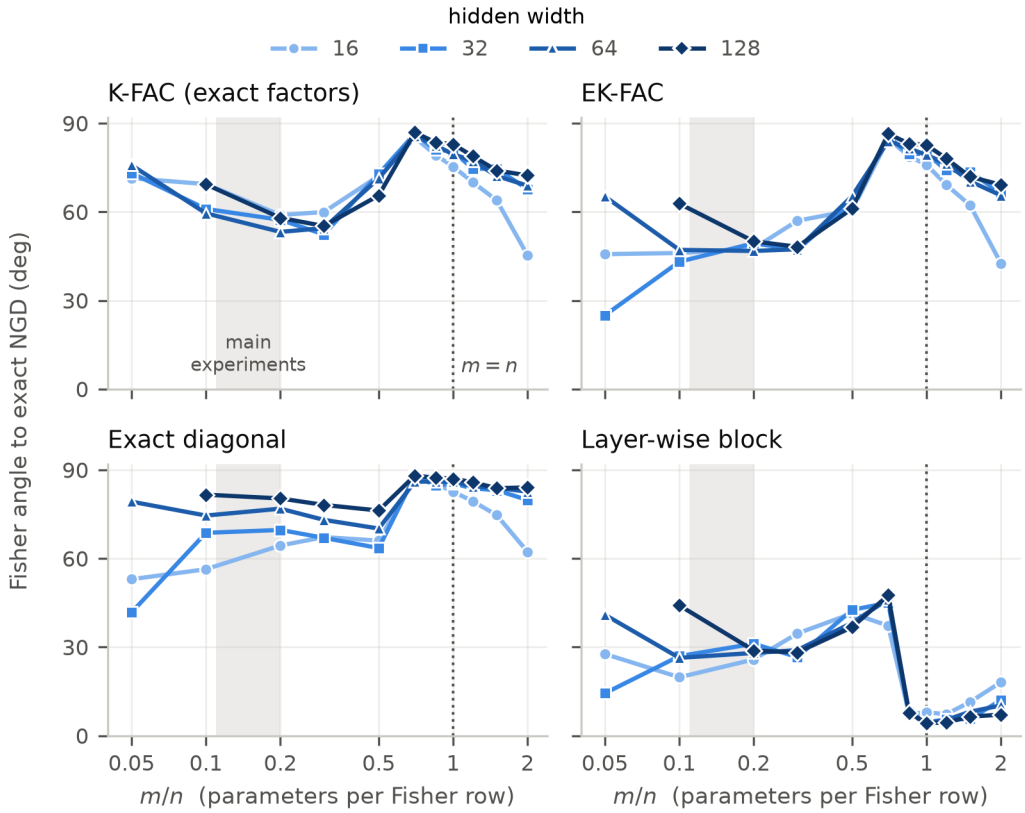

Figure 3: Fisher-metric angle to exact NGD at the last iterate of the exact-NGD trajectory, against m/n, with the width held fixed and the number of samples varied. Each line corresponds to a single hidden-layer width; each point is the median over six dataset–seed pairs (two synthetic datasets $\times$ three seeds). The shaded band marks the m/n of the main dense experiments (0.11–0.20); the dotted line marks $m = n$, above which F_w is rank-deficient.

block, and width for at most 3%; for the exact diagonal the split is 63% and 22%. Up to $m/n = 0.5$ the cheap approximations sit at 25°–82°. At $m/n = 0.7$ they jump to 84°–88° and stay above 62° at every larger m/n, while the layer-wise block falls below 10° from $m/n = 0.85$. The jump coincides with exact NGD reaching an interpolating fit before the horizon (median stopping iteration 20–55 from $m/n = 0.85$, against 300 below $m/n = 0.5$). Without the variance floor of Appendix M, the same sweep showed the cheap approximations' angles *collapsing* between $m/n = 0.3$ and 0.85; that collapse came from runs dominated by variance collapse and does not survive the floor.

