THE SECOND LAYER OF LINEAR SELF-ATTENTION IMPLEMENTS A CONTEXT-ADAPTIVE UPDATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work interprets transformer depth as iterations of an optimisation algorithm. We demonstrate that context-adaptive computation first occurs at layer two in linear self-attention. We prove that one-layer models cannot implement context-adaptive updates, that two-layer models can, and that, for every finite depth, there are task distributions where context-adaptive updates achieve strictly lower risk than any composition of the same number of fixed optimisation steps. These results identify the first layer where linear self-attention can condition its update rule on the observed context rather than only the training distribution. Depth therefore enables more than additional iterations of a fixed optimisation algorithm. We introduce a framework for measuring update behaviour in trained transformers by comparison with reference optimisation algorithms. The framework is validated on models with known mechanisms and applied across a range of transformer architectures and in-context learning tasks. One-layer models match the best single fixed optimisation step. Two-layer models show a context-adaptive correction, which disappears when the second layer is ablated. Our framework identifies the known two-layer induction transition from the update kernel in associative recall and extends to in-context classification and softmax attention.

1 INTRODUCTION

A line of work interprets transformers trained on prediction tasks as executing optimisation algorithms in the forward pass (Garg et al., 2022; Akyürek et al., 2023; von Oswald et al., 2023). Prior work has shown that, for linear self-attention (LSA), one layer implements a preconditioned gradient-descent step (Ahn et al., 2023; Mahankali et al., 2024; Zhang et al., 2024), and that deeper models are commonly interpreted as performing additional iterations. Furthermore, recent work has shown that deeper models exhibit context-dependent optimisation behaviour (Fu et al., 2024; Vladymyrov et al., 2024; Dragutinović et al., 2025). These results leave open when context-dependent preconditioning first becomes representable and when trained models begin to exhibit such updates. We characterise both transitions.

For linear self-attention, the transition between fixed and adaptive occurs at depth two. We extend the optimality analysis of Mahankali et al. (2024) and Zhang et al. (2024), showing that no one-layer LSA can implement a context-adaptive update for any parameter setting. We prove that, on an explicit family of task distributions, the optimal unrestricted two-layer LSA has strictly lower risk than every predictor composed of two fixed preconditioned gradient steps, so two layers cannot be reduced to two optimisation steps. More generally, for every finite depth there are task distributions on which context-adaptive updates achieve strictly lower risk than any composition of that many fixed optimisation steps.

1. **Theory.** A representation theorem shows that a single-layer LSA cannot implement a context-adaptive preconditioner. A separation theorem shows that an unrestricted two-layer LSA can achieve strictly lower risk than any composition of two fixed preconditioned steps, with a corollary extending the bound to any finite depth. An identification result shows that a positive context-conditional gap cannot be produced by any fixed preconditioned update.
2. **Measurement framework.** We introduce a nested oracle ladder of fixed-preconditioner gradient-descent predictors. Comparing held-out error with the ladder measures how many

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.

fixed optimisation steps a trained model implements. A context-adaptive one-step predictor reaches the Bayes floor, providing a reference for models that use context-dependent updates. The adaptive share and context-conditional gap are validated on simulated models with known mechanisms.

3. **Empirical transition and interventions.** Single-layer LSA behaves as a fixed bilinear update and matches $\mathcal{O}_1^{\mathrm{fix}}$. Two-layer LSA exhibits an adaptive correction. At $d \leq 16$, this correction is well described by a context-dependent natural-gradient family and not by any fixed preconditioner. Ablating the second layer removes the measured adaptive computation. Width, parameter-count and layer-graft controls localise the effect to the co-adaptation of the two layers.
4. **Transfer and scope.** The transition persists across input dimensions, task families and distribution shift, although the natural-gradient candidate is supported only at $d \leq 16$. The measurements also recover context-adaptive computation in in-context classification, softmax attention and associative recall, including the known two-layer induction transition of Olsson et al. (2022).

2 Related work

Transformers as in-context optimisation. A line of work interprets transformer in-context learning as an optimisation procedure. The controlled regression setting of Garg et al. (2022) established a standard framework for studying this behaviour. Akyürek et al. (2023) and von Oswald et al. (2023) showed that trained transformers can recover classical estimation algorithms. For linear self-attention, subsequent work characterised one-layer solutions and optimisation dynamics in terms of preconditioned gradient descent (Ahn et al., 2023; Mahankali et al., 2024; Zhang et al., 2024). These analyses have been extended to correlated designs (Li et al., 2024) and to the high-dimensional limit (Lu et al., 2025). Other work has considered context-dependent optimisation. This includes Newton-like updates in deeper transformers (Fu et al., 2024), noise-adaptive preconditioning (Vladymyrov et al., 2024), and context-dependent learning rates in softmax attention (Dragutinović et al., 2025). The algorithm-selection results of Bai et al. (2023) further characterise the algorithms that transformers can implement. On the task families considered here, a fully context-adaptive natural-gradient predictor attains the Bayes floor.

These results leave open when context-dependent preconditioning first becomes representable and when trained models first exhibit such updates. Theorems 1 and 2 show that a one-layer LSA cannot condition its preconditioner on the realised context, while a two-layer LSA can. Section 6 shows that trained models exhibit this transition at depth two.

Mechanistic evaluation and emergence. The empirical methodology uses interventions to examine the computation implemented by trained models. Elhage et al. (2021); Olsson et al. (2022) studied in-context copying and identified an induction-head mechanism requiring two attention layers. Subsequent work proposed associative-memory formation as an explanation for this depth requirement (Bietti et al., 2023). The same requirement has also been observed on recall tasks (Okpekpe & Orvieto, 2025). Section 8 revisits this setting using the oracle-referenced instrumentation developed here.

3 Setting

Tasks. We randomly sample a covariance matrix $\Sigma \succ 0$, weights $w \sim \mathcal{N}(0, I_d)$, context $x_i \sim \mathcal{N}(0, \Sigma)$ and $y_i = w^\top x_i + \tau\varepsilon_i$ for $i = 1, \ldots, n$. The model predicts y_q at a query x_q. We write $\hat{\Sigma} = \frac{1}{n}\sum_i x_i x_i^\top$ and $b = \frac{1}{n}\sum_i x_i y_i$.

Architecture. We use the LSA architecture of Ahn et al. (2023), where attention is linearised by removing the softmax rather than by the kernel approximation of Katharopoulos et al. (2020). The input is the token matrix $Z \in \mathbb{R}^{(d+1)\times(n+1)}$ with columns $(x_i; y_i)$ and query $(x_q; 0)$. Each layer updates Z according to

$$Z \leftarrow Z + \tfrac{1}{n} PZM(Z^\top QZ),$$

where $M = \mathrm{diag}(I_n, 0)$ masks the query token. After L layers, the prediction is the negative of the label coordinate of the query token, following Ahn et al. (2023). The sign is absorbed by P in the unrestricted parametrisation and matters only for the canonical parametrisation of Section 5.

