IN-CONTEXT CLASSIFICATION: CALIBRATION AND NOISE IN LINEAR ATTENTION WITH SOFTMAX READOUT

Anonymous authors
Paper under double-blind review

ABSTRACT

We study in-context binary classification in a simple attention model trained to minimise a logistic loss. Each prompt has a latent logistic teacher, and the learner must infer its direction from Bernoulli-labeled context examples. We derive an asymptotic prediction for the test logit in a high-dimensional regime with context length $\ell \propto d$ and total training prompts $n \propto d^2$. Our formula cleanly separates calibration, test-time context noise, and non-isotropic parameter-estimation noise as interpretable effects on in-context performance. Our analysis also identifies the ridgeless separability regime and addresses the effects of context length, pretraining-set size, regularisation, and sampling temperature on in-context classification performance. The theory's predictions closely agree with finite-dimensional simulations and fully trained linear-attention models, while experiments with standard softmax attention show qualitatively related behaviour beyond the regime assumed by our analysis.

1 INTRODUCTION

In-context learning (ICL) is the ability of a model to adapt its predictions from examples in a prompt, without updating its parameters (Brown et al., 2020; Von Oswald et al., 2023; Wei et al., 2022). It is a striking capability of modern sequence models, which can, for instance, implement a range of statistical algorithms in-context (Bai et al., 2023). However, the statistical mechanism that turns a context into a new-task prediction is not fully understood.

Simplified ICL models offer a useful route to this question because they make explicit which prompt statistics an architecture can use (Akyürek et al., 2023; Von Oswald et al., 2023; Zhang et al., 2023; Lu et al., 2025; Oko et al., 2024; Vasudeva et al., 2025; Letey et al., 2026b;a). In particular, a significant body of recent work has explored ICL in the context of the test problem of linear regression. For these tasks, linear attention (Katharopoulos et al., 2020) has been shown to be a good surrogate for understanding how and what transformers learn (Ahn et al., 2024; Zhang et al., 2024; Wu et al., 2024; Lu et al., 2025; Letey et al., 2026b;a; Bordelon et al., 2025). Together, these works demonstrate how reduced models can yield substantial insight into the mechanisms of ICL.

Despite extensive progress in the theory of in-context learning, many works focus on regression; binary classification introduces qualitatively different questions and phenomena. In the setting studied here, each prompt is generated by a latent task direction, and its context labels are Bernoulli observations determined by that direction. A successful predictor must therefore output a calibrated probability rather than merely recover a sign. Logistic regression, a canonical high-dimensional inference problem, is a natural target for such investigation. This problem has been widely studied (Mai et al., 2019; Sur & Candès, 2019), including its phase transitions (Candes & Sur, 2018; Barbier et al., 2019), universality (Pesce et al., 2023), and generalisation (Gerace et al., 2020; Mei & Montanari, 2022), leading to a rich understanding of a single instance of this estimation problem.

By contrast, the in-context setting involves learning a general estimator that can approximately solve new instances of logistic regression. Recent works have investigated whether transformers can learn to do this, perhaps by performing normalised gradient descent (Zhang & Cao, 2026). Moreover, studies of multiclass classification have shown that properties of the data greatly affect the model's ability to learn in-context (Chan et al., 2022). Yet, a detailed understanding of in-context classification, comparable to what has been achieved for regression, is still lacking.

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.

Here, we study in-context logistic regression using a reduced linear-attention model with a softmax readout. We derive a high-dimensional characterisation of the output logit, and corresponding test errors, in a proportional regime where context length scales with token dimension and the number of pretraining sequences scales with its square. We characterise the joint law of the output logit and teacher logit, leading to the following findings.

- The form of the output logit is highly interpretable, and decomposes into three parts: scaling of the teacher-aligned signal, noise from test-time context, and anisotropy induced by finitely many pretraining sequences.
- This separation distinguishes the roles of context length and training-set size, and recovers the Bayes logit in the joint large-context, large-pretraining-data ridgeless limit.
- We additionally predict the separability boundary at which the ridgeless logistic estimator ceases to have a finite solution.
- Finally, we verify the theory through finite-dimensional simulations and show that its predicted logit distribution describes the output of a trained linear-attention model.

2 Data and model setup

Data generation. Following prior theoretical work on ICL (Zhang et al., 2024; Wu et al., 2024; Lu et al., 2025), we train the model on a dataset of n sequences of examples. Each sequence consists of (x, y) pairs sampled from a latent task, and the model aims to learn the corresponding conditional distribution. Each sequence has a latent task vector $\boldsymbol{w} \in \mathbb{R}^d$ drawn uniformly from the unit sphere. Fix an inverse temperature $\beta > 0$. Fixing $\boldsymbol{w}$, draw independent covariates $\boldsymbol{x}_i \sim \mathcal{N}(0, I_d)$ and a query $\boldsymbol{x}_{\mathrm{q}} \sim \mathcal{N}(0, I_d)$, with

$$y_i \mid \boldsymbol{x}_i, \boldsymbol{w} \sim \mathrm{Bernoulli}\big(\sigma(\beta \boldsymbol{w}^\top \boldsymbol{x}_i)\big)\,, \qquad y_q \mid \boldsymbol{x}_{\mathrm{q}}, \boldsymbol{w} \sim \mathrm{Bernoulli}\big(\sigma(\beta \boldsymbol{w}^\top \boldsymbol{x}_{\mathrm{q}})\big)\,, \tag{1}$$

where $\sigma(t) = (1 + e^{-t})^{-1}$ is the sigmoid function.

