A DEPTH-INDEXED SAMPLE-COMPLEXITY EXCHANGE RATE BETWEEN IN-CONTEXT LEARNING AND LOW-RANK FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learning (ICL) and low-rank (LoRA) fine-tuning are rarely compared on equal footing. We place both on one statistical model, a rank-r linear map estimated from K demonstrations or m fine-tuning examples, and ask how many demonstrations $K^*(m, L)$ match an m-example fine-tune at depth L. The answer depends on what attention can store. Any in-context estimator that is linear in the labels, including gradient-descent linear attention and any head bottleneck, needs asymptotically at least $d^2/[r(2d - r)]$ demonstrations per fine-tuning example, at every depth. With r register tokens, the role played by system prompts, soft prompts and memory tokens, a fixed weight-tied block implements factored gradient descent exactly, so ICL and LoRA become the same algorithm. The exchange rate then becomes depth-indexed: it diverges below a critical depth $L^* = \log(\sigma_r/\lambda_0)/[2\log(1 + 2\eta\sigma_r)]$ and reaches 1:1 above it, and we prove a two-sided escape-time bound whose constants are independent of the initialization scale λ_0, which the registers set. Task families with $k < rd$ effective parameters lower the achievable rate to $k/[r(2d - r)]$, and pretrained registers shift L^* by roughly $4\times$ without removing the transition. Trained linear transformers confirm the separation.

1 INTRODUCTION

Two questions dominate the practice of adapting large pretrained models: how many in-context demonstrations does a prompt need, and how many labeled examples does a fine-tune need. The first has received a sharp mechanistic answer. Linear self-attention (LSA) implements gradient descent on the in-context regression loss (Von Oswald et al., 2023; Ahn et al., 2023), and architecture-level sample-complexity bounds follow for the resulting predictor (Yang et al., 2026). The second question is, in its statistical essence, much older. Restricting a fine-tuning update to a rank-r factorization, as LoRA does (Hu et al., 2022), is an instance of reduced-rank regression, for which sharp minimax rates were established long before either literature existed: the excess risk of a rank-r-constrained least-squares fit on m examples is $\Theta(rd/m)$ (Negahban & Wainwright, 2011; Bunea et al., 2011).

These answers describe the same quantity, how many labeled examples an adaptation mechanism needs, but were derived for structurally different mechanisms using different tools. Nothing in the literature places them side by side. This is a curious gap, because a separate result already tells us the mechanisms are doing the same thing: Dherin et al. (2026) show that a context prompt induces an implicit low-rank update to a downstream layer, so conditioning *is* an implicit gradient step. If ICL and fine-tuning are two views of one computation, the exchange rate between them, how many demonstrations one fine-tuning example is worth, should be derivable rather than folkloric.

The thesis, practical meaning, and scope. Our answer is that the exchange rate is 1:1 only as an asymptotic endpoint: when both mechanisms solve the same rank-r estimation problem to convergence, their leading sample-complexity constants agree, but at finite depth the exchange rate is controlled by optimization dynamics and changes sharply near a critical depth. A model can have enough demonstrations in context yet fail to use them if its effective depth is below the escape threshold, so the question is not only "how many examples fit in context?" but "how much task-specific

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

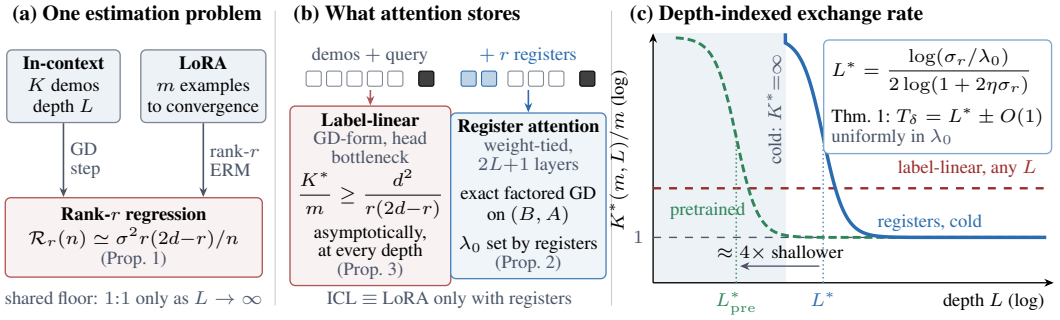


Figure 1: **Overview.** (a) ICL and LoRA estimate one rank-r regression problem. (b) What attention stores decides whether its rate is reachable. (c) Schematic of the resulting depth-indexed exchange rate.

computation can the model perform on them?" Every formal result below is derived and tested inside one idealization: linear self-attention, linear adaptation, and linear ground truth. We adopt it because this is the setting in which conditioning and gradient descent are related by an established identity rather than an analogy (Von Oswald et al., 2023), while retaining an exactly analyzable low-rank optimization problem; we claim nothing directly about softmax attention. Sections 6 and 7 quantify these practical implications and delimit where they do and do not follow.

What attention must store. The premise that ICL and LoRA solve the same rank-r problem is usually stated rather than derived. Deriving it exposes a requirement. The gradient-descent construction of linear attention is linear in the labels, and a head-dimension bottleneck confines the update to a subspace fixed by the weights; neither can select a task-specific rank-r subspace, and we prove that no label-linear in-context estimator reaches the rank-r rate at any depth. What suffices is r *register tokens*, key tokens not derived from the demonstrations, which deployed models already have as system prompts, soft prompts and memory tokens. With them a fixed attention block runs factored gradient descent exactly, the initialization scale in our depth law becomes a property of the registers, and the comparison with LoRA stops being circular: the matched-budget agreement and 1:1 endpoint follow from a construction, while the separation from label-linear attention holds at every depth. The finite-depth law is accordingly a property of factored descent that register ICL inherits exactly: depth is the training budget and the registers are the initialization, so the same transition governs LoRA stopped after L steps.

Contributions. Figure 1 summarizes the three results and how they connect.

- **Registers, not head bottlenecks, make the surrogate exact.** A label-linear lower bound gives $K^*/m \geq d^2/[r(2d-r)]$ asymptotically, at every depth; an explicit register construction realizes factored descent exactly, with λ_0 set by register contents (Propositions 2 and 3).
- **A two-sided discrete escape-time theorem.** Explicit bounds for factored gradient descent, with constants independent of the initialization scale and without small-λ linearization; the residual offset is additive and $2\eta\sigma_1 < 1$ is the local-stability threshold.
- **A depth-indexed exchange rate.** The quantity $K^*(m, L)/m$, a function of $(L, \lambda_0, \eta, \sigma_r, k)$, diverges below the critical depth and approaches 1 above it. The 1:1 statistical rate is the $L \to \infty$ endpoint, not a universal finite-depth claim.
- **Two delimitations.** Task families with $k < rd$ effective parameters move the achievable rate to $k/[r(2d-r)]$; family-level initialization, i.e. pretrained registers, moves L^* by roughly $4\times$ without softening the transition.

2 Related Work

ICL as implicit optimization. Von Oswald et al. (2023) and Akyürek et al. (2023) show LSA can implement gradient-descent steps on in-context linear regression; Ahn et al. (2023) extend this to preconditioned and multi-step GD, and Fu et al. (2024) find second-order updates. Dherin et al.

(2026) generalize the context-as-weight-update picture beyond linear layers. Xie et al. (2025) show the GD/ICL correspondence depends on initialization, which is central for us: our critical depth is a function of the initialization the implicit optimizer starts from, and Section 7 varies exactly that. Yang et al. (2026) give VC-based architecture-level bounds; ours is a distribution-specific statistical bound in the same spirit.

Registers and prompts. Key tokens that do not come from the input are standard: memory tokens (Burtsev et al., 2020), vision-transformer registers (Darcet et al., 2024), and learned soft or prefix prompts (Lester et al., 2021; Li & Liang, 2021); trained language models also route attention to fixed initial tokens (Xiao et al., 2024). We give such tokens a precise role: they hold the factors of the implicit optimizer.

Low-rank adaptation theory. The fine-tuning side of our comparison is what LoRA (Hu et al., 2022) approximates at convergence. Negahban & Wainwright (2011) and Bunea et al. (2011) independently establish matching-order rates $\sigma^2 r(d_1 + d_2)/n$ with matching lower bounds. We build on this and additionally give a from-scratch derivation of the same rate (Appendix A.3) so the central claim does not rest on one source.

Pretraining task distribution. Raventós et al. (2023) show in-context learning of linear-regression tasks emerges only once the pretraining task distribution is sufficiently diverse, with a threshold in the number of pretraining tasks. That diversity axis is the empirical counterpart of the task-family dimension k we sweep in Section 7.

Dynamics of matrix factorization. Saxe et al. (2013) solve the gradient-flow dynamics of two-layer linear networks under small aligned initialization, giving each mode an independent logistic trajectory. Gidel et al. (2019) analyze discrete gradient dynamics for two-layer linear networks and, under vanishing initialization and sufficiently small step size, connect the dynamics to sequential reduced-rank learning; Arora et al. (2019) study implicit regularization in deeper matrix factorizations, and Stöger & Soltanolkotabi (2021); Jacot et al. (2021) small-initialization saddle escape. Our result is complementary: we isolate an exact scalar recursion in an aligned population setting and obtain a two-sided finite-depth escape-time bound with explicit constants in the non-infinitesimal regime $\eta\sigma = O(1)$. Thus the contribution is not a new implicit-regularization principle, but a quantitative depth law for the transient before the reduced-rank solution is reached.

3 Problem Setup

Why reduced-rank regression is the shared object. That ICL and LoRA solve one estimation problem is the premise everything else rests on, so we set it out before using it. On the fine-tuning side it is immediate: LoRA replaces ΔW with a factored BA, $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times d}$, and training that to convergence under squared loss is, by definition rather than by analogy, rank-r-constrained least squares, the classical reduced-rank regression problem (Reinsel & Velu, 1998).

The ICL side turns on an identity of Von Oswald et al. (2023). With each demonstration a token $e_i = (x_i, y_i)$, key and query weights that read only the x-part, and value weights that emit the residual, a linear self-attention layer's output on the query equals one gradient step on $\frac{1}{K}\sum_i \|y_i - Wx_i\|^2$, term for term; stacking L layers reproduces L steps (Ahn et al., 2023). Two consequences make the mechanisms commensurable. First, a forward pass over K demonstrations *is* a gradient step on them, taken without any weight update. The context window thus plays the role of a training set and depth the role of an optimization budget, exactly parallel to how many steps a fine-tune is run for. Second, LoRA's estimator is rank-constrained with a subspace chosen by the data, and the implicit optimizer must be too. The obvious candidate, a head whose value matrix has rank r, does not achieve this: every linear-attention update lies in the span of the value matrices, a subspace fixed by the weights rather than by the task (Lemma 8), and the gradient-descent construction is linear in the labels, so its risk is at least the ridge risk at every depth (Proposition 3). Selecting the subspace requires label-nonlinear state held somewhere other than the weights. We supply it with register tokens.

Assumption 1 (Register-augmented attention)**.** *Besides the K demonstrations and the query, attention reads r key tokens whose content is not derived from the demonstrations (registers). Their contents (a_k, b_k) are the rows of A_0 and columns of B_0.*

The assumption holds in practice. Deployed models prepend demonstration-independent tokens to every prompt as a matter of course: system prompts, learned soft or prefix prompts (Lester et al., 2021; Li & Liang, 2021), and memory or register tokens (Burtsev et al., 2020; Darcet et al., 2024). Under Assumption 1, Proposition 2 constructs fixed attention weights, not assumed dynamics, that run joint gradient descent on (B, A) exactly. Unwound to their statistical content, register ICL and LoRA are then the same procedure, differing only in which pool of examples supplies the gradient (K context tokens against m fine-tuning examples) and how large a budget is spent (L forward-pass steps against training to convergence). Registers are sufficient rather than the only route to label nonlinearity (Appendix J); we use them because they make the surrogate an identity.

Data-generating process. Let $x \sim \mathcal{N}(0, I_d)$. The adaptation is $\Delta W^* = U^* A^*$ with $U^* \in \mathbb{R}^{d\times r}$ Haar-distributed orthonormal and A^* having i.i.d. $\mathcal{N}(0, \tau^2)$ entries, so ΔW^* has rank exactly r. Observations are $y = \Delta W^* x + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$. Given n pairs stacked as $X, Y \in \mathbb{R}^{n\times d}$ we compare four estimators.

(1) Full OLS (unconstrained fine-tuning or infinite-depth unconstrained attention): $\widehat{W}_{\text{ols}} = Y^\top X (X^\top X)^{-1}$.

(2) RRR, the exact rank-r-constrained ERM, standing for LoRA fine-tuned to convergence. With $\hat{C}_{ols} = \widehat{W}_{\text{ols}}^\top$ and $\hat{Y} = X\hat{C}_{ols}$, the minimizer of $\|Y - XC\|_F^2$ subject to $\text{rank}(C) \le r$ is $\hat{C}_r = \hat{C}_{ols} V_r V_r^\top$ for V_r the top-r right singular vectors of $\hat{Y}$; we set $\widehat{W}_r = \hat{C}_r^\top$ (Appendix A).

(3) Unconstrained L-step GD-ICL. A linear-attention layer consuming K context pairs implements L gradient steps on $\frac{1}{K}\left\|Y_{ctx} - X_{ctx} W^\top\right\|_F^2$ from $W_0 = 0$. The loss is quadratic in the unconstrained W, so

$$W_L = \widehat{W}_{ols}^{(K)} \left[I - (I - 2\eta \hat{\Sigma}_K)^L\right], \qquad \hat{\Sigma}_K = \tfrac{1}{K} X_{ctx}^\top X_{ctx}, \tag{1}$$

the classical gradient-descent-as-spectral-shrinkage picture.