Task families. There are two task families. In the fixed-design control, the context inputs are fixed across all tasks, so $\widehat{\Sigma}$ is identical in every prompt. A fixed update is therefore optimal, and any detected adaptivity is a false positive. In the evidence-sensitive family, each task randomly samples its own covariance (random basis, log-uniform condition number in $[1.5, 40]$) and noise level, so the optimal update depends on the realised context.

The oracle ladder. To measure how many fixed optimisation steps a trained transformer implements, we compare it against a hierarchy of fixed-step predictors. The l-th rung, $\mathcal{O}_l^{\mathrm{fix}}$, is fitted by nested least squares on training draws. At step k, the preconditioner is chosen as the least-squares optimum given the preceding steps. For $l = 1$, this is the best fixed bilinear predictor. For $l \geq 2$, the stagewise construction provides an upper bound on the error of the best jointly optimised chain of l fixed steps. We therefore also report the jointly optimised chain in Appendix C. All predictors are evaluated on the same held-out tasks.

We also include a fully context-adaptive predictor as a reference. A single adaptive step attains the Bayes floor, providing a benchmark for context-dependent updates. A transformer with lower error than $\mathcal{O}_L^{\mathrm{fix}}$ has lower error than the stagewise L-step fixed predictor. Appendix C shows that the depth-two model also has lower error than the jointly optimised two-step chain.

4 Instruments

The following diagnostics are functions of the update kernel, $K(x_q, x_i) = \partial\hat{y}/\partial y_i$, which measures how each context label influences the prediction. This kernel is computed by automatic differentiation and characterises the form of the update applied by the model independently of its predictive accuracy.

Adaptive share. We fit the update kernel with the best bilinear kernel

$$K_{\mathrm{fix}}(x_q, x_i) = x_q^\top \Gamma x_i,$$

using a single matrix Γ shared across all contexts and optimising over its d^2 free parameters. Every fixed preconditioned update $\hat{y} = x_q^\top G b$ has a kernel of exactly this form, so an adaptive share of zero is the signature of a fixed update. Theorem 1 further shows that a one-layer LSA can produce no other form of context dependence. The fit R^2_{fix} measures how much of the evidence response is explained by the best fixed preconditioner, and we define the adaptive share as $1 - R^2_{\mathrm{fix}}$.

Context-conditional gap. A positive adaptive share does not necessarily imply context-adaptive optimisation, since a nonlinear but context-independent response also fails the bilinear fit. The two statistics therefore serve different purposes. The adaptive share measures departure from the fixed bilinear class. The registered statistic, the context-conditional gap $R^2_{\mathrm{ad}} - R^2_{\mathrm{fix}}$, tests whether a specified context-dependent candidate explains that departure. Here, R^2_{ad} fits the two-parameter family $c x_q^\top (\hat{\Sigma} + \lambda I)^{-1} x_i$, with (c, λ) shared between contexts and $\hat{\Sigma}$ recomputed per context. A positive context-conditional gap states that this context-dependent law explains the kernel better than the best d^2-parameter fixed matrix.

Choice of candidate family. The adaptive share is candidate-independent, while the context-conditional gap depends on the proposed update law. We use the family $c x_q^\top (\hat{\Sigma} + \lambda I)^{-1} x_i$ because it is motivated independently of both statistics. It attains the Bayes floor for this task distribution (Section 3). The family is also motivated by related Newton-like interpretations of deeper transformer stacks (Fu et al., 2024). The instrumentation can therefore reject a candidate as well as support one. Sections 7 and 8 provide examples of both outcomes.

Validation. Table 1 evaluates the instrumentation on simulated students with known update mechanisms. The fixed and two-step-fixed students produce shares of at most 0.036 and non-positive

Table 1: Instrument validation on simulated students with known update mechanisms. A positive context-conditional gap identifies context dependence beyond any fixed preconditioner, while R^2_{ad} measures how well the proposed candidate describes the observed update. In these validation experiments, shares below about 0.04 are not evidence of adaptivity.

STUDENT	TRUE MECHANISM	SHARE	GAP	R^2_{ad}
fixed	fixed	0.000	-0.121	0.88
two-step fixed	fixed	0.036	-0.040	0.92
pure adaptive	adaptive, proposed candidate	0.167	$+0.154$	0.99
30% adaptive mixture	adaptive, proposed candidate	0.015	-0.048	0.94
wrong-form adaptive	adaptive, different candidate	0.425	$+0.231$	0.81

context-conditional gaps, so two fixed optimisation steps do not register as adaptivity. The pure adaptive and wrong-form students both produce positive gaps. The 30% adaptive mixture gives a share of 0.015, matching the predicted variance fraction $0.3^2 \times 0.167 = 0.015$. It lies below the two-step-fixed student. Shares below about 0.04 are therefore not evidence of adaptivity.

The wrong-form student is a negative control for law identification. Its gap of $+0.231$ exceeds that of the pure adaptive student ($+0.154$), but the natural-gradient candidate fits its kernel with $R^2_{\mathrm{ad}} = 0.81$, compared with 0.99 for the pure adaptive student. A positive gap thus detects context-dependence without identifying the update law, which is read from R^2_{ad}.

The fitting procedure recovers known preconditioning matrices to relative error below 10^{-4}. Bootstrap confidence intervals achieve empirical coverage of 0.963 at the nominal 0.95 level. Comparisons between independently fitted instances of the same oracle predictor yield a false-positive rate of 2 in 30 trials (0.067), within the 0.10 bound for the two-directional test at $\alpha = 0.05$ per direction.

5 THEORY

All results in this section concern the architecture introduced in Section 3. The architecture is single-head linear self-attention without normalisation. Residual token updates are used, with a zero-initialised query-label channel. Complete proofs appear in Appendix A. Within this architecture we establish a depth-one impossibility result and a depth-two separation. The representation theorem characterises every one-layer LSA and shows that its update cannot depend on the observed covariance. The separation theorem shows that this limitation is absent from the unrestricted depth-two representable class. Corollary 3 extends the separation to arbitrary finite depth, showing that for every finite depth there is an ensemble on which no sequence of that many fixed updates can substitute for a single context-adaptive update.

These results provide the theoretical basis for the measurements introduced in Section 4. In particular, a positive context-conditional gap cannot be produced by any fixed preconditioned update, while the two-step fixed student of Section 4 empirically shows that the measurement does not arise from a composition of two fixed updates. The results also identify a transition from affine to quadratic dependence on the observed covariance, the form that arises in second-order interpretations of prior work (Fu et al., 2024). All four statements follow from the same one-layer identity, $\Delta Z = PSQZ$, where $S = \frac{1}{n}\sum_i z_i z_i^\top$ is the context Gram matrix and the only point at which the observed context enters.