For this paper, what matters is where the experiments sit. The main dense experiments ($m/n = 0.11$–0.20), the scale ladder of Appendix G ($m/n = 0.2$) and the convolutional arm ($m/n_{\text{eff}} = 0.217$) all lie below $m/n = 0.5$, where every exact-NGD run completes its horizon; the overparameterised regime behaves differently and is left open.

G Robustness of Retention

Damping and grid density. The intervals in Table 2 are wide and adjacent rows are not separable: K-FAC's $[61\%, 84\%]$ overlaps EK-FAC's $[61\%, 88\%]$. The levels move by up to 22 points across two decades of damping (K-FAC reads 71.7%, 66.9%, 49.5% at $\lambda_{\text{rel}} = 10^{-5}, 10^{-4}, 10^{-3}$) and by up to a third of their own value between the 16- and 31-point grids, because densifying the grid also improves the denominator. The roughly 40%/67%/71% figures hold within this protocol at this damping. The ordering is what generalises beyond it. The order of EK-FAC and K-FAC in particular

is unstable: EK-FAC leads in $5/9$ pairs at the reported damping, $7/9$ and $6/9$ at the neighbouring two, and $2/9$, $5/9$ and $5/9$ across the 8-, 16- and 31-point grids.

Scale. Damping and grid density are perturbations of the protocol; *scale* is a larger change, and retention is expensive at size because the exact reference recurs at every iterate of every step-size trial. We replace the dense $O(nm^2)$ build and $O(\min(m,n)^3)$ solve with Jacobi-preconditioned conjugate gradients on Fisher-vector products, which need only the action of F_w. That extends the range of parameter counts to $m = 6362$, $3.8\times$ the largest dense arm and $60\times$ the smallest. A reduced protocol ($T = 50$, the same 16-point grid, one synthetic generator, three seeds, m/n pinned at 0.2) is run identically at every parameter count, making the comparison internal. The reference is validated against the dense pipeline at $m = 106$, agreeing to 10^{-9} in final loss and reproducing dense retention to three decimals. Angles alone, which need no step-size grid, reach $m = 35{,}640$ at the same fixed m/n (30 exact-NGD steps, one to three seeds per rung): K-FAC's and EK-FAC's Euclidean angles rise from 64° and 59° at $m = 106$ to 74° and 71°, and the diagonal's from 81° to 89°, while all three Fisher angles stay between 33° and 43°.

Table 9: Retention against m under one reduced protocol run identically at every rung. Levels are **not** comparable with Table 2 ($T = 300$, three datasets); the comparison is across rungs. All 16 grid points stay finite for every approximation at every rung, and the loss gap between GD and exact NGD *grows* with m ($0.43 \to 0.70$), so the trend is not a shrinking-denominator artefact.

approximation	$m = 106$	$m = 922$	$m = 6362$	change
K-FAC (exact factors)	72.4%	74.9%	61.4%	-11.0
EK-FAC	63.0%	78.2%	**68.3%**	$\mathbf{+5.3}$
exact diagonal	36.6%	10.5%	2.8%	$\mathbf{-33.8}$
Kronecker ahead of diagonal	3/3	3/3	3/3	**9/9**

The separation holds at every parameter count. Retention for the Kronecker approximations stays between 61% and 78% across the range, while the exact diagonal decreases by a factor of thirteen, monotonically. In these runs the gap therefore *widens* rather than eroding, and it is the diagonal that degrades. This is evidence that the qualitative separation survives scale, on one generator and three rungs. Because m/n is pinned at 0.2 here, the ladder stays in the regime of the main experiments, below the m/n at which exact NGD starts to reach an interpolating fit before the horizon (Appendix F). Scale moves the levels and leaves the ordering alone, the same verdict damping and grid density return, now on the quantity the headline claims rest on.

Retention at exact NGD's step. Table 10 gives the full split of Section 7 on dense networks.