(4) Rank-r factored GD-ICL. Register attention (Proposition 2), or equivalently LoRA during training, parameterizes $W = BA$ at initialization scale ε_0, the register norm, and takes L joint steps on both factors. This objective is bilinear and nonconvex; its dynamics are the subject of Section 5. At matched budget (4) and LoRA are the same map (Lemma 1), so their agreement is an identity rather than a finding; the substantive comparisons are (4) against its converged value (Section 5) and against label-linear attention (Proposition 3).

Risk throughout is $\left\|\hat{W} - \Delta W^*\right\|_F^2$, which equals expected squared test error under $x \sim \mathcal{N}(0, I_d)$. Rows (4) and (2) differ only in budget; rows (4) and (3) differ in label-linearity (Figure 1b).

4 The infinite-depth endpoint and what attention must store

The finite-depth theory below is anchored by the statistical limit that both mechanisms approach when the rank constraint is fully optimized. The leading constant is derived asymptotically in Appendix A.4.

Proposition 1 (Shared reduced-rank statistical floor)**.** *For $n > d + 1$, $\mathbb{E}\left\|\widehat{W}_{\text{ols}} - \Delta W^*\right\|_F^2 = \sigma^2 d^2/(n - d - 1)$ exactly. For the rank-r estimator, $\mathbb{E}\left\|\widehat{W}_r - \Delta W^*\right\|_F^2 = \Theta(\sigma^2 r d/n)$. If register ICL (Proposition 2) and LoRA fine-tuning are evaluated under the same rank-r task prior, both are run to their rank-r optima, and the leading risk term has the same constant C_r as a function of sample size, i.e. $R_r(n) = C_r/n + o(1/n)$, then risk matching defines $K^*(m)$ through $R_r(K^*) = R_r(m)$ and yields $K^*/m \to 1$. The shared constant is not an extra hypothesis: it is $C_r = \sigma^2 r(2d - r)$ (Proposition 5).*

The OLS formula follows from $\mathbb{E}[(X^\top X)^{-1}] = I_d/(n-d-1)$ applied to each row. The rank-r order is the classical reduced-rank regression rate (Negahban & Wainwright, 2011; Bunea et al., 2011), and we re-derive the upper bound from scratch in Appendix A.3. The sharper constant $r(2d - r)$, the dimension of the tangent space of the rank-r manifold, is derived for both RRR and truncated OLS in Appendix A.4 and verified in Table 1.

The 1:1 rate is an infinite-depth endpoint, and it holds only for a mechanism that can reach $\widehat{W}_r$. The next two results say which mechanisms can (proofs in Appendix J).

Proposition 2 (Registers realize factored descent exactly)**.** *Under Assumption 1, with d further fixed anchor tokens, there are fixed weights for a two-head linear-attention* refresh *layer and a two-head* update *layer such that refresh followed by L repetitions of (update, refresh), i.e. $2L+1$ weight-tied layers, leaves the registers holding exactly the iterates $B_{l+1} = B_l - \eta G_l A_l^\top$, $A_{l+1} = A_l - \eta B_l^\top G_l$, $G_l = 2(B_l A_l \hat{\Sigma}_K - \hat{S}_K)$, and every query reading out $B_L A_L x_q$. The rank equals the number of registers, and mode strengths $\lambda_i(0) = \sum_k \langle b_k, u_i\rangle\langle a_k, v_i\rangle$ are set by register contents.*

Proposition 3 (Label-linear in-context estimators)**.** *Let U^* be Haar-distributed. Every in-context estimator that is linear in Y given X, in particular estimator (3) at every depth and any linear-attention stack of the gradient-descent form, satisfies $\mathbb{E}\left\|\hat{W} - \Delta W^*\right\|_F^2 \geq R_{\mathrm{lin}}(K) := d\sigma^2\,\mathbb{E}\,\mathrm{tr}[(X^\top X + \lambda I)^{-1}]$ with $\lambda = \sigma^2 d/(\tau^2 r)$, attained by ridge. Hence $K^*(m, L) \geq K_{\mathrm{lin}}(m)$ for every L, where $R_{\mathrm{lin}}(K_{\mathrm{lin}}(m)) = R_r(m)$, and $K_{\mathrm{lin}}(m)/m \to d^2/[r(2d-r)]$.*

At $(d, r, m) = (20, 3, 140)$, $R_{\mathrm{lin}} = 3.16$ against the RRR floor 0.89, and $K_{\mathrm{lin}}/m = 3.34$. Unlike the 1:1 endpoint, which follows once Proposition 2 makes the mechanisms coincide, this ratio compares different mechanisms, and no depth closes it. Head bottlenecks do not help: a weight-fixed subspace of dimension s incurs bias $\tau^2 r(d-s)$ (Proposition 9).

5 Critical depth in discrete time

We analyze estimator (4), realized by register attention (Proposition 2); one step is two layers, so a depth of L steps costs $2L+1$ layers. Let $\Delta W^* = \sum_i \sigma_i u_i v_i^\top$ and initialize the factors in the corresponding singular subspaces with balanced positive strengths. The following assumptions are deliberately stronger than those needed for the empirical sweeps; they isolate the regime in which the discrete recursion is monotone and permit a two-sided theorem.

Assumption 2 (Aligned balanced initialization)**.** *Let $B_0 = \sum_i b_i^{(0)} u_i e_i^\top$ and $A_0 = \sum_i a_i^{(0)} e_i v_i^\top$ with $b_i^{(0)} = a_i^{(0)} = \sqrt{\lambda_i(0)} > 0$.*

Assumption 3 (Monotone stability regime)**.** *For the mode under consideration, $0 < 4\eta\sigma \leq 1$.*

Proposition 4 (Exact scalar recursion)**.** *Under Assumption 2, each singular mode evolves independently as*

$$\lambda_{l+1} = \lambda_l\big(1 + 2\eta(\sigma - \lambda_l)\big)^2. \tag{2}$$

In particular, the aligned balanced subspace is invariant and no approximation is involved in the mode recursion.

Theorem 1 (Two-sided discrete escape time)**.** *Let Assumptions 2–3 hold, let $\lambda_0 \in (0, \sigma/2)$ and $\delta \in (0, 1/2]$, and define*

$$L^* = \frac{\log(\sigma/\lambda_0)}{2\log(1+2\eta\sigma)}, \qquad T_\delta = \min\{l : \lambda_l \geq (1-\delta)\sigma\}. \tag{3}$$

Then

$$\boxed{L^* - C_1 \leq T_\delta \leq L^* + C_2} \tag{4}$$

with

$$C_1 = \frac{\log(1/(1-\delta))}{2\log(1+2\eta\sigma)}, \qquad C_2 = \frac{(1+\eta\sigma)^2}{2\log(1+2\eta\sigma)} + \frac{\log(1/(2\delta))}{-\log(1-2\eta\sigma)} + 1. \tag{5}$$

Both constants depend only on $(\eta\sigma, \delta)$ and are independent of λ_0.

Proof idea. The argument needs no small-λ linearization. Taking logarithms of the exact recursion gives

$$T = \frac{\log(\lambda_T/\lambda_0) + 2D_T}{2\log(1+2\eta\sigma)}, \qquad D_T = \sum_{l<T} \log\frac{1+2\eta\sigma}{1+2\eta u_l}, \quad u_l = \sigma - \lambda_l. \tag{6}$$

The entire discrepancy from L^* is the nonnegative drag term. During the growth phase, geometric amplification implies $\sum_l \lambda_l$ is dominated by its final terms, giving $D_T \leq (1+\eta\sigma)^2/2 \leq 25/32$.

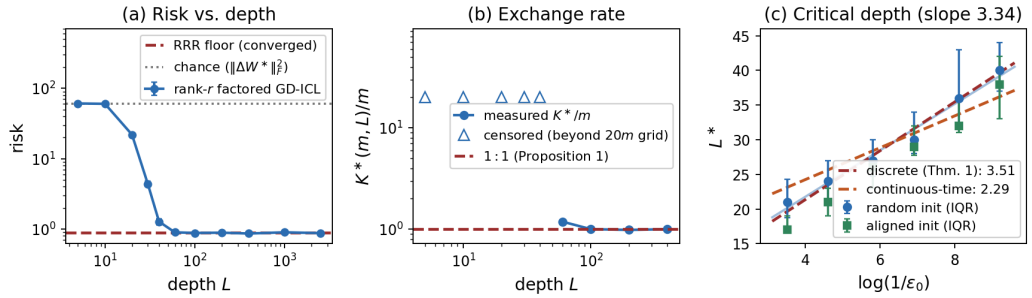


Figure 2: **(a)** Risk of rank-r factored GD-ICL against depth ($K = 140$, 120 trials, mean $\pm$ SEM). **(b)** Exchange rate $K^*(m, L)/m$ at $m = 140$; hollow markers are censored (crossing beyond $20m$). **(c)** Critical depth against initialization scale (60 trials/point, IQR bars) with the discrete (Theorem 1) and continuous-time predictions.

The remaining saturation phase contracts the deficit geometrically. These two bounds produce the stated additive constants uniformly in the initialization scale; the full proof is in Appendix H.

Consequences. The exact map has $g'(\sigma) = 1 - 4\eta\sigma$, so the target is locally attracting if and only if $2\eta\sigma < 1$; Assumption 3 is the stricter no-overshoot regime. Because the bound is additive and uniform in λ_0, T_δ grows in $\log(1/\lambda_0)$ with asymptotic slope exactly $1/[2\log(1+2\eta\sigma)]$ (Corollary 2). Relative to the continuous-time value $\log(\sigma/\lambda_0)/(4\eta\sigma)$, the discrete law is longer by the factor $2\eta\sigma/\log(1+2\eta\sigma) = 1 + \eta\sigma + O((\eta\sigma)^2)$, at most 1.233 in this regime.

Theorem 1 is a population, aligned-initialization result. The finite-K, randomly initialized case remains open because alignment introduces an additional rotation cost; the binding mode satisfies Assumption 3 in every configuration ($4\eta\sigma_r \approx 0.68$ at the base setting), and Appendix H.4 extends the theorem to the full stable range $2\eta\sigma < 1$. For multiple modes, L^* is decreasing in σ, so the smallest nonzero singular value σ_r is the binding mode. Through Proposition 2, λ_0 is the overlap of the registers with the task, so the depth a prompt needs grows logarithmically in how poorly its registers are aligned with the task.

6 Experiments

All experiments use $d = 20$, $r = 3$, $\sigma = \tau = 1$, closed-form OLS and RRR, and explicit gradient recursions, in numpy with fixed integer seeds. Every number is a mean or median over at least 40 independent trials with standard errors, IQRs or bootstrap intervals reported. The suite was implemented with generative-AI assistance and cross-checked against the closed forms of Appendix A; Appendix G discloses what was AI-assisted and how it was verified. The attention constructions of Section 4 are additionally run as actual linear-attention forward passes on GPU (Appendix J).

The transition (Figure 2a,b). At $K = n = 140$, factored GD-ICL risk is at chance (≈ 60) for $L \leq 10$, falls through 21.9, 4.4 and 1.3 at $L = 20, 30, 40$, and matches the RRR floor (0.883 ± 0.008) for every $L \geq 60$ tested up to $L = 2500$. The exchange rate mirrors this: for $L \leq 40$ the crossing is censored beyond $20 \times m$, meaning context alone cannot compensate even at twenty times the fine-tuning sample size, and for $L \geq 100$ it is flat within Monte Carlo noise at 0.99–1.00. Running unconstrained GD-ICL to convergence instead gives risk 3.34 ± 0.02 against RRR's 0.89, a $3.8\times$ penalty that, by Proposition 3, no depth removes: its minimum over depth, 3.23, sits just above $R_{\text{lin}} = 3.16$ (Table 3).

Which time domain (Figure 2c). Sweeping ε_0 over $[10^{-4}, 3 \times 10^{-2}]$ and measuring the median depth at which risk first drops below twice the RRR floor, the measured slope is 3.35 with bootstrap 95% interval $[3.04, 3.77]$. Eq. 3 predicts 3.51, inside that interval. The continuous-time expression evaluated as this literature conventionally would, at the mean of the top-r singular values, predicts 2.29, which lies outside it. Most of the gap is the singular-value plug-in: using σ_r, the binding mode, gives 3.10, and discretization gives 3.51.

Table 1: Proposition 1 verification. **Left:** full-OLS risk against $\sigma^2 d^2/(n-d-1)$ (300 trials/row). **Right:** the RRR constant $r(2d-r)$, read off the largest n and extrapolated by fitting $n\cdot\text{risk} = c+b/n$ over $n \geq 220$.

Full OLS ($d=20$)				RRR constant $r(2d-r)$				
n	empirical	theory	ratio	(d,r)	pred.	largest n	extrapolated	ratio
25	96.79 ± 2.47	100.00	0.968	$(10,2)$	36	36.6	36.2 ± 0.4	1.006
30	43.67 ± 0.56	44.44	0.983	$(20,3)$	111	112.5	112.0 ± 0.6	1.009
40	21.27 ± 0.19	21.05	1.010	$(20,6)$	204	207.1	203.5 ± 0.9	0.997
60	10.33 ± 0.07	10.26	1.007	$(40,4)$	304	306.9	304.3 ± 1.1	1.001
100	5.07 ± 0.03	5.06	1.002					

Table 2: Theorem 1 per mode: half-rise depth against continuous-time and discrete predictions, binned by singular value ($K = 4000$, near-noiseless, 40 trials, 120 crossings).