Theorem 1 (Depth-one representation)**.** *For arbitrary parameters, the depth-one update kernel has the form*

$$K(x_q, x_i) = \frac{1}{n} x_q^\top \Gamma_0 x_i + \frac{2\beta}{n} y_i (v^\top x_q),$$

where Γ_0 is a fixed matrix determined by the model parameters. The kernel is independent of the observed covariance $\hat{\Sigma}$, so a one-layer LSA cannot implement a context-adaptive preconditioner. Under the canonical parametrisation, the second term vanishes. The kernel is then exactly the fixed bilinear form, giving $R^2_{\mathrm{fix}} = 1$ and adaptive share 0.

Theorem 2 (Depth-two expressivity and separation)**.** *In the canonical parametrisation, depth two coincides with the two-fixed-step class $\mathcal{O}_2^{\mathrm{fix}}$. With unrestricted parameters, the effective preconditioner acquires a non-zero quadratic dependence on $\widehat{\Sigma}$, and the representable family strictly contains $\mathcal{O}_2^{\mathrm{fix}}$. There exists an explicit three-atom context ensemble, and an open neighbourhood of it, on which the optimal depth-two LSA attains strictly lower risk than every member of $\mathcal{O}_2^{\mathrm{fix}}$. On this ensemble, the Bayes preconditioner is*

$$\gamma^\star(s) = (s+\lambda)^{-1},$$

and every member of $\mathcal{O}_2^{\mathrm{fix}}$ incurs strictly positive excess risk relative to the Bayes predictor. The Bayes preconditioner $\gamma^\star(s)$ is strictly convex as a function of the context scale s. Every member of $\mathcal{O}_2^{\mathrm{fix}}$ is affine in s. Hence an affine rule can coincide with $\gamma^\star(s)$ for at most two distinct values of s.

The canonical parametrisation therefore cannot account for the observed transition, since it is equivalent to two fixed steps. The unrestricted parametrisation introduces cross-coordinate interactions that allow the effective preconditioner to depend quadratically on the observed covariance.

Corollary 3 (Lower bound at every fixed depth)**.** *Theorem 2 extends to every finite depth. An l-step fixed composition represents an effective preconditioner that is a polynomial of degree at most $l-1$ in $\widehat{\Sigma}$. There exists an ensemble of $l+1$ distinct atoms $s_1 < \cdots < s_{l+1}$ with weights w_k (defined in Appendix A) on which every predictor in $\mathcal{O}_l^{\mathrm{fix}}$ incurs excess risk of at least*

$$\Big[\prod_k (s_k+\lambda)\Big]^{-2} \Big/ \sum_k \delta_k^2/w_k, \qquad \delta_k = \prod_{j\neq k}(s_k - s_j)^{-1},$$

whereas a single context-adaptive step attains zero excess risk. The bound increases as λ decreases and is attained by the optimal degree $(l-1)$ polynomial rule.

Proposition 4 (Identification)**.** *In population, every fixed preconditioned update rule $\hat{y} = x_q^\top G b$ satisfies $R^2_{\mathrm{fix}} = 1$ and therefore has a non-positive context-conditional gap. The natural-gradient candidate with non-degenerate $\widehat{\Sigma}$ gives a strictly positive context-conditional gap. In population, a positive context-conditional gap is therefore incompatible with any fixed preconditioned update. An estimated positive gap provides evidence against the fixed class, subject to the uncertainty procedure in Appendix C.*

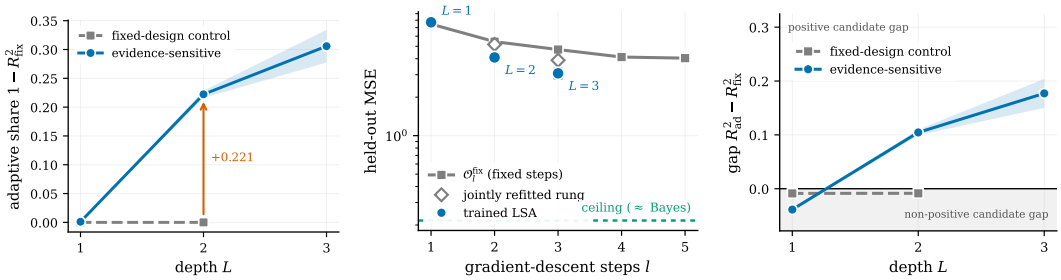
6 Experimental results

Unless otherwise stated, all reported results use 20 independently seeded models per depth with input dimension $d = 16$, trained for 30,000 optimisation steps and evaluated on $4{,}000$ common held-out tasks. Results are reported over converged seeds. The primary analysis concerns depths one and two, as specified before the confirmatory runs. Deeper models are less stable to train because of the higher-order parameter dependence of unnormalised linear self-attention (Appendix B). The convergence criterion in Appendix B uses held-out error only and does not use the adaptive share, context-conditional gap, candidate fit or oracle-rung placement. On the evidence-sensitive family, all 20 seeds converge at depths one to three and none at depth four. On the fixed-design control, 8 of the 20 depth-two runs fail the criterion, leaving $N = 12$ for that condition, and no depth-three or depth-four run converges.

Depth one is a fixed update. The depth-one model has $R^2_{\mathrm{fix}} = 1.000$ and an adaptive share of 0.001. Its held-out error matches that of $\mathcal{O}_1^{\mathrm{fix}}$ (ratio 1.00), and placement at $\mathcal{O}_1^{\mathrm{fix}}$ passes the equivalence and superiority tests in all 20 seeds (Figure 1b).

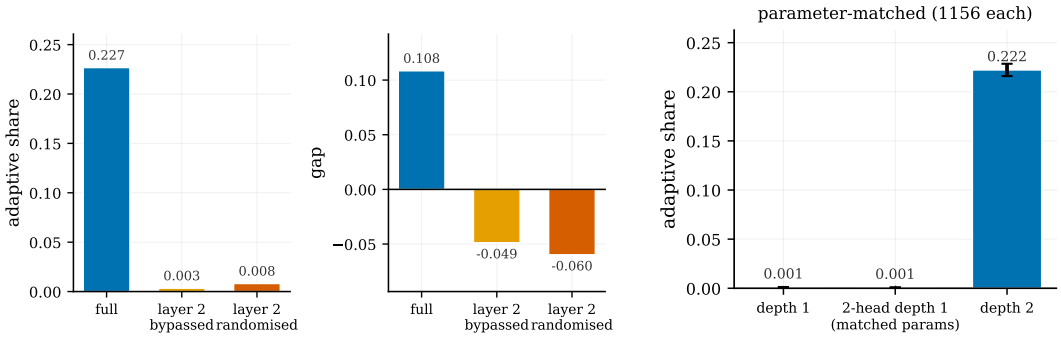
Emergence at depth two. At depth two, the adaptive share increases from 0.001 to 0.222 ± 0.006, six times the 0.036 detection floor established in Section 4 (Figure 1a). The context-conditional gap becomes positive (Figure 1c), increasing from -0.039 at depth one to $+0.105$ at depth two. The seed-level bootstrap interval for the increase is $[0.215,\ 0.227]$, showing that the increase is consistent across training runs.