Table 10: Retention at each method's own best step size and at exact NGD's best step size, on dense networks at horizons $T = 50$ and $T = 300$. Median over nine dataset–seed pairs. Fractions quoted in the text are medians of per-pair ratios, not ratios of these medians. Layer block (layer-wise block), Exact d. (exact diagonal), Sq.-root d. (square-root diagonal).

	$T = 50$		$T = 300$	
Method	Own	NGD's	Own	NGD's
Layer block	0.97	0.87	1.15	0.72
EK-FAC	0.63	0.57	0.70	0.46
K-FAC (exact)	0.66	0.55	0.67	0.52
Exact d.	0.31	0.20	0.41	0.23
Sq.-root d.	0.12	0.07	0.17	0.05

H The First-Order Identity for a Damped Reference

A first-order identity singles out a Fisher-type angle. Stating it precisely is important here because the reference is *damped*. Maximising the first-order decrease $-g_w^\top d$ subject to $d^\top F_w d = 1$ gives

$d \propto -F_w^{-1} g_w$, and the fraction of that maximum achieved by any unit-Fisher-norm d is exactly $\cos \angle_{F_w}\big(d, -F_w^{-1} g_w\big)$. This is an identity about the *undamped* direction, and we verify it numerically to 2.9×10^{-13}. Under damping $g_w = (F_w + \lambda I) d_{\mathrm{NG}}$, so

$$-g_w^\top d \;=\; \langle d, d_{\mathrm{NG}} \rangle_{F_w} \;+\; \lambda\, d^\top d_{\mathrm{NG}}, \tag{1}$$

introducing an additional Euclidean term. This term is not a rounding artifact. The direction d_{NG} carries 51% to 92% of its energy in the bottom decile of the spectrum, where λ dominates μ_i at $\mathrm{cond}(F_w) \gtrsim 10^8$. The corresponding statement for a damped reference uses the *damped* metric. Constraining $d^\top (F_w + \lambda I) d = 1$ makes d_{NG} itself the maximiser, and $\cos \angle_{F_w + \lambda I}(d, d_{\mathrm{NG}})$ is then the exact fraction of the attainable decrease. We report both angles throughout. Pooled over methods, the damped Fisher angle yields $\rho = -0.94$ ($p = 0.005$) and the undamped one $\rho = -0.71$ ($p = 0.11$); the Euclidean angle yields $\rho = -0.94$ and the inverse-Fisher angle $\rho = -0.37$. Within pairs (Section 5), the damping correction has no effect on the count: both Fisher angles rank correctly in $8/9$ pairs (Table 11).

Table 11: Pairs in which the angle and retention rankings of six approximations are positively correlated, with two-sided sign-test p-values. I-, D- and UD-Fisher are the inverse damped, damped and undamped Fisher metrics. The six approximations are the layer-wise block, K-FAC with exact and with Monte Carlo factors, EK-FAC and both diagonals. The test uses only the sign of each within-pair correlation.

Angle metric	Pairs	p
Euclidean	$9/9$	0.004
I-Fisher	$6/9$	0.51
D-Fisher	$8/9$	0.04
UD-Fisher	$8/9$	0.04

I Band-Analysis Details

Table 12: Alignment with exact damped NGD *by spectral band* of F_w (bottom decile, middle 80%, top decile), along each method's own trajectory, median over the nine dataset–seed pairs. *F-weight on top* is the share of the *Fisher* angle's numerator contributed by the top decile. Exact NGD's own direction carries less than 10^{-4} of its energy in that band. The spread row covers the six ranked approximations; output-space preconditioning and GD are shown for reference.