σ range	$2\eta\sigma$	measured	continuous, rate 2	continuous, rate 4	discrete (Eq. 3)
2.19–3.54	0.306	28.4	49.4	24.7	**28.3**
3.54–4.22	0.388	23.2	39.2	19.6	**23.2**
4.22–4.91	0.457	20.5	33.6	16.8	**20.4**
4.91–6.81	0.549	17.8	28.4	14.2	**17.8**
mean measured/predicted			0.600	1.201	$\mathbf{1.001 \pm 0.002}$
mean absolute error			66.9%	16.6%	**1.9%**

Theorem 1, mode by mode (Table 2). Tracking each singular mode from a balanced aligned initialization, the discrete prediction is accurate to 1.9% over 120 mode-crossings (ratio 1.001 ± 0.002), while the continuous limit is low by 17% and worsens as $\eta\sigma$ grows.

Discretization and threshold checks. Across an order of magnitude in η, the ratio of continuous to discrete error (1.25) matches the predicted correction (1.24; Appendix B). Changing the convergence threshold $\kappa \in [1.1, 5]$ moves the intercept by more than half but the slope by only 4.7%, as a logarithm over a rate requires (Appendix C).

The residual offset is additive, not multiplicative. Random initialization lies outside Theorem 1, so this is an empirical finding. Trial by trial, the ratio of measurement to Eq. 3 drifts with $\log(1/\varepsilon_0)$ while the difference does not; the additive model's RMS residual is 2.2× smaller. The fitted offset is $c_0 = 5.55$ $[4.83, 6.41]$ steps for random and 2.89 $[2.30, 3.48]$ for aligned initialization; the 2.7-step gap is the alignment cost (Appendix C).

Generality. Sweeping each axis independently around the base configuration (Appendix D), three predictions hold. L^* is independent of K and of the noise level within ±6% while the RRR floor moves by an order of magnitude, so the transition is optimizational rather than statistical; it rises with r as σ_r falls and is flat in d at fixed $2\eta\sigma_1$. Factored GD diverges once $2\eta\sigma_1 \gtrsim 1$, and since $\sigma_1 \sim \tau(\sqrt{d} + \sqrt{r})$ the critical depth cannot be driven to zero by larger steps.

The transition survives preconditioning. Trained linear transformers implement preconditioned rather than vanilla descent (Ahn et al., 2023), which could equalize the per-mode rates. With $x \sim \mathcal{N}(0, \Sigma)$, preconditioned descent is factored descent on $\left\| BA - \widehat{W}_{\text{ols}} \right\|_F^2$, so the recursion holds exactly with the singular values of $\widehat{W}_{\text{ols}}$ (Appendix A.6). As predicted, at $\text{cond}(\Sigma) = 1, 10, 100$ the preconditioned L^* stays at $29.5, 29.0, 28.0$ while the vanilla one grows $32.0, 38.0, 54.5$ and the RRR floor moves by 3.6×.

What attention stores decides the reachable rate (Table 3). Every construction matches its recursion to 10^{-15} (Appendix J). Label-linear full heads never go below R_{lin}; a fixed rank-r head subspace plateaus at 50.5, near the bias bound $\tau^2 r(d-r)$, and matches RRR only when handed the task's U^*. Label nonlinearity without registers (Proposition 11) dips below R_{lin} but converges to OLS rather than RRR. Only register attention shows the plateau-then-drop transition and settles on the RRR floor.

Table 3: What attention stores decides the reachable rate. Mean risk against depth in gradient steps for five constructed linear-attention mechanisms, run as actual forward passes ($d = 20$, $r = 3$, $K = 140$, $\eta = 0.05$, 200 tasks; register attention uses $2L + 1$ layers). Reference values on the same tasks: RRR 0.89, $R_{\text{lin}} = 3.16$, OLS 3.35, chance 58.9.

depth	1	10	20	30	40	60	200	min
label-linear, full heads (3)	47.8	10.8	4.56	3.43	3.23	3.26	3.36	3.23
label-linear, fixed rank-r heads	57.3	51.7	50.7	50.5	50.5	50.5	50.5	50.5
label-linear, oracle rank-r heads	47.8	9.85	2.75	1.15	0.71	0.53	0.51	0.51
register-free, label-nonlinear	57.8	48.0	37.4	26.4	16.8	7.48	2.89	2.89
register attention (4)	58.9	58.6	22.2	4.99	1.38	0.89	0.89	0.89

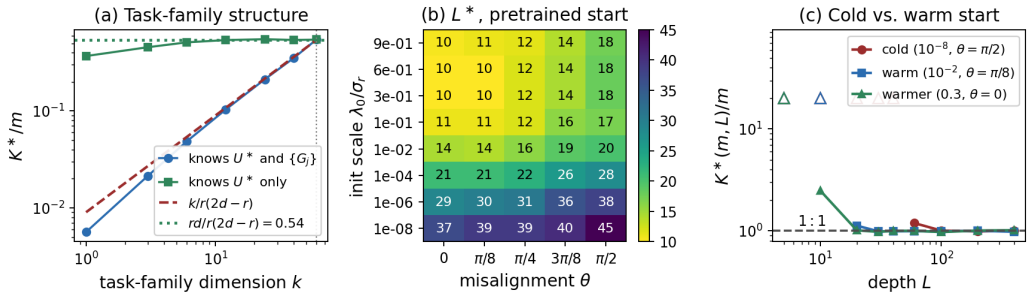


Figure 3: **(a)** Exchange rate achievable by estimators with increasing knowledge of the task family, as its effective dimension k varies ($n = 300$, 200 trials/point). **(b)** Critical depth over initialization scale and misalignment, where alignment uses only the family-level subspace ($K = 140$, 50 trials/cell). **(c)** The exchange rate of Figure 2b recomputed under family-level initializations.

Trained linear transformers. We also train multi-layer linear attention on the in-context task ($d = 10$, $r = 2$, $K = 70$, 2 seeds; Appendix J.6) with full heads, rank-r heads, or $r + d$ learned registers. Head-bottlenecked models stay above R_{lin} through 12 layers ($1.18R_{\text{lin}}$ at $L = 12$), as Lemma 8 predicts. Full-head models become label-nonlinear and reach the RRR floor at four layers, concentrated on the task subspace (alignment 0.98). The transition is present ($1.53R_{\text{lin}}$ at two layers) but far shallower than factored descent needs, and learned registers change the result by under 0.3%. Training thus finds label-nonlinear rank selection that keeps its state in the demonstration tokens: registers are one exact mechanism, not the one training selects here.

7 What amortization can and cannot buy

The 1:1 rate is specific to the task prior (Figure 3a). With A^* having i.i.d. entries the task family has exactly rd free parameters: it is the maximum-entropy family given the rank constraint, so beyond U^* there is nothing left to amortize, and Proposition 1's conclusion follows partly by construction. Real pretraining sees structured families. Giving the coefficient matrix a k-dimensional prior, $A^* = \sum_{j=1}^{k} c_j G_j$ with $c \sim \mathcal{N}(0, I_k)$ and $\{G_j\}$ a fixed orthonormal basis shared across tasks, so that $k = rd$ recovers our original setting, we compare three estimators: RRR, one given U^*, and one given U^* and $\{G_j\}$. At $k = rd$ the two informed estimators coincide at $K^*/m = 0.546$, against the predicted $rd/r(2d - r) = 0.541$, confirming Proposition 1's accounting. But the fully informed estimator's risk is $\sigma^2 k/n$ to measurement error, with $n \cdot \text{risk} = 1.0, 3.0, 6.1, 12.4, 24.9, 41.1, 63.3$ for $k = 1, 3, 6, 12, 24, 40, 60$, so its exchange rate tracks $k/r(2d - r)$ across two orders of magnitude, reaching 0.006 at $k = 1$.

The informed estimator is an oracle; unlike $\widehat{W}_r$, for which Proposition 2 constructs a mechanism, we do not show attention can realize it. The sweep shows the 1:1 rate is the $k = rd$ endpoint of a curve, so gains beyond a constant factor need structure over and above rank, the quantity Raventós et al. (2023) manipulate empirically.

Family-level initialization acts on the initialization (Figure 3b,c). Theorem 1 makes L^* a function of λ_0, and for register attention λ_0 is the register contents, so family-level initialization means pretrained or prompt-tuned registers. We let the task family share an output subspace U, which pretraining could learn, draw $\Delta W^* = UA^*$, and let the initialization align only B with U, never with the task-specific A^*, which would be oracle information.

The transition does not disappear under a warm start, it moves, and remains sharp (90%-to-10% width relative to L^*) at every initialization tested. The effect is substantial, with the depth at which K^*/m reaches parity dropping from $L \approx 100$ to $L \approx 20$. And L^* saturates near 10 steps rather than falling to zero, however favourable the initialization, because the task-specific coefficients must still be estimated from context. Family-level pretraining buys roughly a $4\times$ reduction in required depth and no more.

Where the exchange rate is not an identity. At matched budget register ICL and LoRA are the same algorithm (Lemma 1; Appendix E is an implementation check), so their agreement, and the 1:1 endpoint, follow from Proposition 2; the finite-depth transition measures only ICL's depth-capped budget. The exchange rate that compares different mechanisms is against label-linear attention: $K^*(m, L) \geq K_{\text{lin}}(m) = 3.34\, m$ at every depth, closed by what attention stores rather than by depth.

8 Discussion and Limitations

What is established and what is not. Theorem 1 is proved in the population limit under aligned balanced initialization and $4\eta\sigma \leq 1$, without linearization; the full qualifications and proof audit are in Appendix I and Appendix H. The finite-K, randomly initialized case and oscillatory regime remain outside the theorem, and the $r(2d - r)$ constant holds asymptotically (Appendix A.4).

The register assumption. Assumption 1 is architectural and met by deployed models, and Proposition 2 makes factored descent an identity rather than a surrogate. Label nonlinearity is proved necessary (Proposition 3); registers are proved sufficient but not necessary, and trained linear transformers reach the rank-r floor without them at shallower depth. Theorem 1 is therefore an exact guarantee for one realizable mechanism rather than a lower bound on the depth every trained model needs; whether the depth trained models require also scales with $\log(\sigma_r/\lambda_0)$ under controlled initialization is the natural next test. The construction uses hand-set weights and d anchor tokens; Proposition 3 assumes a Haar task subspace; trained models deeper than six layers diverged and are reported as missing; softmax attention is untested.

9 Conclusion

Placing in-context learning and low-rank fine-tuning on equal statistical footing requires deciding what attention can store. Label-linear attention, including any head bottleneck, pays asymptotically at least $d^2/[r(2d - r)]$ demonstrations per fine-tuning example, at every depth. Register attention, the architecture that system prompts and soft prompts already provide, implements factored descent exactly, and its exchange rate is a depth-indexed transition: 1:1 at the shared optimum, divergent below a critical effective depth, and at parity above it. We prove that transition's location in discrete time, two-sidedly and with constants independent of the initialization scale, which registers set. The 1:1 rate is specific to task families with no structure beyond rank; with $k < rd$ effective parameters it becomes $k/[r(2d-r)]$, and pretrained registers reduce critical depth by roughly $4\times$ while leaving the transition sharp. For register attention, amortization buys optimization budget rather than statistical efficiency. All of this holds for linear attention; softmax attention remains an empirical question.

Reproducibility statement

All estimators are implemented from scratch in numpy with fixed integer seeds, so results reproduce across processes. The complete source is in the supplementary material (Appendix K) and generates every number and figure. No external data or pretrained models are required; the numpy suite runs in about twenty minutes on one CPU core, and the attention experiments of Appendix J (`assumption1_gpu.py`) run on one GPU, with a float64 CPU reference for every construction. Appendix G discloses our use of generative AI and the verification applied to each AI-assisted component.

References

P.-A. Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. *Implicit regularization in deep matrix factorization*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 2(1):53–58, 1989.

Florentina Bunea, Yiyuan She, and Marten H. Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2), April 2011. ISSN 0090-5364. doi: 10.1214/11-aos876. URL http://dx.doi.org/10.1214/11-AOS876.

Mikhail S. Burtsev, Yuri Kuratov, Anton Peganov, and Grigory V. Sapunov. Memory transformer. *arXiv preprint arXiv:2006.11527*, 2020.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

Benoit Dherin, Michael Munn, Hanna Mazzawi, Michael Wunder, and Javier Gonzalvo. Learning without training: The implicit dynamics of in-context learning, 2026. URL https://arxiv.org/abs/2507.16003.

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.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/f39ae9ff3a81f499230c4126e01f421b-Paper.pdf.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4582–4597, 2021.

Sahand Negahban and Martin J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, 39(2):1069 – 1097, 2011. doi: 10.1214/10-AOS850. URL https://doi.org/10.1214/10-AOS850.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 14228–14246. Curran Associates, Inc., 2023. doi: 10.52202/075280-0626. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2e10b2c2e1aa4f8083c37dfe269873f8-Paper-Conference.pdf.

Gregory C. Reinsel and Raja Velu. Multivariate reduced-rank regression: Theory and applications. 1998. URL https://api.semanticscholar.org/CorpusID:120576794.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *CoRR*, abs/1312.6120, 2013. URL https://api.semanticscholar.org/CorpusID:17272965.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

Shifeng Xie, Rui Yuan, Simone Rossi, and Thomas Hannagan. The initialization determines whether in-context learning is gradient descent. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=fvqSKLDtJi.

Chenxiao Yang, Nathan Srebro, and Zhiyuan Li. Tight sample complexity of transformers. In Steve Hanneke and Tor Lattimore (eds.), *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 6887–6923. PMLR, 29 Jun–03 Jul 2026. URL https://proceedings.mlr.press/v336/yang26a.html.