Depth-two models have lower error than $\mathcal{O}_2^{\mathrm{fix}}$ (5.73) and the jointly optimised two-step chain (5.17). Fitting the affine family $x_q^\top(C + D\hat{\Sigma}E)b$, which contains every two-step fixed chain, gives no



(a) Adaptive share by depth. (b) Depth-oracle ladder. (c) Context-conditional gap.

Figure 1: The depth-two transition. (a) The adaptive share is 0.001 at depth one and 0.222 ± 0.006 at depth two on the evidence-sensitive family, and 0.000 at all depths on the fixed-design control. (b) The depth-one model is equivalent to the first fixed rung. The depth-two and depth-three models lie below their corresponding stagewise fixed rungs. Jointly refitted rungs at $l = 2, 3$ have lower error, but the relative ordering of the trained models is unchanged (Appendix C). (c) The context-conditional gap becomes positive at depth two. Error bars show 95% seed-level bootstrap intervals over converged seeds. The evidence-sensitive family has $N = 20$ at each depth shown. The fixed-design control has $N = 20$ at depth one and $N = 12$ at depth two.



(a) Layer-two intervention. (b) Capacity control.

Figure 2: Interventions. (a) Bypassing layer two of a trained depth-two model reduces the adaptive share from 0.227 to 0.003 and the gap from $+0.108$ to -0.049. The ladder position moves from $\mathcal{O}_4$ to $\mathcal{O}_1$. Randomising layer two gives a further reduction. (b) A two-head depth-one model with the same parameter count as the depth-two model (1156 parameters each) has an adaptive share of 0.001.

further held-out improvement. The trained model has error 4.16 on the same seeds and beats the jointly optimised chain in three of three seeds under the paired test (Appendix C).

On the fixed-design control, the adaptive share is 0.000 and the context-conditional gap is non-positive at depths one and two.

Layer-two ablation. The architecture has no normalisation and residual token updates, so evaluating the first layer in isolation is a true ablation. Bypassing the second layer reduces the adaptive share from 0.227 to 0.003 and the context-conditional gap from $+0.108$ to -0.049, while the ladder position moves from $\mathcal{O}_4$ to $\mathcal{O}_1$ (Figure 2a). These placement labels refer to the stagewise ladder. Against the jointly refitted rungs, the depth-two model lies between the two-step and three-step chains. The measurements are from a representative seed, whose full-model adaptive share is 0.227 rather than the 20-seed mean of 0.222. Randomising the second layer further reduces both adaptivity and predictive accuracy, and a retrained depth-one model matches the bypassed first layer. All five predefined intervention checks satisfy their acceptance criteria.

Capacity controls. A two-head depth-one LSA has the same parameter count as the depth-two model (1156 parameters each) but still performs only a single computational step. Applying The-

orem 1 head by head, the combined kernel remains fixed. The measured adaptive share is 0.001, as predicted (Figure 2b). Together with the layer-two ablation, this localises the measured adaptive computation to the second layer. Thus, the effect is not explained by parameter count.

Layer grafts. The second layer is exchanged between a model trained on the evidence-sensitive family and one trained on the fixed-design control. This intervention is well defined because the model operates directly on the input tokens. The control-trained model gives an adaptive share of 0.30 on the evidence-sensitive family despite having learned no adaptive update. The adaptive share alone is therefore not informative under this distribution shift. The context-conditional gap distinguishes the models. The co-adapted model gives a gap of $+0.108$, whereas the two hybrid models give gaps of -0.65 and -0.08. Combined with the ablation results above, these measurements localise the adaptive computation to layer two. The ablation shows that layer two is necessary, while the graft experiment provides a negative control showing that its weights alone are not sufficient. As a result, the measured correction depends on their co-adaptation with layer one.

7 GENERALITY

Dimension. The adaptive-share transition is present at every input dimension tested, but identification of its update law is not. The share increases by 0.211, 0.208, 0.186 and 0.071 at $d = 8, 16, 32$ and 64 (Figure 3a), with a smaller increase at the highest dimension. The context-conditional gap is positive at $d \leq 16$ ($+0.103$ and $+0.053$) and negative at $d \geq 32$ (-0.030 and -0.032). The data support the natural-gradient candidate at $d \leq 16$ and not at $d \geq 32$. A reduced-training-budget sweep reproduces the same pattern.

Context-length control. At fixed $d = 32$, increasing the context-to-dimension ratio n/d from 2 to 16 changes the context-conditional gap only from -0.032 to -0.016 (Figure 3b), while the adaptive share remains between 0.151 and 0.203. The adaptive share remains above the 0.036 detection floor throughout the sweep, whereas the context-conditional gap stays non-positive. Based on the validation results of Section 4, this combination indicates a context dependence that is not explained by the covariance-inverse candidate.

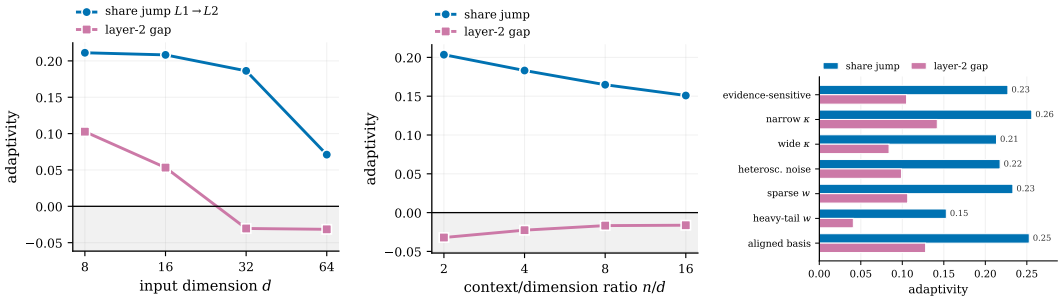
Breadth and shift. The depth-two transition is observed across a range of task families, including narrow ($[1.2, 6]$) and wide ($[4, 120]$) condition-number ranges, a shared eigenvector basis, heteroscedastic observation noise, and sparse and heavy-tailed weight priors. Across these settings, the adaptive share ranges from 0.153 to 0.256 (Figure 3c). The measurements show the same pattern under a train-to-test distribution shift in which both covariance and noise parameters differ between training and evaluation, giving a depth-two adaptive share of 0.304.