method	cos top	cos mid	cos bot	F-weight on top	retention
layer-wise block	0.968	**0.681**	0.617	**0.054**	114.7%
EK-FAC	0.973	0.285	0.132	0.751	70.5%
K-FAC (exact factors)	0.979	0.222	0.293	0.817	66.9%
K-FAC (Monte Carlo factors)	0.964	0.275	0.260	0.761	58.6%
exact diagonal	0.966	0.176	0.297	0.879	40.7%
square-root diagonal (Adam-style)	0.973	0.194	0.351	**0.970**	17.0%
spread across approximations	0.016	**0.504**	**0.485**	—	—
output-space preconditioning	0.861	0.251	0.802	0.962	−12.0%
GD	0.977	0.404	0.923	0.973	0%

The decomposition is fixed. The $10/80/10$ split is a fixed coarse decomposition, chosen to match the band edges of the tangent-kernel spectrum analysis and applied unchanged to every approximation, every damping, every width and every scale reported here. Every band edge is set in advance of any retention measurement and held common across approximations and pairs. We make no claim that $10/80/10$ is optimal, and a fitted split would almost certainly separate better. The evidence that it is not merely descriptive is that the same fixed decomposition transfers to settings whose retention ordering is reversed (*Generalisability* below). Nor does the matched-angle result depend on the split. Rerunning the $60°$ construction with $5/90/5$, $10/80/10$ and $20/60/20$ bands, bottom-band error outperforms top-band error in $9/9$ pairs under every split, and top-band retention is negative in every pair (medians 1.05, 1.10, 1.00 for the bottom band against -3.58, -3.06, -3.07 for the top).

Reweighting the angle does not repair it. The repair the band analysis suggests, weighting by $1/(\mu_i + \lambda)$ where NGD's energy actually is, is not supported by Table 12. The bottom band is high-energy but undiscriminating in the *other* direction: GD aligns there at 0.923, better than any cheap approximation, because where $\mu_i \approx 0$ the two differ by a near-constant factor rather than in direction, so a metric concentrated at the bottom favours the method that retains nothing. Nor is any weighting meaningfully better at prediction: ranking the six ranked approximations within each pair, identity scores $9/9$, inverse-Fisher $6/9$, damped Fisher $8/9$. The failure is of the premise that any single angle determines how much an approximation retains, not of one weighting, and what separates the three is where their weight sits, which is visible only band by band.

Generalisability across settings and scale. The middle-band result generalises. In three held-out settings ($\lambda_{\text{rel}} = 10^{-5}, 10^{-3}$ and width 32, the last one where K-FAC outperforms EK-FAC), middle-band alignment ranks the approximations correctly in $7/9$, $9/9$ and $9/9$ pairs, and the Euclidean angle in $9/9$ in each. The top band's lack of discriminative power is specific to the baseline setting, with correct rankings of $2/9$, $5/9$ and $9/9$ across the same three settings. Extending the decomposition to $m = 6362$ separates a general finding from a local one. The Fisher angle's concentration on the top decile strengthens with m ($0.69 \to 0.93$ for K-FAC), indicating that the misplacement of weight is persistent. However, the top decile ceases to be the band where approximations agree. Alignment there falls from about 0.97 to 0.15–0.26 for every approximation, while the Kronecker approximations' middle-band alignment remains stable ($0.38 \to 0.40$ for K-FAC, $0.46 \to 0.47$ for EK-FAC). The misplacement is the persistent finding. The particular explanation offered for it is specific to $m \sim 10^2$.

J The Convolutional Arm

Architecture and angles. The convolutional arm uses three 3×3 convolutions with mean pooling, a global average pool and a dense read-out to $k = 10$ logits, with a categorical head, $N = 800$, $m = 1506$–1614, $m/n_{\text{eff}} = 0.217$, and $\text{cond}(F_w)$ median 2.3×10^8, on MNIST and CIFAR-10. The cheap approximations sit $64°$–$87°$ from exact NGD. K-FAC's angle is the most architecture-invariant of any approximation ($+6.0°$ against the dense arm) while the exact diagonal moves $+34.5°$ to GD-like angles; a matched dense-versus-convolutional control at equal m/n isolates that collapse to *weight sharing* rather than off-diagonal structure, since the off-diagonal Frobenius mass of F_w is 0.99 in both.