Tian Ye and Simon S. Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

A Proofs

A.1 Closed form for reduced-rank regression (estimator 2)

We minimize $\|Y - XC\|_F^2$ over $C \in \mathbb{R}^{d\times d}$ with $\mathrm{rank}(C) \le r$, given X full column rank. Let $\hat{C}_{ols} = (X^\top X)^{-1} X^\top Y$ and $\hat{Y} = X\hat{C}_{ols}$, the orthogonal projection of Y onto $\mathrm{col}(X)$. For any

candidate C we have $XC \in \mathrm{col}(X)$, so by the Pythagorean identity

$$\|Y - XC\|_F^2 = \left\|Y - \hat{Y}\right\|_F^2 + \left\|\hat{Y} - XC\right\|_F^2.$$

The first term is independent of C, so we minimize the second. Since X has full column rank, $\mathrm{rank}(XC) = \mathrm{rank}(C)$, so this is the best rank-r approximation of $\hat{Y}$ within $\mathrm{col}(X)$. With $\hat{Y} = \hat{U}\hat{S}\hat{V}^\top$ and V_r the top-r columns of $\hat{V}$, Eckart–Young gives $\hat{Y}V_rV_r^\top$; its left singular vectors lie in $\mathrm{col}(\hat{Y}) \subseteq \mathrm{col}(X)$, so it is achievable, and by $X^+X = I_d$,

$$C^* = X^+\hat{Y}V_rV_r^\top = \hat{C}_{ols}V_rV_r^\top. \qquad \blacksquare$$

A.2 Closed form for L-step unconstrained GD (estimator 3)

The loss $\mathcal{L}(W) = \frac{1}{K}\left\|Y_{ctx} - X_{ctx}W^\top\right\|_F^2$ has gradient $\nabla_W\mathcal{L} = 2(W\hat{\Sigma}_K - \hat{S})$ with $\hat{\Sigma}_K = X_{ctx}^\top X_{ctx}/K$ and $\hat{S} = Y_{ctx}^\top X_{ctx}/K$. Gradient descent gives $W_{l+1} = W_l(I - 2\eta\hat{\Sigma}_K) + 2\eta\hat{S}$, whose fixed point is $W^* = \hat{S}\hat{\Sigma}_K^{-1} = \widehat{W}_{ols}^{(K)}$. Subtracting, $W_{l+1} - W^* = (W_l - W^*)(I - 2\eta\hat{\Sigma}_K)$, so $W_L = W^*[I - (I - 2\eta\hat{\Sigma}_K)^L]$ from $W_0 = 0$. $\blacksquare$

A.3 An independent derivation of the rank-r rate

We derive $\mathbb{E}\left\|\widehat{W}_r - \Delta W^*\right\|_F^2 = O(\sigma^2 rd/n)$ using only the closed form above and two standard facts about Gaussian matrices, stated as black boxes. The goal is not the tight minimax constant of Negahban & Wainwright (2011) or Bunea et al. (2011) but the same rate independently, so that Proposition 1 does not rest on trusting a single source. We work in the $C = W^\top$ parameterization.

Basic inequality. $\hat{C}_r$ exactly minimizes $\|Y - XC\|_F^2$ over rank-$\le r$ matrices, and $C^* = (\Delta W^*)^\top$ is feasible, so $\left\|Y - X\hat{C}_r\right\|_F^2 \le \|E\|_F^2$ where $Y = XC^* + E$. Writing $\Delta = C^* - \hat{C}_r$ and expanding,

$$\|X\Delta\|_F^2 \le 2\langle X\Delta, E\rangle = 2\langle \Delta, X^\top E\rangle \le 2\,\|\Delta\|_*\,\|X^\top E\|_{op},$$

by trace duality between the nuclear and operator norms.

Black-box facts. (a) Δ has rank at most $2r$, being a difference of two rank-r matrices, so Cauchy–Schwarz on its singular values gives $\|\Delta\|_* \le \sqrt{2r}\,\|\Delta\|_F$. (b) Since X acts on one side only, $\|X\Delta\|_F^2 \ge \lambda_{\min}(X^\top X)\,\|\Delta\|_F^2$ for every Δ, and for X with i.i.d. $\mathcal{N}(0,1)$ entries $\lambda_{\min}(X^\top X) \ge n/2$ with high probability once $n \gtrsim d$ (Vershynin, 2018); no restricted-isometry argument is needed. (c) For E with i.i.d. $\mathcal{N}(0,\sigma^2)$ entries independent of X, conditioning on X gives each column of $X^\top E$ distributed as $\sigma(X^\top X)^{1/2}g$, so $\|X^\top E\|_{op} = O(\sigma\sqrt{nd})$ with high probability (e.g. Vershynin, 2018).

Combining,

$$\tfrac{n}{2}\,\|\Delta\|_F^2 \le \|X\Delta\|_F^2 \le 2\sqrt{2r}\,\|\Delta\|_F \cdot O(\sigma\sqrt{nd}) \quad\Longrightarrow\quad \|\Delta\|_F = O(\sigma\sqrt{rd/n}),$$

matching the order of the rate used in Proposition 1. Fact (b) needs only $n \gtrsim d$. The exact leading constant is derived next. $\blacksquare$

A.4 The asymptotic reduced-rank constant

Write $\hat{\Sigma} = X^\top X/n$ and $\widehat{W}_{\mathrm{ols}} = Y^\top X(X^\top X)^{-1}$, so $\widehat{W}_{\mathrm{ols}} - \Delta W^* = E^\top X(X^\top X)^{-1}$. Let Π_r denote the best rank-r approximation in Frobenius norm and $\widehat{W}_{\mathrm{tr}} = \Pi_r(\widehat{W}_{\mathrm{ols}})$ the truncated OLS estimator. Since $\left\|Y - XW^\top\right\|_F^2 = \left\|Y - X\widehat{W}_{\mathrm{ols}}^\top\right\|_F^2 + n\left\|(\widehat{W}_{\mathrm{ols}} - W)\hat{\Sigma}^{1/2}\right\|_F^2$ and $\mathrm{rank}(W\hat{\Sigma}^{1/2}) = \mathrm{rank}(W)$, the reduced-rank estimator is

$$\widehat{W}_r = \Pi_r\big(\widehat{W}_{\mathrm{ols}}\hat{\Sigma}^{1/2}\big)\hat{\Sigma}^{-1/2}, \tag{7}$$

equivalent to the closed form derived above.

Proposition 5 (Asymptotic reduced-rank constant)**.** *Fix ΔW^* of rank r with $\sigma_r(\Delta W^*) > 0$. For $\widehat{W} \in \{\widehat{W}_r, \widehat{W}_{\mathrm{tr}}\}$, as $n \to \infty$, $n\left\|\widehat{W} - \Delta W^*\right\|_F^2 \Rightarrow \sigma^2 \chi^2_{r(2d-r)}$ and $n\,\mathbb{E}\left\|\widehat{W} - \Delta W^*\right\|_F^2 \to \sigma^2 r(2d-r)$. The same limit holds after averaging over the task prior. Hence the two mechanisms of Proposition 1 share the leading constant $C_r = \sigma^2 r(2d-r)$.*

Proof. Step 1. Conditional on X, the rows of $\widehat{W}_{\mathrm{ols}} - \Delta W^*$ are independent $\mathcal{N}(0, \sigma^2(X^\top X)^{-1})$, and $n(X^\top X)^{-1} \to I$ almost surely, so $\sqrt{n}(\widehat{W}_{\mathrm{ols}} - \Delta W^*) \Rightarrow Z$ with i.i.d. $\mathcal{N}(0, \sigma^2)$ entries.

Step 2. The set $\mathcal{M}_r$ of rank-r matrices is a smooth embedded manifold of dimension $r(2d-r)$, with tangent space $T_W = \{UA + BV^\top\}$ at $W = U\Sigma_r V^\top$. On the open set $\{\sigma_r > \sigma_{r+1}\}$, Π_r is single-valued and smooth, and its derivative at any $W \in \mathcal{M}_r$ is the orthogonal projection P_{T_W} (Absil & Malick, 2012).

Step 3. Let $F(Z, S) = \Pi_r(ZS^{1/2})S^{-1/2}$, which is smooth near $(\Delta W^*, I)$, and note $F(\Delta W^*, S) = \Delta W^*$ for all S near I. By (7), $\widehat{W}_r - \Delta W^* = F(\widehat{W}_{\mathrm{ols}}, \hat{\Sigma}) - F(\Delta W^*, \hat{\Sigma}) = D_Z F(\Delta W^*, \hat{\Sigma})[\widehat{W}_{\mathrm{ols}} - \Delta W^*] + O(\left\|\widehat{W}_{\mathrm{ols}} - \Delta W^*\right\|_F^2)$ on an event of probability tending to one, and $D_Z F(\Delta W^*, S) \to P_T$ as $S \to I$. By Step 1 and Slutsky, $\sqrt{n}(\widehat{W}_r - \Delta W^*) \Rightarrow P_T Z$. The truncated estimator is the case $S = I$. Since Z is isotropic and P_T has rank $r(2d-r)$, $\|P_T Z\|_F^2 \sim \sigma^2 \chi^2_{r(2d-r)}$.

Step 4. For convergence of the mean it suffices that $n\left\|\widehat{W} - \Delta W^*\right\|_F^2$ has bounded second moments.

- For $\widehat{W}_{\mathrm{tr}}$: ΔW^* has rank r, so $\left\|\widehat{W}_{\mathrm{tr}} - \Delta W^*\right\|_F \le 2\left\|\widehat{W}_{\mathrm{ols}} - \Delta W^*\right\|_F$, and $\mathbb{E}[\left\|\widehat{W}_{\mathrm{ols}} - \Delta W^*\right\|_F^4 \mid X] \le 3\sigma^4 d^2 \operatorname{tr}((X^\top X)^{-1})^2$, whose n^2-scaled expectation is bounded for $n > d + 3$ by inverse-Wishart second moments.
- For $\widehat{W}_r$: ΔW^* is feasible in the X-metric, so $\left\|X(\widehat{W}_r - \Delta W^*)^\top\right\|_F \le 2\|P_X E\|_F$ and $n\left\|\widehat{W}_r - \Delta W^*\right\|_F^2 \le 4\|P_X E\|_F^2\, n \operatorname{tr}((X^\top X)^{-1})$, with $\|P_X E\|_F^2 \sim \sigma^2 \chi^2_{d^2}$ independent of X.

Step 5. These bounds do not depend on ΔW^*, and $\sigma_r(\Delta W^*) > 0$ almost surely under the prior, so dominated convergence gives the prior-averaged limit. □

Table 1 verifies the mean. Rerunning the same extrapolation for $\widehat{W}_{\mathrm{tr}}$ gives 36.2, 111.9, 203.3 and 303.8 against 36, 111, 204 and 304. The variance of $n\cdot$risk matches the χ^2 value $2r(2d-r)$ to within 14% (ratios 0.86, 1.10, 1.00 and 1.06 for RRR).

Endpoint of factored descent. Under Assumption 1 and Proposition 2, the infinite-depth ICL estimator is the global minimizer of the factored objective whenever descent converges to one. For the vanilla update the landscape has no spurious local minima (Baldi & Hornik, 1989), the minimizer is $\widehat{W}_r$, and convergence is observed in every run (Appendix F). For the preconditioned update the objective is $\left\|BA - \widehat{W}_{\mathrm{ols}}\right\|_F^2$ (Appendix A.6), the minimizer is $\widehat{W}_{\mathrm{tr}}$, and global convergence from small random initialization is proved by Ye & Du (2021). In either case Proposition 5 supplies the shared constant used in Proposition 1.