A downstream comparison with fixed-gradient descent is reported in Appendix D (Figure 4b).

8 TRANSFER OF THE MEASUREMENT FRAMEWORK

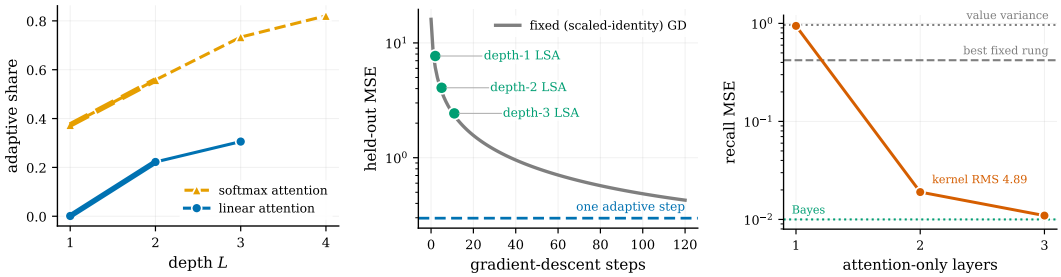
The preceding sections establish the depth-two transition for linear self-attention. We now test the measurement framework outside that setting. These experiments do not extend Theorems 1 and 2. We test the task-independence result of Theorem 1 on binary classification. We then compare the predicted behaviour of linear and softmax attention. Finally, we evaluate the instrumentation on an attention-only model with a known update mechanism.

Softmax attention. The softmax transition differs from that of linear self-attention. Softmax models have adaptive shares of 0.374, 0.558, 0.734 and 0.821 at depths one through four, respectively. The fraction of the total increase occurring at the first depth increment is 0.62 for linear attention and 0.41 for softmax (Figure 4a), consistent with a smoother transition. These experiments lie outside the scope of the theory. Theorem 1 applies only to linear self-attention, and a softmax layer can produce a non-bilinear update kernel without depending on the observed covariance, as in the adaptive nearest-neighbour predictor of Collins et al. (2024).



(a) Dimension. (b) Context-length control. (c) Task-family breadth.

Figure 3: Scope. (a) The adaptive share increases at depth two across the input dimensions tested. The gap for the natural-gradient law decreases and becomes negative at $d \geq 32$ under the full training budget. (b) At fixed $d = 32$, the gap changes from -0.032 to -0.016 as n/d increases from 2 to 16. (c) The adaptive share increases at depth two across all task-family variants.



(a) Transition sharpness. (b) Downstream consequence. (c) Associative recall.

Figure 4: Extensions. (a) Linear attention has 0.62 of the total share increase at the first depth increment, compared with 0.41 for softmax attention. Depth four is excluded for linear attention by the convergence criterion in Appendix B. (b) One natural-gradient step reaches an error not attained by fixed scaled-identity gradient descent within the evaluation horizon. Each trained depth-L model is shown on the same convergence curve. (c) Recall error by number of attention-only layers. The one-layer model has error close to the value-variance baseline. The two-layer model has error twice the Bayes floor and below every rung of the fixed ladder. The three-layer model has error close to the Bayes floor. The two-layer update-kernel root-mean-square is 4.89, compared with $\sqrt{n} = 4.90$ for a one-hot match selector.

Layer-wise analysis. Prefix predictors are constructed from a single trained softmax model using the embedding, the first k encoder layers, and a ridge read-out calibrated at each depth. The adaptive share increases smoothly as additional layers are included. The context-conditional gap first becomes positive at the second layer (-0.059, $+0.166$, $+0.453$, $+0.527$ across the four prefixes). Bypassing the upper layers removes the contribution they make to the measured adaptivity.

Binary classification. The Bayes predictor for in-context binary classification is non-linear in the context moments, so the depth-two theory does not apply. Theorem 1, however, makes no assumptions about the label distribution. Consistent with this prediction, the depth-one model remains a fixed update, with adaptive share 0.000, and is placed on the first fixed rung of the oracle ladder. At depth two, the adaptive share increases to 0.149, and the model is placed on the second rung across all seeds. The natural-gradient candidate is not evaluated because it lies outside the scope of this task.

Associative recall. We evaluate the instrumentation on sequential associative recall using an attention-only transformer, the minimal architecture of the induction-head analysis of Olsson et al. (2022). The Bayes predictor for this task is a one-hot match selector, whereas the best fixed bilinear

update is a kernel smoother. Therefore, the entire fixed-step ladder remains at error 0.421, well above the Bayes floor of 0.010.

The one-layer model fails the task (error 0.943), close to the value-variance baseline of 0.964 and worse than the 0.421 smoother floor. That floor is not reachable by a single layer under this encoding. The two-layer model reduces the error to 0.019, below every fixed rung, and a third layer has error 0.011, close to the Bayes floor of 0.010. The transition therefore occurs at two layers, with the third refining a mechanism that is already in place.

The match-selector kernel has a root-mean-square value $\sqrt{n}$, equal to 4.90 at the length of the evaluation context. The measured two-layer kernel has root-mean-square 4.89, whereas the one-layer response is a diffuse value-averaging kernel with value 0.96. The match-selection mechanism can therefore be identified directly from the update kernel without inspecting attention maps.

As in the layer-grafting experiment, the adaptive share alone is not informative here. It saturates because the one-layer response is non-bilinear despite being context independent. The context-conditional gap remains informative when computed with the match-indicator candidate. The candidate explains none of the one-layer kernel ($R^2_{\mathrm{ad}} = 0.00$) and 99.2% of the two-layer kernel, exceeding the best fixed bilinear fit by $+0.575$.

9 Discussion

The update kernel requires only a differentiable predictive mean, so the same measurements can be applied beyond the linear regression setting of the theory. We apply them to binary classification, softmax attention and associative recall. In the latter setting, they recover the match-selection mechanism directly from derivatives and reproduce a transition previously established through attention and weight analyses (Olsson et al., 2022).

The theoretical results are derived for linear self-attention, where several optimisation-based characterisations of in-context learning have been obtained (Ahn et al., 2023; Mahankali et al., 2024; Zhang et al., 2024; Vladymyrov et al., 2024). The experiments provide evidence that analogous depth-two transitions occur beyond the linear setting. Across the settings studied in this paper, the transition is associated with a second layer acting on context statistics computed by the first. A related mechanism has been proposed for attention followed by a feed-forward layer, in which the context acts as an implicit low-rank update to the feed-forward weights (Dherin et al., 2025). In linear self-attention, the depth-two transition is characterised exactly by the theory. In softmax attention and associative recall, it is supported by empirical measurements but is not covered by the theoretical results.