Where Kronecker factorisation is accurate. K-FAC on a convolution is KFC (Grosse & Martens, 2016), which discards cross-location terms. The discarded mass is almost the whole block: the cross-location fraction of the exact layer block is 0.995, 0.990 and 0.978 on the three convolutions. On the dense read-out, where the assumption is vacuous, K-FAC's relative error against the exact block is 0.019, against 0.62–0.79 on the convolutions. Kronecker factorisation is 33–42 times more accurate on the read-out than on these convolutions, so a last-layer or adapter-only probe measures K-FAC precisely where it is almost exact, and reports a small angle that says nothing about the network.

Table 13: Retention against horizon T on the convolutional arm, each approximation's own best over the step-size grid at that T. Mean over two seeds per dataset. Values above 1.0 beat exact damped NGD. † marks cells compressed by the MNIST floor, which should be read as "both reach the floor".

	CIFAR-10			MNIST		
T	K-FAC	EK-FAC	diagonal	K-FAC	EK-FAC	diagonal
20	**1.26**	**1.80**	0.02	**1.69**	**2.29**	−0.03
50	0.81	**1.07**	0.01	**1.18**	**1.30**	0.08
100	0.61	0.78	0.01	0.98†	1.00†	0.16

Replication on a second architecture. What replicates on a second architecture (a flat $(8, 8, 8)$ channel profile against the baseline's widening $(6, 8, 12)$, matched m/n) is the asymmetry the mechanism rests on: exact NGD keeps three of six step sizes on *both*, and K-FAC's winning step exceeds

it in every pair. What does not replicate is the magnitude or the overtaking: the grid resolves the headroom to one or two steps ($25\times$ in three of four pairs, $5\times$ in the fourth), and on the flat profile neither approximation overtakes. A milder version appears on the dense arm: within 2° of exact NGD, 15 of 27 interpolated runs outperform it, eight of them with a larger step (median $2.0\times$ over the fifteen); under the protocol of Table 10 exact NGD stays finite at every grid step there, so on that arm a larger step is a better choice rather than one exact NGD is denied.

Minibatches. Re-forming every method's direction from a minibatch leaves the step-size ratio unchanged at every batch size, while from $B = 800$ to $B = 64$ exact NGD's final loss rises by 1.17 NLL against K-FAC's 0.53. The exact solve is about twice as noise-fragile as the approximation of it.

K TANGENT-MAP ANISOTROPY

The same quantity that makes the exact direction valuable also makes it hard to approximate: $\mathrm{cond}(J^\top J)$ has geometric mean 1.3×10^5 over nineteen designs, and the anisotropy is a broad tilt rather than a few extreme modes. Figure 4 shows the tangent-kernel spectrum of five of these designs. In each, the eigenvalues drop steeply over the first few percent of modes and then decline steadily across the rest, so the middle 80% of modes carries 42%–51% of the spectrum's log-range against the top decile's 29%–42%.