A.5 The matched-budget identity

Lemma 1 (Matched budget is an identity)**.** *Under Assumption 1, with ICL realized by Proposition 2, a fine-tune run for L factored steps on m examples, with the same step size and initialization law, has exactly the risk distribution of the ICL estimator with $K = m$ and depth L.*

Proof. Both are the same measurable map applied to m i.i.d. pairs from the same law. □

A.6 FINITE-K EXACTNESS OF THE PRECONDITIONED UPDATE AND A K-FREE DEPTH FLOOR

Proposition 6 (Finite-K exactness). *For any $K \geq d$ with $\hat{\Sigma}_K$ invertible, $\nabla_W \widehat{\mathcal{L}}_K \hat{\Sigma}_K^{-1} = 2(W\hat{\Sigma}_K - \hat{S}_K)\hat{\Sigma}_K^{-1} = 2(W - \widehat{W}_{\text{ols}}^{(K)})$. Preconditioned factored descent on K examples is therefore factored descent on $\left\| BA - \widehat{W}_{\text{ols}}^{(K)} \right\|_F^2$. With Assumption 2 on the top-r singular pairs of $\widehat{W}_{\text{ols}}^{(K)} = \sum_i \hat{\sigma}_i \hat{u}_i \hat{v}_i^\top$, Proposition 4 and Theorem 1 hold exactly with $\sigma_i \mapsto \hat{\sigma}_i$, and if $4\eta\hat{\sigma}_1 \leq 1$ then $BA \to \widehat{W}_{\text{tr}}$.*

Proof. In Lemma 2, the components $i > r$ of the target are annihilated, because $\hat{v}_i^\top A_l^\top = 0$ and $B_l^\top \hat{u}_i = 0$. The rest of Appendix H applies verbatim, and Proposition 8(i) gives the limit. □

Corollary 1 (A K-free depth floor, any initialization). *Let $s_0 = \max(\|B_0\|_{\text{op}}, \|A_0\|_{\text{op}}) \leq \varepsilon_0$ and $\theta = \|\Delta W^*\|_F / (2\sqrt{r})$. For the vanilla update, let $\rho_K = 1 + 2\eta(\|\hat{S}_K\|_{\text{op}} + \|\hat{\Sigma}_K\|_{\text{op}}\theta)$; for the preconditioned update, let $\rho_K = 1 + 2\eta(\|\widehat{W}_{\text{ols}}^{(K)}\|_{\text{op}} + \theta)$. If $L \leq \log(\theta/\varepsilon_0^2)/(2\log\rho_K)$, then $\|B_L A_L - \Delta W^*\|_F^2 \geq \frac{1}{4}\|\Delta W^*\|_F^2$, for every K.*

Proof. With $G_l = 2(B_l A_l \hat{\Sigma}_K - \hat{S}_K)$ and $s_l = \max(\|B_l\|_{\text{op}}, \|A_l\|_{\text{op}})$, $s_{l+1} \leq s_l(1 + 2\eta\|\hat{S}_K\|_{\text{op}} + 2\eta\|\hat{\Sigma}_K\|_{\text{op}} s_l^2)$. Induction gives $s_l \leq \varepsilon_0 \rho_K^l$ as long as $\varepsilon_0^2 \rho_K^{2l} \leq \theta$. Then $\|B_L A_L\|_{\text{op}} \leq \theta$, and since $B_L A_L$ has rank at most r, $\|B_L A_L\|_F \leq \|\Delta W^*\|_F / 2$. The preconditioned case is identical. □

For $K \geq d$, standard Gaussian singular-value bounds (Vershynin, 2018) give $\rho_K \leq 1 + 2\eta c(\sigma_1 + \sigma + \theta)$ with $c = (2 + \sqrt{2})^2$, with probability at least $1 - 4e^{-d}$. Below depth $\Theta(\log 1/\varepsilon_0)$ the risk therefore stays at a constant fraction of chance for every context length, so $K^*(m, L) = \infty$ there. This is the divergence seen in Figure 2b, and it holds for random initialization and vanilla descent. The bound is not tight: its slope involves σ_1 rather than σ_r.

B STEP-SIZE DEPENDENCE OF THE DISCRETIZATION CORRECTION

The η sweep in Appendix D is the cleanest test of the discrete-time correction because the continuous and discrete predictions coincide as $\eta \to 0$ and separate as $\eta\sigma$ grows. At $\eta = 0.005$ the measured-to-continuous and measured-to-discrete ratios are both 1.29; at $\eta = 0.08$ they are 1.70 and 1.36, respectively. The ratio of these two errors, $1.70/1.36 = 1.25$, agrees with the predicted correction $0.526/0.423 = 1.24$. Thus the discretization effect is independently visible across more than an order of magnitude in step size, without relying on the aggregate slope test.

C CRITICAL-DEPTH EXPERIMENT DETAIL

Base configuration $d = 20$, $r = 3$, $K = 140$, $\eta = 0.05$, 60 trials per point, one checkpointed trajectory per trial so that the first crossing is a genuine first crossing rather than a minimum over independently initialized runs.

Threshold invariance. Slope and intercept of L^* against $\log(1/\varepsilon_0)$ as the convergence threshold κ varies:

κ	slope	intercept	R^2	$\kappa\cdot$floor $> \sigma_r^2$
1.1	3.225	12.74	0.990	0.0%
1.5	3.378	9.27	0.980	0.0%
2.0	3.340	8.43	0.982	0.0%
3.0	3.340	7.43	0.982	0.0%
5.0	3.354	6.09	0.984	0.6%
20.0	2.692	2.88	0.993	95.8%

Over $\kappa \in [1.1, 5]$ the slope varies by 4.7% while the intercept falls by more than half, which is the behaviour Eq. 3 requires. The $\kappa = 20$ row is expected to deviate, for a reason the last column makes explicit. A trial whose threshold satisfies $\kappa \cdot \text{floor} > \sigma_r^2$ can cross it while the slowest mode is still

unlearned, because the residual risk at that point is about σ_r^2. The crossing is then set by mode $r-1$, and Eq. 3 at σ_r does not apply. This happens in at most 0.6% of trials for $\kappa \leq 5$ and in 95.8% at $\kappa = 20$ (floor 0.894, median $\sigma_r^2 = 11.5$). The fractions are computed on the same seeded tasks as the sweep.

Functional form of the residual. Measured critical depth against the Eq. 3 prediction, per initialization scale, at $\kappa = 2$:

	random init				aligned init			
ε_0	emp.	pred.	ratio	diff.	emp.	pred.	ratio	diff.
3×10^{-2}	21.0	15.1	1.393	5.93	17.0	14.3	1.191	2.73
10^{-2}	24.0	18.3	1.314	5.74	21.0	18.1	1.162	2.93
3×10^{-3}	27.0	22.4	1.205	4.60	26.0	21.8	1.194	4.23
10^{-3}	30.0	26.2	1.146	3.83	29.0	27.3	1.063	1.73
3×10^{-4}	36.0	30.1	1.196	5.90	32.0	30.3	1.057	1.72
10^{-4}	40.0	34.0	1.177	6.02	38.0	34.7	1.095	3.31

The ratio drifts monotonically (-0.037 per log unit for random initialization) while the difference does not ($+0.005$ per log unit). Fitting each one-parameter model, the additive form leaves an RMS residual of 0.83 steps against 1.82 for the multiplicative form, and for aligned initialization 0.88 against 1.42. Bootstrap over trials gives $c_0 = 5.55\,[4.83, 6.41]$ steps for random initialization and $2.89\,[2.30, 3.48]$ for aligned. The difference between the arms is the alignment cost, 2.7 steps. The convergence tolerance, computed as $\log((1-\delta)/\delta)/2\log(1+2\eta\sigma_r)$ with $\delta = \sqrt{(\kappa-1)\cdot \text{floor}}/\sigma_r$, predicts 1.54 steps. The two together account for 4.2 of the measured 5.55.

Slope estimates. Bootstrap over trials, 4000 resamples: random initialization $3.354\,[3.040, 3.773]$ against a prediction of 3.515; aligned initialization $3.487\,[3.140, 3.821]$ against 3.570. The continuous-time expression evaluated at the mean top-r singular value gives 2.293, outside both intervals.

D GENERALITY SWEEP

One-at-a-time sweeps around the base configuration, $\varepsilon_0 = 10^{-3}$, 40 trials per point. Columns e/c and e/d are the ratios of the measured critical depth to the continuous-time and discrete-time predictions.

sweep	value	$2\eta\sigma_1$	σ_r	L^* emp.	contin.	e/c	e/d
η	0.005	0.05	3.29	302	234	1.29	1.29
	0.01	0.11	3.28	148	116	1.27	1.27
	0.02	0.22	3.30	74	58	1.28	1.26
	0.05	0.56	3.41	32.5	22.8	1.43	1.23
	0.08	0.84	3.29	24	14.1	1.70	1.36
	0.10	1.08		diverges in 31/40 trials			
d, $K = 7d$, $\eta \propto d^{-1/2}$	10	0.57	1.93	32.0	28.7	1.12	0.99
	20	0.54	3.41	32.5	22.8	1.43	1.23
	40	0.51	5.05	33.0	21.6	1.53	1.32
	80	0.49	7.75	31.0	20.1	1.54	1.30
r	1	0.44	4.41	26.0	18.6	1.40	1.16
	2	0.51	3.71	30.0	21.2	1.41	1.21
	3	0.55	3.49	32.0	22.2	1.44	1.23
	6	0.61	2.47	39.5	28.9	1.37	1.22
	10	0.69	1.79	49.0	39.7	1.23	1.15
K	70	0.54	3.36	32.0	23.0	1.39	1.21
	140	0.54	3.41	32.5	22.8	1.43	1.23
	280	0.55	3.48	31.0	22.3	1.39	1.20
	560	0.54	3.33	32.0	22.8	1.41	1.22
σ	0.5	0.54	3.4	34.5	22.0	1.57	1.34
	1.0	0.54	3.4	34.0	23.7	1.44	1.24
	2.0	0.54	3.4	29.0	22.5	1.29	1.12

The K and σ rows are the substantive checks. The RRR floor moves from 0.20 to 2.02 across the K sweep and from 0.22 to 3.72 across the σ sweep, an order of magnitude in each, while L^* stays within $\pm 6\%$. The critical depth is set by the optimization problem, not the statistical one.

E MATCHED-BUDGET CONSISTENCY CHECK

By Lemma 1, the ICL and LoRA arms below have identically distributed risk under Assumption 1. The table therefore checks the implementation; it is not evidence for any claim about ICL. The third arm is not covered by the identity. At $K = m = 140$, $\eta = 0.05$, $\varepsilon_0 = 10^{-3}$, 120 trials per depth. The ICL arm runs L factored-GD steps on K context examples; the LoRA arm runs L steps on m fine-tuning examples from an independent draw; the third arm uses the standard LoRA initialization $B = 0$, $A \sim \mathcal{N}(0, 1/d)$. The converged RRR floor is 0.892 ± 0.010. Intervals are two one-sided tests at $\alpha = 0.05$, equivalently a 90% confidence interval on the difference.

L	ICL	LoRA	LoRA (B=0)	90% CI on difference	bound as % of LoRA
5	60.92	60.92	38.9	$[-2.361,\ 2.361]$	3.9
10	60.20	60.16	12.6	$[-2.097,\ 2.165]$	3.6
20	21.95	21.50	2.24	$[-1.064,\ 1.972]$	9.2
30	4.386	4.498	0.980	$[-0.999,\ 0.776]$	22.2
40	1.276	1.370	0.884	$[-0.446,\ 0.258]$	32.6
60	0.900	0.902	0.902	$[-0.032,\ 0.027]$	3.5
100	0.881	0.863	0.863	$[-0.010,\ 0.046]$	5.4
200	0.889	0.877	0.877	$[-0.015,\ 0.040]$	4.6
400	0.864	0.900	0.900	$[-0.065,\ -0.007]$	7.2
1000	0.900	0.885	0.885	$[-0.014,\ 0.045]$	5.1
2500	0.876	0.906	0.906	$[-0.059,\ -0.002]$	6.5

All intervals are consistent with the identity. Two of eleven exclude zero narrowly, which matches the 10% per-interval error rate of 90% intervals. The substantive observation is the third arm that the standard $B = 0$ initialization, which converges about twice as fast as the symmetric small-random one, because only one factor starts small, which is itself a mild instance of the warm-start effect of Section 7.

F ADDITIONAL EXPERIMENTAL DETAIL

Gradient step size $\eta = 0.05$ was selected by a stability sweep. At $d = 20$, $r = 3$, $n = 140$, factored GD from initialization 10^{-3} run to $L = 4000$: $\eta = 0.2$ diverges by step 16; $\eta = 0.1$ has not converged by $L = 4000$ (risk 2.95 against an RRR floor of 0.87); and $\eta \in \{0.05, 0.02, 0.01, 0.005\}$ all converge to the floor to ten decimal places, with smaller η requiring proportionally more steps. Larger step sizes remain stable for the unconstrained parameterization while diverging for the factored one, which is a small illustration of the qualitatively different landscape induced by the rank constraint.

G DISCLOSURE OF GENERATIVE AI USE

Per ICLR policy we describe how generative AI tools were used in preparing this work, and, for each use, what verification was applied. This section is not counted towards the page limit.

What was done without AI assistance. The research question and the framing of the comparison; the problem formulation of Section 3, including the identification of reduced-rank regression as the object shared by in-context learning and low-rank fine-tuning and the argument establishing it, except its attention-level step, which is disclosed below; Proposition 1 and the exchange-rate statement; and the closed forms in Appendix A, namely the reduced-rank closed form, the L-step gradient-descent closed form, and the from-scratch derivation of the rank-r rate. These are the authors' own work and were not produced with generative AI.

Required disclosures. We used generative AI tools for the following.

Formulating mathematical claims and assisting in a derivation. An earlier version of this work stated the critical depth in continuous time, following the gradient-flow convention standard in the matrix-factorization literature. AI assistance identified that this idealization can carry $O(1)$ error in the regime where the rank bottleneck binds, because $2\eta\sigma_r$ is not small there, and suggested analyzing the corresponding discrete escape-time form, together with the observation that the initialization enters as ε_0^2 through the product BA. It also suggested that the residual should be additive rather than multiplicative. The authors subsequently developed and proved Theorem 1: the exact telescoping identity, the drag-term bounds, the explicit constants, the monotone-regime hypothesis, and the stability analysis are the authors' own work. The per-mode test in Table 2 and the independent verification sweep were then used to check the theorem numerically against the exact recursion and measurements. Thus AI assistance helped identify the continuous-time limitation and motivate the discrete formulation, whereas the central theorem and its proof are attributable to the authors. In preparing the final version, AI assistance was also used to identify circular or inconsistent steps in a draft. These were a shared leading constant that was assumed rather than derived, a matched-budget comparison that is an identity under Assumption 1, and an explanation of the $\kappa = 20$ deviation that appealed to linearization. AI assistance also drafted the proofs of Proposition 5, Lemma 1, Proposition 6, Corollary 1 and the extension of Theorem 1 to $2\eta\sigma < 1$ (Appendix H.4). The authors checked each proof, and each is verified numerically (Appendix H.6).

Designing and giving feedback on experiments. The experiment suite of Sections 6 and 7 was specified with AI assistance, including the threshold-invariance test, the extrapolation of the leading finite-sample correction in Table 1, the task-family structure sweep, the matched-budget control of Appendix E, and the preconditioning test. AI-generated adversarial review of drafts also identified two errors that we corrected: an earlier warm-start experiment initialized the optimizer using the singular subspace of the individual test task rather than the subspace shared by the task family, which is oracle information no pretrained model could hold; and an earlier critical-depth sweep drew a fresh initialization for each candidate depth, so that a first threshold crossing was in effect a minimum over independent runs.