10 Limitations

The theoretical results apply to the linear self-attention architecture and training setting of Section 3. The experiments reported in Section 8 are not covered by the theorems. The natural-gradient candidate explains the measured update at $d \leq 16$ but not at $d \geq 32$ in the experiments reported here. We do not claim to have identified the update law in the high-dimensional regime. Linear self-attention beyond depth three is difficult to train reliably. Depths beyond two are subject to a predefined convergence criterion, and failed runs are excluded before analysis. At $d = 16$, none of the 20 depth-four evidence-sensitive runs converge, and 8 of the 20 depth-two fixed-design runs fail the criterion. The principal empirical results concern depths one and two. Finally, the context-manipulation experiments beyond context length rely on a scalar effective-ridge read-out that is sensitive to estimation error, and are reported as exploratory in Appendix D.

Reproducibility statement

All decision thresholds, stopping rules, exclusion criteria, and subsequent protocol modifications are documented in the accompanying reproducibility record. The experimental pipeline is organised as a sequence of independent stages, each of which can be run, resumed, and verified separately. All tables and figures are generated from the run logs, and an included script checks every reported statistic against the logs.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with grammar and writing, editing the paper to improve readability, and creation of some artifacts. Additionally, we used generative AI tools for editing software, creating or editing software code, drafting parts of the research paper, and suggesting the structure of the paper. We have reviewed all AI-assisted work. AI-assisted text was reviewed and revised by the authors to ensure consistency with the underlying research. AI-assisted software and artifacts were checked by the authors before use. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2306.00297.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023. arXiv:2211.15661.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2306.04637.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems*, volume 36, 2023. arXiv:2306.00802.

Liam Collins, Advait Parulekar, Aryan Mokhtari, Sujay Sanghavi, and Sanjay Shakkottai. In-context learning with transformers: Softmax attention adapts to function Lipschitzness. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2402.11639.

Benoit Dherin, Michael Munn, Hanna Mazzawi, Michael Wunder, and Javier Gonzalvo. Learning without training: The implicit dynamics of in-context learning. *arXiv preprint arXiv:2507.16003*, 2025.

Sara Dragutinović, Andrew M. Saxe, and Aaditya K. Singh. Softmax $\geq$ linear: Transformers may learn to classify in-context by kernel gradient descent. *arXiv preprint arXiv:2510.10425*, 2025.

Nelson Elhage, Neel Nanda, Catherine Olsson, et al. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. URL `https://transformer-circuits.pub/2021/framework/index.html`.

Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn to achieve second-order convergence rates for in-context linear regression. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2310.17086.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, 2022. arXiv:2208.01066.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, 2020. arXiv:2006.16236.

Yingcong Li, Ankit Singh Rawat, and Samet Oymak. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2407.10005.

Yue M. Lu, Mary I. Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28), 2025. doi: 10.1073/pnas.2502599122.

Arvind V. Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *International Conference on Learning Representations*, 2024. arXiv:2307.03576.

Destiny Okpekpe and Antonio Orvieto. Revisiting associative recall in modern recurrent models. *arXiv preprint arXiv:2508.19029*, 2025.

Catherine Olsson, Nelson Elhage, Neel Nanda, et al. In-context learning and induction heads. Transformer Circuits Thread, 2022. arXiv:2209.11895.

Lingfeng Shen, Aayush Mishra, and Daniel Khashabi. Position: Do pretrained transformers learn in-context by gradient descent? In *International Conference on Machine Learning*, 2024. arXiv:2310.08540.

Max Vladymyrov, Johannes von Oswald, Mark Sandler, and Rong Ge. Linear transformers are versatile in-context learners. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2402.14180.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, 2023. arXiv:2212.07677.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. arXiv:2306.09927.

A PROOFS

Proof of Theorem 1. By the one-layer identity of Section 5,

$$\hat{y}_1 = e_{d+1}^\top PSQz_q,$$

where $z_q = (x_q; 0)$. We first compute the three factors in turn. Since

$$Qz_q = (Ax_q;\, v^\top x_q),$$

it follows that

$$SQz_q = \big(\widehat{\Sigma}Ax_q + b(v^\top x_q);\, b^\top Ax_q + \widehat{c}(v^\top x_q)\big).$$

Applying the final row of P gives

$$\hat{y}_1 = w^\top \widehat{\Sigma}Ax_q + x_q^\top(vw^\top + \beta A^\top)b + \beta\widehat{c}(v^\top x_q).$$

The update kernel is obtained by differentiating with respect to the support labels. Since

$$\frac{\partial b}{\partial y_i} = \frac{x_i}{n}, \qquad \frac{\partial \widehat{c}}{\partial y_i} = \frac{2y_i}{n}, \qquad \frac{\partial \widehat{\Sigma}}{\partial y_i} = 0,$$

we obtain

$$K(x_q, x_i) = \frac{1}{n}x_q^\top \Gamma_0 x_i + \frac{2\beta}{n}y_i(v^\top x_q),$$

where

$$\Gamma_0 = vw^\top + \beta A^\top$$

is independent of the observed context. Thus, the update kernel contains no dependence on $\widehat{\Sigma}$.

Under the canonical parametrisation

$$P = e_{d+1}e_{d+1}^\top, \qquad Q = \mathrm{diag}(A, 0),$$

we have $\beta v = 0$, so the label-dependent term vanishes. The kernel therefore reduces to

$$K(x_q, x_i) = \frac{1}{n}x_q^\top A^\top x_i,$$

which is exactly a fixed bilinear form. Consequently $R^2_{\text{fix}} = 1$, and the adaptive share is zero. □

Proof of Theorem 2. Canonical parametrisation. Under the canonical parametrisation, $P_\ell = e_{d+1}e_{d+1}^\top$ and $Q_\ell = \mathrm{diag}(A_\ell, 0)$, the feature coordinates remain unchanged. The label coordinates satisfy

$$y_i \mapsto y_i + x_i^\top A_1^\top b,$$

so

$$\widehat{\Sigma}^{(1)} = \widehat{\Sigma}, \qquad b^{(1)} = (I + \widehat{\Sigma}A_1^\top)b.$$

Applying the second layer and the negated read-out gives

$$\hat{y}_2 = -x_q^\top(A_1^\top + A_2^\top + A_2^\top\hat{\Sigma}A_1^\top)b.$$

With $M_0 = -A_1^\top$ and $M_1 = -A_2^\top$, this becomes

$$x_q^\top(M_0 + M_1 - M_1\hat{\Sigma}M_0)b,$$

which is the two-fixed-step class.