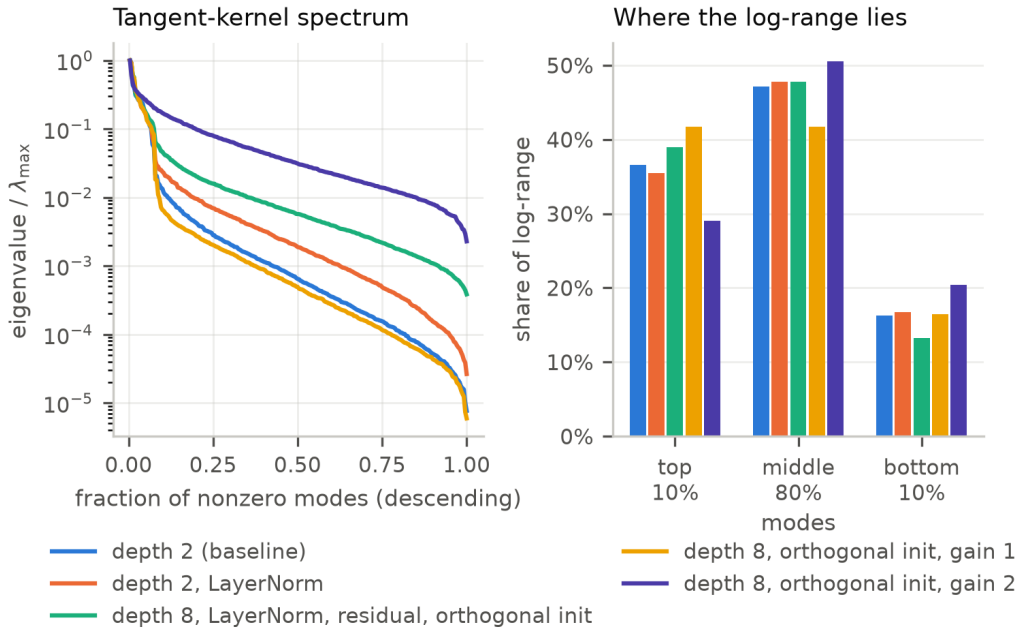

Figure 4: Tangent-kernel spectrum at initialisation for five of the nineteen designs, on the first 120 diabetes samples (width 32, $\tanh$). *Left:* nonzero eigenvalues of $JJ^\top$ (equivalently of $J^\top J$), divided by the largest and sorted in descending order, against their position as a fraction of the nonzero spectrum. *Right:* the share of the spectrum's total $\log_{10}$ range spanned by the top 10% of modes, the middle 80% and the bottom 10%, with colours as on the left.

An interpretation from this is that a correction repairing only a small number of extreme spectral modes offers limited scope for improvement. However, this does not preclude the effectiveness of alternative low-rank corrections. Architecture changes also do not eliminate the anisotropy. Across nineteen variants, $\mathrm{cond}(\mathbf{J}^\top \mathbf{J})$ varies over three orders of magnitude, yet the measured angle remains within a narrow 55.3°–74.9° band. Furthermore, a critical-gain sweep indicates that dynamical isometry in $\partial f/\partial x$ does not imply a well-conditioned $\partial f/\partial w$ (Saxe et al., 2014; Pennington et al., 2017).

L BAND-CONFINED ERROR AT A MATCHED DAMPED-FISHER ANGLE

Construction. Section 6 holds the Euclidean angle fixed, so the Fisher angle moves with the band. Here we hold the damped Fisher angle fixed instead. In the coordinates $\sqrt{\mu_i + \lambda}\, u_i^\top d$, in which the damped Fisher metric is Euclidean, we build the perturbation of Definition 10, confined to band $\mathcal{B}$ and orthogonal to d_{NG}, and map the result back. The damped Fisher angle to exact NGD is then exactly θ (verified to 10^{-6}), the error outside $\mathcal{B}$ is zero, and by Corollary 1 every such direction achieves the same fraction $\cos\theta$ of exact NGD's first-order decrease per unit damped-Fisher norm. Steps are still normalised in the undamped Fisher norm (Convention 2) and the step is retuned per run, so the equality holds up to that per-run scale.

Table 14: Retention under band-confined error at a matched *damped-Fisher* angle θ, with the realised Euclidean angle below each band. Median over nine dataset–seed pairs (compare Table 5).