Implementing methods and generating synthetic data. The code in Appendix K was written with substantial AI assistance. All data in this paper is synthetic and is generated by that code. This was a deliberate choice: AI assistance let us test the claims far more thoroughly than we otherwise would have, which is why the paper reports a (d, r) sweep, a 21-configuration generality sweep, bootstrap intervals, equivalence intervals, and a stability boundary rather than a single working configuration.

Interpreting results. AI assistance was used in diagnosing several discrepancies between prediction and measurement, including the decomposition of the theory-experiment gap in the critical depth into a singular-value plug-in error and a discretization error.

We did not use generative AI to propose the research topic or the hypotheses. Assistance with translation, cleaning or reformatting external datasets, qualitative or thematic analysis, survey or interview design, and transcription is not applicable to this work.

Recommended disclosures. We additionally used generative AI to create the figures, which are produced by the plotting code in Appendix K; to draft and edit prose for readability; to suggest the arrangement of sections; and to generate critical reviews of the manuscript, several of whose objections we adopted, including the scoping of the 1:1 result to task families without structure beyond their rank.

Verification. Every AI-assisted component was checked against something independent of the assistance that produced it. Estimators with a closed form are verified against it: the OLS risk matches $\sigma^2 d^2/(n-d-1)$ to within 1.7% at every sample size tested (Table 1), and the reduced-rank estimator is computed from the Eckart–Young characterization proved in Appendix A. The central rate is corroborated three ways: two independent classical sources, our own derivation, and a four-configuration numerical sweep. Theorem 1 is proved analytically from the exact recursion and is additionally tested per mode rather than only in aggregate; the verification grid checks all 160 tested parameter cells over the full stable range. Seeds are fixed integers and no result depends on process-dependent hashing, so every number reproduces. Where a claim rests on an assumption rather than a proof, the paper says so: Proposition 1 relies on factored descent converging to its global minimizer (proved for the preconditioned update, observed for the vanilla one), Theorem 1 is an aligned-initialization result, and the estimator of Section 7 is an oracle rather than a mechanism.

Responsibility. We have reviewed all AI-assisted work and we take responsibility for the final content of this paper, including its text, claims, figures, and code.

H Proof of Theorem 1

For completeness, the factor update used by the proof is

$$B_{l+1} = B_l - 2\eta(B_l A_l - \Delta W^*)A_l^\top, \qquad A_{l+1} = A_l - 2\eta B_l^\top(B_l A_l - \Delta W^*). \tag{8}$$

Under the aligned ansatz this gives

$$b_i^{(l+1)} = b_i^{(l)} + 2\eta a_i^{(l)}(\sigma_i - \lambda_i^{(l)}), \qquad a_i^{(l+1)} = a_i^{(l)} + 2\eta b_i^{(l)}(\sigma_i - \lambda_i^{(l)}). \tag{9}$$

We give the full proof in the aligned population setting. The argument follows from the exact factor recursion, conservation of balancedness, the invariant interval $[0, \sigma]$, the telescoping identity, a geometric bound on the growth-phase drag, and geometric contraction of the saturation deficit.

H.1 Exact reduction to a scalar recursion

Lemma 2 (The aligned ansatz is invariant)**.** *Under (8), if $B_l = \sum_i b_i^{(l)} u_i e_i^\top$ and $A_l = \sum_i a_i^{(l)} e_i v_i^\top$, then B_{l+1}, A_{l+1} have the same form, with $\lambda_i^{(l)} := b_i^{(l)} a_i^{(l)}$ and*

$$b_i^{(l+1)} = b_i^{(l)} + 2\eta a_i^{(l)}\big(\sigma_i - \lambda_i^{(l)}\big), \qquad a_i^{(l+1)} = a_i^{(l)} + 2\eta b_i^{(l)}\big(\sigma_i - \lambda_i^{(l)}\big). \tag{10}$$

Proof. $B_l A_l = \sum_i \lambda_i^{(l)} u_i v_i^\top$, so $B_l A_l - \Delta W^* = \sum_i (\lambda_i^{(l)} - \sigma_i) u_i v_i^\top$. Using $v_i^\top v_j = \delta_{ij}$,

$$\big(B_l A_l - \Delta W^*\big)A_l^\top = \Big(\sum_i (\lambda_i^{(l)} - \sigma_i) u_i v_i^\top\Big)\Big(\sum_j a_j^{(l)} v_j e_j^\top\Big) = \sum_i a_i^{(l)}(\lambda_i^{(l)} - \sigma_i)\, u_i e_i^\top,$$

which gives the first identity in (10); the second follows symmetrically from $u_i^\top u_j = \delta_{ij}$. In particular no mode jei enters the update of mode i: the modes are exactly decoupled, not approximately so. □

Lemma 3 (Balancedness is conserved). *If $b_i^{(0)} = a_i^{(0)}$ then $b_i^{(l)} = a_i^{(l)}$ for every $l \geq 0$.*

Proof. Immediate induction: (10) adds the same increment $2\eta b_i^{(l)}(\sigma_i - \lambda_i^{(l)})$ to both when $b_i^{(l)} = a_i^{(l)}$. □

Proposition 7 (Exact scalar recursion). *Under Assumptions 2, each mode evolves independently as*

$$\lambda_{l+1} = g(\lambda_l), \qquad g(\lambda) := \lambda\big(1 + 2\eta(\sigma - \lambda)\big)^2, \tag{11}$$

where we have dropped the mode index.

Proof. By Lemma 3, (10) reads $b_{l+1} = b_l(1 + 2\eta(\sigma - \lambda_l))$ and identically for a_l, so $\lambda_{l+1} = b_{l+1}a_{l+1} = \lambda_l(1 + 2\eta(\sigma - \lambda_l))^2$. □

Everything below concerns (11) and involves *no* approximation of g.

H.2 The deficit form and the three regimes

Write $u_l := \sigma - \lambda_l$ for the deficit.

Lemma 4 (Exact deficit recursion). *For all $\lambda \in [0, \sigma]$, with $u = \sigma - \lambda$,*

$$\sigma - g(\lambda) = u\big[1 - 4\eta\lambda(1 + \eta u)\big]. \tag{12}$$

Proof. Expanding, $(\sigma - u)(1 + 2\eta u)^2 = \sigma + 4\eta\sigma u + 4\eta^2\sigma u^2 - u - 4\eta u^2 - 4\eta^2 u^3$, so

$$\sigma - g(\lambda) = u\big[1 - 4\eta\sigma + 4\eta u - 4\eta^2\sigma u + 4\eta^2 u^2\big] = u\big[1 + 4\eta(u - \sigma)(1 + \eta u)\big],$$

the last step by factoring $-\sigma + u - \eta\sigma u + \eta u^2 = (u - \sigma)(1 + \eta u)$. Substituting $u - \sigma = -\lambda$ gives (12). □

Proposition 8 (Invariant interval, monotonicity, stability). *Fix $\sigma, \eta > 0$.*

(i) If $4\eta\sigma \leq 1$ and $\lambda_0 \in (0, \sigma)$, then (λ_l) is strictly increasing, remains in $(0, \sigma)$, and $\lambda_l \uparrow \sigma$.

(ii) $g'(\sigma) = 1 - 4\eta\sigma$, so σ is a locally attracting fixed point if and only if $2\eta\sigma < 1$.

(iii) If $4\eta\sigma > 1$ the iteration overshoots σ; if $2\eta\sigma > 1$ it does not converge to σ from small initialization.

Proof. (i) Let $h(u) := 4\eta(\sigma - u)(1 + \eta u)$, so (12) reads $u' = u[1 - h(u)]$. Then $h'(u) = 4\eta(\eta\sigma - 1 - 2\eta u)$. Since $4\eta\sigma \leq 1$ forces $\eta\sigma \leq \frac{1}{4} < 1$, we get $h'(u) < 0$ on $[0, \sigma]$, hence $h(u) < h(0) = 4\eta\sigma \leq 1$ for every $u \in (0, \sigma]$. Therefore $0 < 1 - h(u) < 1$, giving $0 < u' < u$, i.e. $\lambda < \lambda' < \sigma$. The sequence is increasing and bounded, so it converges to a fixed point of g in $(0, \sigma]$; but $g(\lambda) = \lambda$ with $\lambda > 0$ forces $(1 + 2\eta u)^2 = 1$, i.e. $u = 0$. Hence $\lambda_l \uparrow \sigma$.

(ii) $g'(\lambda) = (1 + 2\eta u)^2 - 4\eta\lambda(1 + 2\eta u)$; at $\lambda = \sigma$, $u = 0$, so $g'(\sigma) = 1 - 4\eta\sigma$, and $|1 - 4\eta\sigma| < 1 \iff 0 < 4\eta\sigma < 2$.

(iii) If $4\eta\sigma > 1$ then $h(0^+) > 1$, so $u' < 0$ once u is small, i.e. λ crosses σ. If $2\eta\sigma > 1$ then $|g'(\sigma)| > 1$ and σ is repelling. □

Remark 1 (The stability boundary is derived). *Section 6 reports that factored GD diverges once $2\eta\sigma_1 \gtrsim 1$ (31/40 trials at $\eta = 0.1$, stable for $\eta \leq 0.05$). Proposition 8(ii) identifies this as $|g'(\sigma_1)| < 1$ and thus explains it exactly. Since $\sigma_1 \asymp \tau(\sqrt{d} + \sqrt{r})$, the admissible step size shrinks like $d^{-1/2}$, which is why the critical depth cannot be driven to zero by taking larger steps.*

H.3 THE TELESCOPING IDENTITY

Lemma 5 (Telescoping identity)**.** *Under Proposition 8(i), for every $T \geq 0$,*

$$\log \frac{\lambda_T}{\lambda_0} = 2 \sum_{l=0}^{T-1} \log\bigl(1 + 2\eta u_l\bigr). \tag{13}$$

Consequently, defining the drag

$$D_T := \sum_{l=0}^{T-1} \log \frac{1 + 2\eta\sigma}{1 + 2\eta u_l} \;\geq\; 0, \tag{14}$$

we have the exact identity

$$T = \frac{\log(\lambda_T/\lambda_0) + 2D_T}{2\log(1 + 2\eta\sigma)}. \tag{15}$$

Proof. Take logarithms in (11) (all factors are positive) and sum; this is (13). Nonnegativity of each summand of D_T follows from $u_l \leq \sigma$. Adding and subtracting $2T\log(1 + 2\eta\sigma)$ in (13) and rearranging gives (15). □

Identity (15) is the engine of the whole argument: it says the escape-time formula $\log(\lambda_T/\lambda_0)/2\log(1 + 2\eta\sigma)$ is exact up to the single term D_T. The theorem reduces to bounding D_T by a constant that does not depend on λ_0.

Lemma 6 (Drag bound in the growth phase)**.** *Let $T_1 := \min\{l : \lambda_l \geq \sigma/2\}$, finite by Proposition 8(i). Then*

$$D_{T_1} \;\leq\; 2\eta \sum_{l<T_1} \lambda_l \;\leq\; \frac{(1 + \eta\sigma)^2}{2} \;\leq\; \frac{25}{32}. \tag{16}$$

Proof. For each summand, using $\sigma - u_l = \lambda_l$ and $\log(1 + x) \leq x$,

$$\log \frac{1 + 2\eta\sigma}{1 + 2\eta u_l} = \log\Bigl(1 + \frac{2\eta\lambda_l}{1 + 2\eta u_l}\Bigr) \leq \frac{2\eta\lambda_l}{1 + 2\eta u_l} \leq 2\eta\lambda_l,$$

which gives the first inequality. For the second, note that $l < T_1$ implies $\lambda_l < \sigma/2$, hence $u_l > \sigma/2$ and $\lambda_{l+1} = \lambda_l(1 + 2\eta u_l)^2 > \rho\lambda_l$ with $\rho := (1 + \eta\sigma)^2 > 1$. Therefore $\lambda_l \leq \lambda_{T_1-1}\rho^{-(T_1-1-l)}$ and

$$\sum_{l<T_1} \lambda_l \;\leq\; \lambda_{T_1-1} \sum_{j\geq 0} \rho^{-j} = \lambda_{T_1-1} \frac{\rho}{\rho - 1} < \frac{\sigma}{2} \cdot \frac{(1 + \eta\sigma)^2}{2\eta\sigma},$$

using $\rho - 1 = 2\eta\sigma + \eta^2\sigma^2 \geq 2\eta\sigma$. Multiplying by 2η gives $(1 + \eta\sigma)^2/2$, and $\eta\sigma \leq \frac{1}{4}$ gives $25/32$. (If $T_1 = 0$ the sum is empty and $D_0 = 0$.) □

This is the step that makes the offset *additive*: geometric growth means $\sum_l \lambda_l$ is dominated by its final terms, which are $O(\sigma)$ regardless of how small λ_0 was.