Unrestricted parametrisation. When $B \neq 0$, the first layer also updates the feature coordinates:

$$x_i^{(1)} = (I + B\widehat{\Sigma}A_1)x_i + \rho_i,$$

where ρ_i collects the terms proportional to b and $\widehat{c}$. Substituting this expression into

$$b^{(1)} = \frac{1}{n}\sum_i x_i^{(1)}y_i^{(1)}$$

introduces, among others, the quadratic term

$$B\widehat{\Sigma}A_1\widehat{\Sigma}A_1^\top b.$$

The second layer therefore contributes

$$x_q^\top A_2^\top B\widehat{\Sigma}A_1\widehat{\Sigma}A_1^\top b$$

to the prediction. Its coefficient is a non-zero polynomial in the model parameters and is therefore non-zero on an open dense subset of parameter space. The effective preconditioner consequently has quadratic dependence on $\widehat{\Sigma}$, whereas every member of $\mathcal{O}_2^{\mathrm{fix}}$ is affine. The representable family therefore strictly contains $\mathcal{O}_2^{\mathrm{fix}}$.

Separation. Take $\Sigma = I_d$, $\lambda = \tau^2/n$, and a three-atom ensemble $\widehat{\Sigma} = sI_d$ with $s \in \{s_1, s_2, s_3\}$ and positive atom probabilities.

Because x_q and b are isotropic, the excess risk of a matrix preconditioner $\Gamma(s)$ is

$$\frac{c(s)}{d}\|\Gamma(s) - \gamma^\star(s)I\|_F^2.$$

Thus it suffices to consider the diagonal entries of $\Gamma(s)$. Every member of $\mathcal{O}_2^{\mathrm{fix}}$ has diagonal entries that are affine in s. The isotropic depth-two subfamily below gives polynomial diagonal entries. The target value for each diagonal entry is

$$\gamma^\star(s) = (s + \lambda)^{-1}.$$

Restricting to isotropic parameters $A_1 = a_1 I$, $A_2 = a_2 I$, and $B = bI$, with the remaining blocks zero, gives the scalar preconditioner

$$\gamma(s) = a_1 + a_2(1 + ba_1 s)^2(1 + a_1 s).$$

At $b = 0$, this reduces to the affine form

$$\gamma(s) = a_1 + a_2 + a_1 a_2 s.$$

On the isotropic ensemble, every two-step fixed predictor is affine in s. Let γ be the risk-minimising affine predictor and define

$$r_k = \gamma^\star(s_k) - \gamma(s_k).$$

By optimality, r is w-orthogonal to $\{1, s\}$. Since $\gamma^\star$ is strictly decreasing, this predictor has negative slope $c_1 = a_1 a_2 < 0$. Any affine rule with $c_0^2 \geq 4c_1$ is realised by some (a_1, a_2), and this condition holds whenever $c_1 < 0$. Thus the optimal affine rule lies in the family with $a_1 a_2 \neq 0$.

The derivative of the depth-two family with respect to b at $b = 0$ is

$$\left.\frac{\partial\gamma}{\partial b}\right|_{b=0} = 2a_1 a_2 s(1 + a_1 s).$$

Since $a_1 a_2 \neq 0$, this has a non-zero s^2 component. The residual r is nonzero and w-orthogonal to $\{1, s\}$. On three distinct atoms, this orthogonal complement is one-dimensional. Since the derivative has a non-zero s^2 component, it has a non-zero projection onto this complement. Hence

$$\sum_k w_k r_k \left.\frac{\partial\gamma(s_k)}{\partial b}\right|_{b=0} \neq 0,$$

so the risk decreases strictly for a suitable perturbation of b. The depth-two family therefore has strictly lower risk than the best affine predictor, and hence than every member of $\mathcal{O}_2^{\mathrm{fix}}$. Continuity of the risks in the ensemble parameters extends the strict inequality to an open neighbourhood. □

Proof of Corollary 3. By induction, every predictor in $\mathcal{O}_l^{\mathrm{fix}}$ has an effective preconditioner that is a polynomial in $\widehat{\Sigma}$ of degree at most $l - 1$. The first step gives $w_1 = M_0 b$, and each subsequent step updates

$$w \mapsto (I - M_k\widehat{\Sigma})w + M_k b,$$

which increases the polynomial degree by at most one.

On the atom ensemble $\widehat{\Sigma} = sI$, the scalar reduction from the proof of Theorem 2 applies. With $w_k = \pi_k\, c(s_k)$, the excess risk of a polynomial rule $\gamma(s)$ is

$$\sum_k w_k \big(\gamma(s_k) - \gamma^\star(s_k)\big)^2.$$

The optimal polynomial is obtained by orthogonal projection in the weighted inner product on $\mathbb{R}^{l+1}$. The orthogonal complement of the degree $(l-1)$ polynomial subspace is spanned by the vector with entries δ_k/w_k, where

$$\sum_k \delta_k s_k^m = 0, \qquad m < l.$$

The squared residual of $\gamma^\star$ is therefore

$$\frac{\left(\sum_k \delta_k \gamma^\star(s_k)\right)^2}{\sum_k \delta_k^2/w_k}.$$

The numerator is the lth divided difference of $s \mapsto (s+\lambda)^{-1}$, which equals

$$(-1)^l \prod_k (s_k + \lambda)^{-1}.$$

Substituting this expression gives the stated lower bound. The context-adaptive rule matches $\gamma^\star$ on every atom and therefore has zero excess risk. □

Proof of Proposition 4. Suppose first that

$$\hat{y} = x_q^\top G b$$

for a fixed matrix G. Its update kernel is

$$K(x_q, x_i) = \frac{1}{n} x_q^\top G x_i,$$

which is a member of the fixed bilinear family with $\Gamma = G$. Hence $R^2_{\mathrm{fix}} = 1$. Since $R^2_{\mathrm{ad}} \le 1$, the context-conditional gap is non-positive.