Band		15°	30°	45°	60°
Bottom decile	Retention	1.13	0.99	0.99	0.62
	Euclidean angle	41°	60°	72°	81°
Middle 80%	Retention	1.17	0.94	0.93	0.90
	Euclidean angle	12°	25°	41°	53°
Top decile	Retention	1.19	1.03	0.93	0.97
	Euclidean angle	0.3°	0.6°	1.2°	2.7°

Result. Up to $45°$, every band retains 0.93–1.19 of exact NGD's improvement and no band is consistently better than another (top above bottom in 4–6 of nine pairs). At $60°$, bottom-decile error retains 0.62 against 0.97 for top-decile error (top ahead in $8/9$ pairs, $p = 0.055$). Across five draws of e at $60°$, top-decile error retains more than bottom-decile error in 39 of 45 runs (pooled medians 0.69, 0.87 and 0.96 for the bottom, middle and top bands). With r drawn once in parameter space and projected onto the band at every step, so that e does not flip with the arbitrary eigenvector signs, top-decile error beats bottom-decile error in 8 or 9 of nine pairs at every angle (bottom 0.49–0.73, top 0.95–1.07). The realised Euclidean angle now moves with the band instead ($41°$–$81°$ for bottom-band error, under $3°$ for top-band error), and retention correlates only weakly with either the undamped Fisher angle ($\rho = -0.20$) or the Euclidean angle ($\rho = -0.24$).

Interpretation. The Euclidean-matched result of Section 6 therefore mostly measures how the Fisher-norm step budget prices error: error in high-curvature directions consumes the budget. Once first-order decrease is equalised, spectral location matters much less, and what remains makes low-curvature error, where exact NGD carries most of its norm, the costlier.

M HELD-OUT LOSS AND THE VARIANCE FLOOR

Variance collapse without a floor. With $\sigma = e^s$ and s clipped at -6 in the forward pass (but not in the derivative), exact NGD lowers the training loss by shrinking σ on fitted training points rather than by fitting: its training mean-squared error is worse than GD's (0.66 against 0.53, medians over pairs), 7% of its training points sit at the clip, and flooring σ at 0.05 after training halves its training-loss advantage over GD. Its held-out NLL then exceeds 10^3 in seven of nine pairs, and in that setting retention mostly measured how aggressively each method collapsed the variance. All results in this paper therefore use $\sigma = 0.05 + e^s$, below the smallest noise level of any dataset (about 0.11 after standardisation), with a derivative that respects the clip. At the final iterate no method has any training point with σ at the floor.

Held-out loss with the floor. Table 15 comes from a rerun of the protocol that records held-out NLL at every iterate. At the final iterate of each method's train-selected run, exact NGD is worse than GD in all nine pairs (median ratio 5.6), and the held-out ranking largely reverses the training ranking. This is mostly a matter of timing. Every method's best held-out iterate sits at a training loss of 0.98–1.04, far above the 0.56–0.76 it reaches at $T = 300$, so by the horizon every method, GD included, has overfitted, and the faster ones further. With the step and the stopping iterate chosen

Table 15: Held-out NLL under four selection rules, median over nine dataset–seed pairs. The last three use the held-out split for selection and are optimistic by construction.

	GD	Sqrt d.	Exact d.	K-FAC	EK-FAC	Exact NGD	Layer b.
Final iterate, train-selected step	1.27	1.46	1.54	2.96	1.81	5.13	12.14
Step selected on held-out, $T = 300$	1.12	1.13	1.13	1.14	1.13	1.21	1.15
Early stopping, train-selected step	1.10	1.10	1.11	1.13	1.11	1.19	1.14
Best step and iterate	1.09	1.10	1.10	1.10	1.09	1.15	1.12

on held-out loss, all methods come within 6% of each other (1.09–1.15). A small residual remains: exact NGD's best held-out loss is still above GD's in eight of nine pairs (median ratio 1.04), as is the layer-wise block's, while K-FAC and EK-FAC match GD (better in five and four of nine pairs). The overfitting is in the mean as well as the scale: at the final iterate exact NGD's held-out mean-squared error is 0.83 against GD's 0.63, and its held-out NLL with a single shared σ is 1.46 against 1.22. Our results therefore concern optimisation of the training objective; under matched early stopping the methods generalise about equally, and exact NGD's direction brings no held-out benefit in this setting.