Lemma 7 (Saturation phase)**.** *Let $\delta \in (0, \frac{1}{2}]$. For every $j \geq 0$, $u_{T_1+j} \leq \frac{\sigma}{2}(1 - 2\eta\sigma)^j$, and hence $\lambda_l \geq (1 - \delta)\sigma$ for some*

$$l \;\leq\; T_1 + T_2, \qquad T_2 := \Bigl\lceil \frac{\log(1/2\delta)}{-\log(1 - 2\eta\sigma)} \Bigr\rceil. \tag{17}$$

Proof. By monotonicity $\lambda_l \geq \sigma/2$ for $l \geq T_1$, so (12) gives $u_{l+1} \leq u_l(1 - 4\eta\lambda_l) \leq u_l(1 - 2\eta\sigma)$, and $1 - 2\eta\sigma \in [\frac{1}{2}, 1)$ by Assumption 3. Since $u_{T_1} \leq \sigma/2$, induction gives the stated decay. Requiring $\frac{\sigma}{2}(1 - 2\eta\sigma)^j \leq \delta\sigma$ yields the bound on T_2. □

Proof of Theorem 1. Lower bound. Apply (13) at $T = T_\delta$. Since $u_l \le \sigma$ for all l, $\log(\lambda_{T_\delta}/\lambda_0) \le 2T_\delta \log(1+2\eta\sigma)$. By definition $\lambda_{T_\delta} \ge (1-\delta)\sigma$, so

$$T_\delta \;\ge\; \frac{\log\big((1-\delta)\sigma/\lambda_0\big)}{2\log(1+2\eta\sigma)} = L^* - C_1.$$

Upper bound. By Lemma 7, $T_\delta \le T_1 + T_2$. Apply the identity (15) at $T = T_1$, using $\lambda_{T_1} \le \sigma$ (Proposition 8(i)) and Lemma 6:

$$T_1 = \frac{\log(\lambda_{T_1}/\lambda_0) + 2D_{T_1}}{2\log(1+2\eta\sigma)} \;\le\; \frac{\log(\sigma/\lambda_0) + (1+\eta\sigma)^2}{2\log(1+2\eta\sigma)} = L^* + \frac{(1+\eta\sigma)^2}{2\log(1+2\eta\sigma)}.$$

Finally $T_2 \le \log\frac{1}{2\delta}\big/\big(-\log(1-2\eta\sigma)\big) + 1$ by removing the ceiling. Summing gives (4). □

H.4 Extension to the full stable range $2\eta\sigma < 1$

Assumption 3 ($4\eta\sigma \le 1$) is used above only to guarantee monotonicity. For the *first* hitting time T_δ, local stability suffices.

Theorem 2 (Theorem 1 on the full stable range)**.** *Let Assumption 2 hold and $0 < 2\eta\sigma < 1$, with $\lambda_0 \in (0, \sigma/2)$ and $\delta \in (0, 1/2]$. Then $L^* - C_1 \le T_\delta \le L^* + C_2'$, where*

$$C_2' = \frac{(1+\eta\sigma)^2 + \log^+\frac{(1+\eta\sigma)^2}{2}}{2\log(1+2\eta\sigma)} + \frac{\log\frac{1}{2\delta}}{-\log(1-2\eta\sigma)} + 1.$$

When $4\eta\sigma \le 1$ the $\log^+$ term vanishes and $C_2' = C_2$.

Proof. Three observations replace monotonicity.

(a) *Before the first hit.* For $l < T_\delta$ we have $\lambda_l \in (0, (1-\delta)\sigma)$ by definition, so $u_l \in (\delta\sigma, \sigma)$ and $\lambda_{l+1} = \lambda_l(1+2\eta u_l)^2 > \lambda_l$. Hence the sequence is increasing up to T_δ, $T_\delta < \infty$, and the telescoping identity (15) holds for every $T \le T_\delta$ with $D_T \ge 0$. The lower bound follows exactly as before.

(b) *Overshoot allowance.* For $\lambda \le \sigma/2$, $g'(\lambda) = (1+2\eta u)(1+2\eta\sigma-6\eta\lambda) \ge (1+2\eta u)(1-\eta\sigma) > 0$, so g is increasing on $[0, \sigma/2]$ and $\lambda_{T_1} \le g(\sigma/2) = \frac{\sigma}{2}(1+\eta\sigma)^2$. The drag bound of Lemma 6 does not use the step-size condition, and gives $D_{T_1} \le (1+\eta\sigma)^2/2 \le 9/8$. Therefore $T_1 \le L^* + [(1+\eta\sigma)^2 + \log^+((1+\eta\sigma)^2/2)]/(2\log(1+2\eta\sigma))$.

(c) *Saturation up to the first hit.* For $T_1 \le l < T_\delta$, (a) gives $\lambda_l \in [\sigma/2, (1-\delta)\sigma)$ and $u_l > 0$, so Lemma 4 gives $u_{l+1} \le u_l(1-2\eta\sigma)$, with $1 - 2\eta\sigma \in (0,1)$. If the bracket turns negative, then $\lambda_{l+1} > \sigma$ and the level has been hit. Lemma 7 therefore holds up to T_δ, and summing gives the upper bound. □

Overshoot past σ, which is possible when $4\eta\sigma > 1$, can occur only at or after the first hit, so it does not affect T_δ. The binding mode satisfies $4\eta\sigma_r \le 1$ in every configuration of the paper. The extension matters for the larger modes and for the η sweep of Appendix D.

H.5 Additive offset

The following consequence of Theorem 1 is used in Section 5.

Corollary 2 (Additive offset)**.** *For fixed (σ, η, δ) satisfying Assumption 3,*

$$T_\delta(\lambda_0) = \frac{\log(1/\lambda_0)}{2\log(1+2\eta\sigma)} + c(\sigma,\eta,\delta) + e(\lambda_0), \qquad |e(\lambda_0)| \le C_1 + C_2, \tag{18}$$

uniformly over $\lambda_0 \in (0, \sigma/2)$. Consequently,

$$\lim_{\lambda_0 \to 0} \frac{T_\delta(\lambda_0)}{\log(1/\lambda_0)} = \frac{1}{2\log(1+2\eta\sigma)}. \tag{19}$$

Under factor-scale initialization $\lambda_0 = \Theta(\varepsilon_0^2)$, the asymptotic slope in $\log(1/\varepsilon_0)$ is $1/\log(1+2\eta\sigma)$. In particular, a multiplicative model $T_\delta = cL^*$ with $c e 1$ is ruled out by the uniform additive bound.

H.6 NUMERICAL VERIFICATION

The script `verify_escape_time.py` (supplementary material) checks every step above against the exact recursion and the matrix dynamics. It covers the lemmas of Appendix H, the two-sided bound over the full stable range (block [T1]), Proposition 6 (block [P5]) and Corollary 1 (block [C4]). Its complete output:

```
[L1-L3] max |mode of BA - scalar recursion| : 3.109e-15
[L1-L3] max off-diagonal leakage            : 6.026e-16
[L1-L3] max |B^T B - A A^T|                 : 1.776e-15
[L4] max relative error, deficit identity    : 5.747e-11

[P4] regime scan (sigma = 1, lam0 = 1e-6, T = 4000)
     4*eta*sigma   monotone   overshoot  lam -> sigma
           0.500       True       False          True
           0.900       True       False          True
           0.990       True       False          True
           1.000       True       False          True
           1.001      False        True          True
           1.300      False        True          True
           1.900      False        True          True
           1.990      False        True          True
           2.000      False        True         False
           2.050      False        True         False
    g'(sigma) numeric 0.160000000   predicted 1-4*eta*sigma 0.160000000

[L5] max relative error, telescoping identity : 1.358e-15
[L6] drag at T1 = first l with lam_l >= sigma/2  (bound (1+eta s)^2/2)
     eta*sigma  log10 lam0     D_T1  2eta*sum lam    bound    ok
         0.001          -2   0.3411        0.3417   0.5010  True
         0.001          -6   0.3457        0.3463   0.5010  True
         0.001         -10   0.3458        0.3464   0.5010  True
         0.010          -2   0.3362        0.3420   0.5101  True
         0.010          -6   0.3431        0.3490   0.5101  True
         0.010         -10   0.3377        0.3435   0.5101  True
         0.050          -2   0.3121        0.3390   0.5513  True
         0.050          -6   0.3000        0.3261   0.5513  True
         0.050         -10   0.3313        0.3598   0.5513  True
         0.100          -2   0.3302        0.3862   0.6050  True
         0.100          -6   0.3085        0.3615   0.6050  True
         0.100         -10   0.2862        0.3359   0.6050  True
         0.250          -2   0.2984        0.4245   0.7812  True
         0.250          -6   0.2323        0.3344   0.7812  True
         0.250         -10   0.1786        0.2593   0.7812  True
         0.350          -2   0.1441        0.2366   0.9113  True
         0.350          -6   0.2013        0.3266   0.9113  True
         0.350         -10   0.2761        0.4407   0.9113  True
         0.450          -2   0.2921        0.5123   1.0513  True
         0.450          -6   0.2366        0.4214   1.0513  True
         0.450         -10   0.1907        0.3438   1.0513  True
         0.490          -2   0.3798        0.6752   1.1100  True
         0.490          -6   0.1429        0.2714   1.1100  True
         0.490         -10   0.2018        0.3770   1.1100  True
    all drag bounds hold : True

[L7] violations of u_j <= (sigma/2)(1-2 eta sigma)^j before crossing : 0
[T1] two-sided bound over (eta*sigma, delta, lam0), full stable range 2 eta s < 1
    160/160 cells satisfy the bound; tightest lower margin 0.28 steps, tightest upper margin 2.06 steps

[C2] offset T_delta - L*, last step interpolated (sigma = 3, delta = 0.1)
     eta*sigma |      1e-3     1e-5     1e-7     1e-9    1e-11    1e-13    1e-15     spread
         0.050 |    9.8272   9.8432   9.8492   9.8501   9.8461   9.8375   9.8279   2.28e-02
         0.100 |    4.3926   4.3635   4.4008   4.3828   4.3799   4.4018   4.3645   3.83e-02
         0.250 |    1.2372   1.2685   1.1195   1.2283   1.2714   1.1440   1.2180   1.52e-01
         0.400 |    0.6280   0.6263   0.6016   0.5332   0.4387   0.4661   0.4938   1.89e-01

[C2] measured slope of T_delta in log(1/lam_0) vs 1/(2 log(1+2 eta sigma))
     eta*sigma   measured   predicted    ratio
         0.050     5.2459      5.2460   1.0000
         0.100     2.7422      2.7424   0.9999
         0.250     1.2320      1.2332   0.9990
         0.400     0.8438      0.8506   0.9919
[P5] K=   25  max rel |mode - recursion(sigma_hat)| 2.413e-14   rel ||BA - T_r(W_ols)|| at l=300 1.129e-14
[P5] K=   40  max rel |mode - recursion(sigma_hat)| 3.480e-15   rel ||BA - T_r(W_ols)|| at l=300 5.063e-15
[P5] K=  140  max rel |mode - recursion(sigma_hat)| 4.345e-15   rel ||BA - T_r(W_ols)|| at l=300 2.894e-15
[P5] K= 1000  max rel |mode - recursion(sigma_hat)| 4.700e-15   rel ||BA - T_r(W_ols)|| at l=300 6.444e-15
[P5] worst over K : 2.413e-14

[C4] K-independent depth floor (preconditioned, aligned, eta=0.02, lam0=1e-8)
         K    L       risk  floor bound   ok
        30   20     64.719       64.716 True
        30   40     44.045        8.162 True
        30   60     18.673        0.000 True
        30   80     20.025        0.000 True
       140   20     66.198       66.198 True
       140   40     65.730       65.136 True
       140   60     14.842        0.000 True
       140   80      0.742        0.000 True
      1000   20     47.129       47.129 True
```

```
 1000   40     47.015       46.871 True
 1000   60     22.457        0.000 True
 1000   80      6.605        0.000 True
10000   20     59.508       59.508 True
10000   40     58.545       57.211 True
10000   60     21.186        0.000 True
10000   80      0.772        0.000 True
violations: 0/484
```

I SCOPE AND MODELLING QUALIFICATIONS

The comparison in the main text is intentionally restricted to a linear surrogate: linear attention, linear adaptation, linear ground truth, Gaussian design, and realizable tasks. We do not claim that real transformer ICL implements literal bilinear gradient descent with the small-initialization schedule assumed here; multi-head softmax attention is not linear attention. The idealization is chosen because it is the largest class in which the estimators considered here are closed-form or have provably matching dynamics, allowing statistical and optimization effects to be separated from architectural approximation error, and because this is the setting in which ICL's implicit optimization has been proved exact (Von Oswald et al., 2023; Ahn et al., 2023).

Three results are further scoped. Proposition 1 relies on factored descent reaching its global minimizer, which is proved for the preconditioned update and observed for the vanilla one (Appendix A.4); the structure-aware estimator of Section 7 is an oracle rather than a mechanism; and factored descent is realized exactly by register attention under Assumption 1 (Proposition 2), with hand-set weights that trained models need not find (Appendix J.6). Theorem 1 is proved under aligned balanced initialization. It covers the population and, for the preconditioned update, every finite K (Proposition 6), on the full stable range $2\eta\sigma < 1$ (Theorem 2). The case of vanilla descent from random initialization at finite K, with its alignment cost, remains outside the theorem; only the depth floor of Corollary 1 covers it. The constant $r(2d-r)$ is derived asymptotically (Proposition 5).

The practical prediction defended by the paper is qualitative and falsifiable: architectures with more effective optimization depth should show a sharp rather than gradual improvement in few-shot data efficiency as depth grows, and the depth of this transition should depend on how well pretraining aligns the model with the task family. Since misalignment imposes a floor on required depth that scale alone cannot remove, a model pretrained on a task distribution close to the downstream one should transition at markedly shallower effective depth than one pretrained on a distant distribution. This comparison can be tested on real transformers without importing the linear-algebraic assumptions of the analysis.