Model and estimator. Following prior work, we embed each sequence as a matrix

$$Z = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_\ell & \boldsymbol{x}_{\mathrm{q}} \\ y_1 & \cdots & y_\ell & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(\ell+1)}.$$

The final zero is a placeholder for the query label y_{q}. We pass Z to a linear-attention operator and map the final, label-coordinate entry of the updated query column to a binary class probability. Specifically, writing $\boldsymbol{z}_i = [\boldsymbol{x}_i^\top \; y_i]^\top$ and $\boldsymbol{z}_{\mathrm{q}} = [\boldsymbol{x}_{\mathrm{q}}^\top \; 0]^\top$, the query column updated by linear attention is

$$\mathrm{LA}_{Q,K,V}(Z)_{:,\ell+1} = \boldsymbol{z}_{\mathrm{q}} + \frac{1}{\ell}\sum_{i=1}^{\ell} \langle Q\boldsymbol{z}_{\mathrm{q}}, K\boldsymbol{z}_i\rangle V\boldsymbol{z}_i. \tag{2}$$

Here $Q, K, V \in \mathbb{R}^{(d+1)\times(d+1)}$ are the query, key, and value matrices, respectively, that parameterise linear attention. Since this is binary classification, the softmax depends only on the difference between the output logits. We use the final, label entry as the class-1 logit and fix the class-0 logit to zero,

$$r_{Q,K,V}(Z) = \big[\mathrm{LA}_{Q,K,V}(Z)\big]_{d+1,\ell+1}, \tag{3}$$

$$\mathrm{Out}(Z) = \big[\mathrm{softmax}(0, r_{Q,K,V}(Z))\big]_2 = \sigma(r_{Q,K,V}(Z)). \tag{4}$$

Prior work on in-context linear regression has shown that linear attention admits a particularly clean reduced representation: by expanding (4) in terms of V and $Q^\top K$, not all components of these matrices contribute to learnable directions, and can be ignored without meaningfully changing the optimisation dynamics (Zhang et al., 2024; Lu et al., 2025). We utilise this established reduction to simplify the output logit,

$$r_{Q,K,V}(Z) \approx \boldsymbol{x}_{\mathrm{q}}^\top \left(A\frac{d}{\ell}\sum_{i=1}^{\ell} y_i\boldsymbol{x}_i + \boldsymbol{a}\frac{1}{\ell}\sum_{i=1}^{\ell} y_i \right) \tag{5}$$

$$A = \frac{1}{d} V_{d+1,d+1}(Q^\top K)_{1:d,1:d}\,, \quad \boldsymbol{a}^\top = V_{d+1,d+1}(Q^\top K)^\top_{1:d,d+1}\,. \tag{6}$$

We will later verify this by directly comparing our results for this reduced model to a fully parameterised, trained linear-attention model. We can write the reduced model more compactly by introducing the new parameter matrix $\Gamma = [A \;\; \boldsymbol{a}] \in \mathbb{R}^{d\times(d+1)}$ and

$$r_\Gamma(Z) = \langle \Gamma, H\rangle, \quad H = \boldsymbol{x}_{\mathrm{q}} \begin{bmatrix} \boldsymbol{v}^\top & c \end{bmatrix}, \quad \boldsymbol{v} = \frac{d}{\ell}\sum_{i=1}^{\ell} y_i \boldsymbol{x}_i, \quad c = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i. \tag{7}$$

Finally, applying softmax gives the class-probability estimator $\mathrm{Out}_\Gamma(Z) = \sigma(r_\Gamma(Z))$.

Optimal parameters given training batch. We consider a training batch with n independent prompts, which yields n independent feature matrices H_μ ($\mu \in [n]$) in the reduced logit r_Γ. Write $\boldsymbol{\gamma} = \mathrm{vec}(\Gamma)$ and $\boldsymbol{h}_\mu = \mathrm{vec}(H_\mu)$ for their row-wise vectorisations in $\mathbb{R}^{d(d+1)}$. We use y_μ as shorthand for the μ^{th} sequence's query label y_{q}, and fit the reduced model by regularised Bernoulli negative log-likelihood:

$$\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}} \left\{ \frac{1}{n}\sum_{\mu=1}^{n} \left[L(\boldsymbol{\gamma}^\top \boldsymbol{h}_\mu) - y_\mu \boldsymbol{\gamma}^\top \boldsymbol{h}_\mu \right] + \frac{\lambda d}{2}\|\boldsymbol{\gamma}\|^2 \right\}, \quad L(t) = \log\bigl(1 + e^t\bigr). \tag{8}$$

The choice of λd scaling ensures that the regularisation is applied evenly and consistently across dimensions, as measured by comparison to the empirical Hessian trace: the average Hessian eigenvalue grows with d and so without this normalisation, the effective regularisation would be 0 at large d.