N Reproducibility Details

Hardware. All experiments run on CPU, on a single Apple M4 Pro machine (12 cores, 24 GB RAM), with Python 3.13, NumPy 2.5, SciPy 1.18, scikit-learn 1.9 and pandas 2.3; no GPU is used. Each dense experiment at $m \leq 1{,}666$ runs in minutes. The matrix-free retention runs at $m = 6{,}362$ take hours per pair, dominated by the conjugate-gradient solve for exact NGD at every iterate of every step-size trial.

Datasets. The diabetes dataset (Efron et al., 2004) has 10 features, which we standardise. The two synthetic datasets draw $N = 500$ inputs $x \sim \mathcal{U}[-1, 1]^5$ and set $y = 1.5\,x_1 + 0.7\,x_3^2 + \varepsilon$ with $\varepsilon \sim \mathcal{N}\big(0, \sigma(x)^2\big)$ and $\sigma(x) = \exp(s\,x_2)$, where $s = 0$ gives the homoscedastic dataset and $s = 1.5$ the heteroscedastic one. The synthetic data are drawn once, with seed 0, and every target is standardised.

Hyperparameters. Dense networks have one hidden layer of width 8 (widths 16–128 in Appendix F), $\tanh$ activations and a heteroscedastic Gaussian head with $\sigma = 0.05 + e^s$ (Appendix M); seeds 0, 1 and 2 set the initialisation and the $60/40$ split. The shared damping is $\lambda_{\mathrm{rel}} = 10^{-4}$ (10^{-5} and 10^{-3} in Appendix G), the step grid holds 16 log-spaced values in $[10^{-4}, 10^{0.5}]$ (31 in Appendix G), and the horizon is $T = 300$ ($T = 25$–300 in Section 7). Monte Carlo K-FAC samples 16 targets. Angles along exact NGD's trajectory are measured at iterations 0, 25, 50, 100, 200 and 300 of an exact-NGD run with backtracking (initial step 1, halved up to 30 times until the loss decreases). The interpolation uses $\beta \in \{0, 0.1, \ldots, 1\}$, refined inside $[0.9, 1]$; the matched-angle construction uses $\theta \in \{15^\circ, 30^\circ, 45^\circ, 60^\circ\}$, with five draws of r at 60°. Convolutional networks use six log-spaced step sizes in $[10^{-3}, 10^{0.5}]$ and $T \in \{20, 50, 100\}$. Matrix-free references solve for exact NGD by Jacobi-preconditioned conjugate gradients to relative tolerance 10^{-10}.

Scripts. Every table is produced by a numbered script in the supplementary code that writes its data to CSV: retention (`exp29`); angles along exact NGD's trajectory (`exp16`), the width sweep (`exp30`) and the m/n sweep (`exp54`, `exp55`); the interpolation (`exp19`, `exp31`, run with `MCNG_LAM=1e-4`); the protocol ablation (`exp18`); band alignment (`exp41`, `exp43`); matched-angle bands (`exp50`, `exp53`) and their damped-Fisher counterpart (`exp56`); the step-size split (`exp51`); the convolutional arm (`exp36`–`exp40`, `exp47`); relative fidelity (`exp16`, `exp52`); solve accuracy and trajectory sensitivity (`exp57`); held-out loss (`exp29`, `exp59`) and its diagnosis (`exp58`); the scale ladder (`exp48`); the angular pilot (`exp42`); and the tangent-kernel spectrum (`exp13`, `exp22`). The figures are drawn from these CSVs by `exp22` (Figure 4), `exp33` (Figure 1), `exp34` (Figure 2) and `exp55` (Figure 3). Every derived statistic in the text (medians, intervals, rank tests, read-offs) is recomputed from the CSVs by `paper_numbers.py`, and `run_floor.sh` reruns every dense experiment.