Conversely, suppose the update kernel is

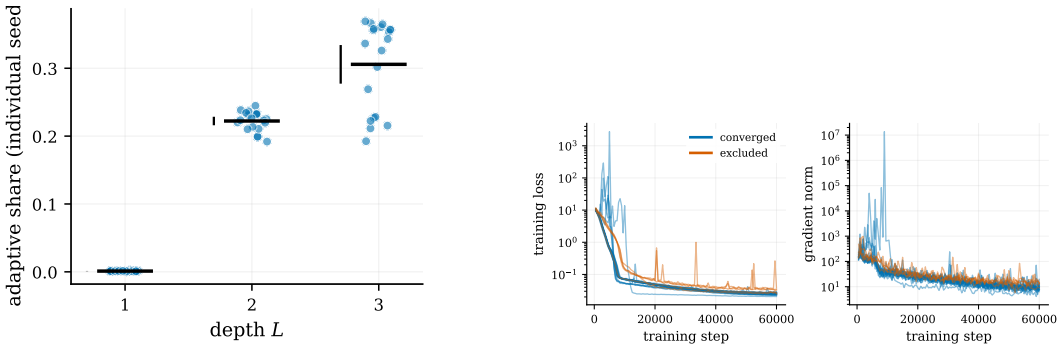
$$K(x_q, x_i) = \frac{c}{n} x_q^\top (\widehat{\Sigma} + \lambda I)^{-1} x_i, \qquad c \neq 0,$$

and that $\widehat{\Sigma}$ is non-degenerate across contexts. The kernel belongs to the candidate family, so $R^2_{\mathrm{ad}} = 1$. If $R^2_{\mathrm{fix}} = 1$, then there exists a fixed matrix Γ such that

$$\Gamma = c(\widehat{\Sigma} + \lambda I)^{-1}$$

for almost every context. This is impossible when $\widehat{\Sigma}$ is non-degenerate across contexts. Therefore $R^2_{\mathrm{fix}} < 1$, and the context-conditional gap is strictly positive. □

B DEEP-MODEL DIAGNOSTICS AND EXCLUSION RULE



(a) Per-seed share distribution.

(b) Deep-model training diagnostics.

Figure 5: (a) Per-seed adaptive share by depth on the evidence-sensitive family. Bars show the mean. (b) Training loss and gradient norm for converged and excluded deep runs. Excluded runs fail the criterion in Appendix B.

Linear self-attention beyond depth two composes a quartic dependence on the model parameters and can diverge from a cold start at standard learning rates. Models with $L \geq 3$ are therefore trained using a damped initialisation, together with a warm-up and cosine learning-rate schedule at a reduced peak learning rate. Learning curves and gradient norms are logged throughout training.

Twenty random seeds are trained for each $(\text{family}, \text{depth})$ configuration. The classification and associative-recall experiments use five seeds per depth. A run is considered not to have converged if its held-out error is not lower than that of the zero predictor, if its error exceeds four times that of the one-step oracle, or if its error is significantly larger than that of the corresponding shallower model trained from the same seed (one-sided bootstrap test, $\alpha = 0.05$). These criteria depend only on held-out prediction error. They do not use the adaptive share, context-conditional gap, candidate fit or oracle-rung placement. Runs that do not converge are excluded from all figures and analyses. Reported seed-level estimates therefore use only converged runs, and the number of retained seeds is stated with each result.

At $d = 16$, every depth-four run on the evidence-sensitive family satisfies at least one exclusion criterion. Depth three is reported only where the training criterion is met. Figure 5 reports the per-seed adaptive-share distributions together with the corresponding learning-curve and gradient-norm diagnostics.

C EXPERIMENTAL DETAILS

Training. Models are trained with Adam at learning rate 10^{-3}. For depths greater than two, training uses a reduced peak learning rate with linear warm-up. The standard training budget is $30{,}000$ optimisation steps. Context lengths are sampled uniformly from 8 to 32 during training and fixed at $n = 24$ for evaluation. Each $(\text{family}, \text{depth})$ configuration uses 20 random seeds. The classification and associative-recall experiments use five seeds per depth.

Reference predictors. Reference predictors are fitted through rung five of the oracle ladder using $60{,}000$ training tasks. Fixed predictors are obtained by nested least squares, which is exact for one step and stagewise for $l \geq 2$. As a check, all step matrices of the two- and three-step rungs were re-optimised jointly by L-BFGS from the stagewise solution on the same training draws. For $l = 2$, we also fitted the affine family $x_q^\top (C + D\hat{\Sigma}E)b$, which contains every two-step chain, from the same initialisation. On seeds 0 to 2, the joint refit lowers the two-step rung from 5.73 to 5.17 and the three-step rung from 4.85 to 3.89. The affine family reaches errors between 5.03 and 5.38. The trained depth-two model has error 4.16 and the depth-three model has error 2.11, both lower than the corresponding jointly refitted chains on all three seeds under the paired superiority test. Placement labels in the main text refer to the stagewise ladder. Against the jointly refitted rungs, the depth-two model has error between the two-step and three-step chains.

The adaptive predictor uses a geometric grid of 40 ridge parameters over $[10^{-3}, 30]$. All predictors are evaluated on the same $4{,}000$ held-out tasks. Variation due to oracle fitting is reported separately and is an order of magnitude smaller than the reported margins.

Uncertainty. The analysis distinguishes three sources of variability: evaluation tasks, training runs, and oracle estimation. Task variability is quantified by paired bootstrap resampling on matched evaluation sets. Training variability is estimated by bootstrap resampling over random seeds and is reported for each headline result. Oracle variability is estimated by refitting the reference predictors on independent training sets. Figures 1 and 3 report seed-level intervals. Table 1 and all oracle margins use the task-level paired bootstrap. Oracle-fitting variability is reported in the text and is an order of magnitude smaller than the reported margins. A hierarchical seed-and-task bootstrap separates within-seed and between-seed contributions.

Statistics. Placement on an oracle rung requires equivalence within a relative margin of 10% by two one-sided tests together with superiority over the preceding rung. A non-significant difference is never interpreted as equivalence. All decision thresholds, stopping rules, exclusion criteria, and subsequent protocol modifications are documented alongside the released code.

D ADDITIONAL TRANSFER EXPERIMENTS

Pretrained transformer. The update-kernel instrumentation requires only predictive means and, as such, can also be used to analyse a pretrained eight-layer transformer for tabular prediction. The measured adaptive share is 0.82. However, a layer-wise interpretation is not appropriate because the linear read-out used to probe intermediate softmax layers introduces nonlinearity unrelated to context-dependent computation. The present results concern models trained directly on in-context objectives. Whether pretrained language models learn in context by anything resembling gradient descent remains an open question (Shen et al., 2024). We therefore report this measurement as an application of the instrumentation rather than as evidence about the underlying training mechanism.

Exploratory context manipulations. We also manipulated the context directly by varying its length, noise level and task diversity. We compared the effective ridge parameter fitted to the full model with that of the layer-one map. Changing the context length shifted the full model's effective ridge from 0.83 to 1.30 while leaving the layer-one map unchanged. Changing the noise level or task diversity did not produce a measurable response. The scalar ridge read-out is sensitive to estimation error, so these results are reported as exploratory and are not used in the main argument.

Downstream comparison. On the evidence-sensitive task family, a single natural-gradient step reaches an error that fixed scaled-identity gradient descent does not attain within 120 steps. The separation increases with the condition number of the task distribution. Relative to the fixed-gradient-descent convergence curve, depth-one, depth-two and depth-three models correspond to approximately 2, 5 and 11 fixed optimisation steps, respectively (Figure 4b).