J REGISTER ATTENTION: WHAT ASSUMPTION 1 NEEDS AND WHAT IT BUYS

This appendix proves the attention-level results of Section 4. Tokens are $e_i \in \mathbb{R}^D$, the key set $\mathcal{K}$ contains the demonstrations and any register or anchor tokens, and query tokens are masked as keys. A multi-head linear-attention layer maps

$$e_j \leftarrow e_j + \textstyle\sum_h P_h \hat{G} Q_h e_j, \qquad \hat{G} = \frac{1}{K} \sum_{i \in \mathcal{K}} e_i e_i^\top, \tag{20}$$

with fixed value–output matrices P_h and merged key–query matrices Q_h. Because queries are masked, the final query token is $\prod_l (I + M_l) e_q$ with $M_l = \sum_h P_{l,h} \hat{G}_l Q_{l,h}$, which is linear in x_q; the implicit predictor $\hat{W}$ is read off with probe queries $x_q = e_m$.

J.1 HEAD BOTTLENECKS GIVE WEIGHT-FIXED SUBSPACES

Lemma 8 (Range confinement)**.** *For any masked multi-layer linear attention with fixed weights and readout R, $\hat{W} - W_{\mathrm{base}}$ has column space inside $R\big(\sum_{l,h} \mathrm{range}(P_{l,h})\big)$, where W_{base} is the context-free map. This subspace has dimension at most $\sum_{l,h} \mathrm{rank}(P_{l,h})$ and does not depend on the demonstrations.*

Proof. $e_q^{(l+1)} - e_q^{(l)} = M_l e_q^{(l)} \in \sum_h \mathrm{range}(P_{l,h})$ for every $\hat{G}_l$; sum over l and apply R. □

A single layer's context-induced update $P\hat{G}Q$ is thus a low-rank update with frozen outer factors and a data-dependent core, whereas LoRA trains both factors.

Proposition 9 (Fixed-subspace bias)**.** *Let U^* be Haar-distributed and A^* have i.i.d. $\mathcal{N}(0,\tau^2)$ entries. Any estimator, linear or not, whose range lies in a fixed subspace $\mathcal{U}$ of dimension $s < d$ has $\mathbb{E}\left\|\hat{W} - \Delta W^*\right\|_F^2 \geq \tau^2 r(d-s)$, at every sample size and depth.*

Proof. $\left\|\hat{W} - \Delta W^*\right\|_F^2 \geq \|(I - \Pi_{\mathcal{U}})U^* A^*\|_F^2$, whose expectation is $\tau^2 d\,\mathbb{E}\,\|(I-\Pi_{\mathcal{U}})U^*\|_F^2 = \tau^2 r(d-s)$. □

Proposition 10 (Gradient-descent-form attention)**.** *With tokens (x_i, y_i), queries $(x_q, 0)$ and weights $P_l = -2\eta\,\mathrm{diag}(0, M_l)$, $Q_l = \mathrm{diag}(N_l, 0)$, the implicit iterate satisfies exactly $W_{l+1} = W_l + 2\eta M_l(\hat{S} - W_l\hat{\Sigma})N_l$ from $W_0 = 0$. Hence W_L is linear in Y given X, and $\mathrm{range}(W_L) \subseteq \sum_l \mathrm{range}(M_l)$. For $M_l \equiv \Pi$, $N_l \equiv I$, $W_L = \Pi\widehat{W}_{\mathrm{ols}}[I - (I - 2\eta\hat{\Sigma})^L]$: geometric contraction from the first layer, with no escape plateau and no initialization scale.*

Proof. By induction the y-channel of demonstration j is $y_j - W_l x_j$ and that of the query is $-W_l x_q$. The layer adds $-2\eta M_l(\frac{1}{K}\sum_i r_i x_i^\top)N_l x_j = -2\eta M_l(\hat{S} - W_l\hat{\Sigma})N_l x_j$ to both. If $\Pi W_l = W_l$ the recursion is $W_{l+1} = W_l(I - 2\eta\hat{\Sigma}) + 2\eta\Pi\hat{S}$. □

J.2 Proof of Proposition 3

$\mathbb{E}[\Delta W^*_{ab}\Delta W^*_{ce}] = \tau^2\delta_{be}\mathbb{E}[(U^*U^{*\top})_{ac}] = (\tau^2 r/d)\delta_{ac}\delta_{be}$ for Haar U^*, so the rows w_a of ΔW^* are uncorrelated with covariance $(\tau^2 r/d)I$, and the columns of Y are $Xw_a + \varepsilon_a$. The linear minimum-mean-square-error estimator decouples across a and is ridge with $\lambda = \sigma^2 d/(\tau^2 r)$, with error covariance $\sigma^2(X^\top X + \lambda I)^{-1}$ per row, giving R_{lin}. The bound uses only second moments, so it holds for every estimator linear in Y given X, in particular for Proposition 10 at any depth and for estimator (3). Since $R_{\mathrm{lin}}(K) \approx \sigma^2 d^2/(K+\lambda)$ and $R_r(m) \approx \sigma^2 r(2d-r)/m$ (Proposition 5), $K_{\mathrm{lin}}(m)/m \to d^2/[r(2d-r)]$. At $(20, 3, 140)$ a Monte Carlo search gives $K_{\mathrm{lin}} = 467$, i.e. $3.34m$, against the asymptotic 3.60; at $(10, 2, 70)$, $K_{\mathrm{lin}} = 176$, i.e. $2.51m$. ■

J.3 Proof of Proposition 2

Use channels $[x|y|p|a|b|z^k|z^v]$, $D = 7d$, with S_c embedding channel c. Demonstrations are $(x_i, y_i, p_i, 0, \dots)$; register k carries $a = a_k$ (row k of A_0) and $b = b_k$ (column k of B_0); anchor m carries $z^k = z^v = e_m$. The two layers are

$$\begin{aligned}\text{refresh:}\quad & P = \{K\,S_p S_b^\top,\ -K\,S_p S_{z^v}^\top\}, \quad Q = \{S_a S_x^\top,\ S_{z^k} S_p^\top\};\\ \text{update:}\quad & P = \{2\eta\,S_a S_x^\top,\ 2\eta\,S_b(S_y - S_p)^\top\}, \quad Q = \{(S_y - S_p)S_b^\top,\ S_x S_a^\top\}.\end{aligned}$$

In refresh, head one contributes $\sum_k b_k(a_k^\top x_j) = BAx_j$ and head two contributes $-\sum_m e_m e_m^\top p_j = -p_j$, so $p_j \leftarrow BAx_j$ on demonstrations and queries, while registers and anchors ($x = p = 0$) are unchanged. In update, head one adds $\frac{2\eta}{K}\sum_i x_i r_i^\top b_k$ to a_k, i.e. $A \leftarrow A + 2\eta B^\top C = A - \eta B^\top G$ with $C = \hat{S} - BA\hat{\Sigma}$ and $r_i = y_i - p_i$; head two adds $2\eta C a_k$ to b_k, i.e. $B \leftarrow B - \eta G A^\top$. Both use pre-update values, so the step is simultaneous. Demonstrations and queries have $a = b = 0$, and registers and anchors have $x = y - p = z = 0$ in every slot the other heads read, so all cross terms vanish. Mode strengths follow from $BA = \sum_k b_k a_k^\top$. ■

The anchors exist only to reset p in pure linear attention; a residual MLP or a constant channel removes them. The block is weight-tied, so the mechanism is a looped transformer, and Theorem 1 with its corollaries transfers verbatim with depth $2L + 1$ layers.

J.4 Label nonlinearity without registers

Proposition 11 (Register-free multiplicative dynamics)**.** *With channels $[x|y|p]$ and heads $P = \{2\eta\alpha S_p(S_y - S_p)^\top, 2\eta\beta S_p(S_y - S_p)^\top\}$, $Q = \{S_x S_x^\top, S_p S_p^\top\}$, one layer implements exactly*

$W_{l+1} = W_l - \eta G_l(\alpha I + \beta W_l^\top W_l)$. *The iterate is nonlinear in Y, and for $\alpha > 0$ its fixed point is $\widehat{W}_{\mathrm{ols}}$.*

Proof. Head two contributes $\frac{2\eta\beta}{K}\sum_i r_i p_i^\top p_j = 2\eta\beta C W^\top W x_j$ because $p_i = W x_i$; head one is the gradient term. Both update p in place. For $\alpha > 0$, $\alpha I + \beta W^\top W \succ 0$, so a fixed point needs $G = 0$. □

Label nonlinearity is therefore necessary for the rank-r rate (Proposition 3) but not by itself sufficient; registers supply the rank-r state that makes the endpoint $\widehat{W}_r$.

Lemma 9 (The LoRA side)**.** *(i) LoRA with full-batch GD under squared loss on a linear map is exactly factored GD; the standard initialization $B_0 = 0$ is unbalanced, $B^\top B - AA^\top = -A_0 A_0^\top$, which is consistent with the faster third arm of Appendix E. (ii) LoRA applied to the weights of gradient-descent-form attention leaves it in that form, so Propositions 10 and 3 still apply; fine-tuning can move the fixed subspace toward a family subspace but cannot make it task-adaptive. (iii) Fine-tuning the registers of Proposition 2 is prompt tuning and changes exactly (B_0, A_0), i.e. λ_0 and the alignment.*

J.5 Numerical verification

`assumption1_gpu.py verify` runs each construction as an actual attention forward pass in float64 (RTX A6000; $d = 20$, $r = 3$, $K = 140$, $L = 80$).

check	max relative error	nonlinearity in Y
register attention vs. factored GD (Prop. 2; W, B, A; 161 layers)	7.7×10^{-16}	0.354
gradient-descent form, fixed rank-r subspace (Prop. 10)	8.2×10^{-16}	6.4×10^{-16}
register-free multiplicative (Prop. 11)	9.7×10^{-16}	0.026

Leakage of the gradient-descent-form estimate outside its fixed subspace is 7.4×10^{-16}. For a random three-layer, two-head stack with rank-one value matrices, the implicit updates span a numerical rank of 6, equal to the bound $\sum \mathrm{rank} P$ of Lemma 8, in ambient dimension 40. Over 4000 tasks the ridge risk is 3.152 ± 0.004 against $R_{\mathrm{lin}} = 3.155$, OLS is 3.355 against 3.361, and RRR is 0.886. In Table 3 the fixed-subspace plateau 50.5 sits above $\tau^2 r(d - r) = 51$ rescaled to the sampled chance risk (58.9 against its expectation 60), i.e. 50.1.

J.6 Trained linear transformers

`assumption1_gpu.py train`: $d = 10$, $r = 2$, $K = 70$, $\sigma = \tau = 1$, token width $7d$, four heads per layer, Adam at 10^{-3} with warmup and cosine decay, 15,000 steps of batch 128, 32 query tokens, two seeds, evaluation on 2048 held-out tasks through probe queries. *full*: unconstrained heads; *headrank*: value and key–query matrices of rank r; *reg*: full heads plus $r + d$ learned register tokens. Nonlinearity is $\|\hat{W}(Y_1 + Y_2) - \hat{W}(Y_1) - \hat{W}(Y_2)\|_F / \|\hat{W}(Y_1 + Y_2)\|_F$; alignment is $\|U^{*\top}\hat{U}_r\|_F^2 / r$. Baselines: OLS 1.695, RRR 0.589, $R_{\mathrm{lin}} = 1.541$, oracle subspace 0.337.

model	L	risk	risk/RRR	risk/R_{lin}	nonlinearity	top-r energy	alignment
full	1	3.723	6.32	2.42	0.002	0.960	0.981
full	2	2.354	4.00	1.53	0.186	0.990	0.980
full	4	0.589	1.00	0.38	0.416	0.999	0.983
full	6	0.573	0.97	0.37	0.706	1.000	0.983
headrank	1	17.45	29.6	11.3	0.009	0.888	0.331
headrank	2	14.40	24.4	9.35	0.323	0.919	0.452
headrank	6	4.863	8.25	3.16	0.495	0.973	0.905
headrank	8	2.704	4.59	1.76	0.110	0.943	0.976
headrank	12	1.823	3.09	1.18	0.110	0.951	0.978
reg	1	3.723	6.32	2.42	0.002	0.960	0.981
reg	2	2.230	3.79	1.45	0.163	0.990	0.981
reg	4	0.588	1.00	0.38	0.493	0.999	0.983
reg	6	0.572	0.97	0.37	0.555	1.000	0.983

headrank at $L = 4$ failed to train (risk 445) and *full* and *reg* at $L \in \{8, 12\}$ diverged at every learning rate tried (10^{-3} and two reductions by 3); these cells are reported as missing rather than tuned. At $L \geq 4$ the trained full-head models are below the RRR risk; RRR is the rank-constrained ERM, not the Bayes estimator, so this does not contradict Proposition 1.

K CODE

The full source is provided as supplementary material, with a single `run_all.sh` entry point that runs every script in dependency order and writes one log per script. All scripts use fixed integer seeds.

script	produces
`common_v2.py`, `exchange_rate_verify.py`	shared harness; the four estimators; Table 1 (left)
`verify_escape_time.py`	Appendix H.6
`e1_revised.py`	Figure 2c, Appendix C
`e2_rrr_extrapolation.py`	Table 1 (right), Appendix A.4
`e3_generality.py`	Appendix D
`e4_warmstart_family.py`	Figure 2b, Figure 3b,c
`e5_task_structure.py`	Figure 3a
`e6_equal_budget.py`	Figure 2a, Appendix E
`e7_permode.py`	Table 2
`e8_precond.py`	preconditioning check (Section 6)
`plot_final.py`	Figures 2 and 3
`assumption1_gpu.py`	Table 3, Appendix J (GPU)
`verify_assumption1_numpy.py`	float64 CPU reference for Appendix J