Asymptotic assumptions. We study the proportional limit $n, \ell, d \to \infty$ with constant

$$\tau := \frac{n}{d^2}, \qquad \alpha_{\mathrm{train}} := \frac{\ell}{d}.$$

3 High-dimensional characterisation

For the fitted, data-dependent parameters $\widehat{\Gamma} = [A^* \;\; \boldsymbol{a}^*]$, define the empirical order parameters

$$\widehat{a}_1 = \mathrm{tr}(A^*), \qquad \widehat{a}_2 = d\|A^*\|_F^2, \qquad \widehat{a}_0 = \|\boldsymbol{a}^*\|^2. \tag{9}$$

We will derive limiting deterministic values for these random parameters. The parameter $\widehat{a}_0$ measures the strength of the vector component of $\widehat{\boldsymbol{\gamma}}$. For the other two, note that

$$\widehat{a}_2 - \widehat{a}_1^2 = d\|A^*\|_F^2 - \bigl(\mathrm{tr}\, A^*\bigr)^2 = d\left\| A^* - \frac{\mathrm{tr}(A^*)}{d} I_d \right\|_F^2 \geq 0.$$

This quantity is minimised at $A^* \propto I_d$, so it measures the squared distance of A^* from isotropy.

We aim to derive a statistically equivalent expression for the test logit, where $\boldsymbol{h}_{\mathrm{test}}$ is drawn fresh from the distribution described above, and the optimal parameters $\widehat{\boldsymbol{\gamma}}$ are set from a finite training set of sequences as above. We highlight that this will provide richer phenomenology than an analysis of the population optimal parameters, which we begin with to provide intuition.

3.1 Population limits

Consider $\Gamma_{\mathrm{pop}} = [A^*_{\mathrm{pop}} \;\; \boldsymbol{a}^*_{\mathrm{pop}}]$ and $\boldsymbol{\gamma}_{\mathrm{pop}} = \mathrm{vec}(\Gamma_{\mathrm{pop}})$ defined by

$$\boldsymbol{\gamma}_{\mathrm{pop}} = \arg\min_{\boldsymbol{\gamma}} \left\{ \mathbb{E}_{\boldsymbol{h}, y_{\mathrm{q}}} \left[L(\boldsymbol{\gamma}^\top \boldsymbol{h}) - y_{\mathrm{q}} \boldsymbol{\gamma}^\top \boldsymbol{h} \right] + \frac{\lambda d}{2}\|\boldsymbol{\gamma}\|^2 \right\} = \arg\min_{\boldsymbol{\gamma}} \mathcal{L}_{\mathrm{pop}}(\boldsymbol{\gamma}). \tag{10}$$

For $\lambda > 0$, this objective is strictly convex in $\boldsymbol{\gamma}$ and therefore has a unique minimiser. The linear features $\boldsymbol{w}^\top \boldsymbol{x}$ are rotationally invariant, since $(O\boldsymbol{w})^\top (O\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ for every orthogonal O. The structure of the linear-attention estimator then gives

$$\mathcal{L}_{\mathrm{pop}}(\mathrm{vec}([A \;\; \boldsymbol{a}])) = \mathcal{L}_{\mathrm{pop}}(\mathrm{vec}([OAO^\top \;\; O\boldsymbol{a}]))$$

for *every* orthogonal O. Uniqueness of the population solution therefore implies $A^*_{\mathrm{pop}} \propto I_d$ and $\boldsymbol{a}^*_{\mathrm{pop}} = \boldsymbol{0}$. Thus isotropy emerges as $\tau \to \infty$, whereas the finite-sample result can retain an anisotropic contribution.

We can separately consider the long-context limit. Holding n and d fixed, take $\ell \to \infty$. Then

$$\frac{1}{d}\boldsymbol{v} = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i \boldsymbol{x}_i \longrightarrow \mathcal{C}_\beta \boldsymbol{w}, \qquad c = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i \longrightarrow \frac{1}{2}, \tag{11}$$

where $\mathcal{C}_\beta := \mathbb{E}_{G\sim\mathcal{N}(0,1)}[G\sigma(\beta G)] = \beta\mathbb{E}[\sigma'(\beta G)]$ is the first Hermite coefficient of the sigmoid activation. Thus the reduced-model logit becomes

$$r_\Gamma(Z) \longrightarrow d\mathcal{C}_\beta \boldsymbol{x}_{\mathrm{q}}^\top A \boldsymbol{w} + \frac{1}{2}\boldsymbol{x}_{\mathrm{q}}^\top \boldsymbol{a}. \tag{12}$$

This predictor averages away within-prompt randomness, but A can still retain anisotropy from the finite collection of training sequences. Only after subsequently taking $\tau \to \infty$ is A isotropic. For the ridgeless ($\lambda = 0$) case, in that limit, calibration requires $a_1 = \beta/\mathcal{C}_\beta$, which yields the Bayes logit βG_{q} for $G_{\mathrm{q}} \sim \mathcal{N}(0,1)$. These two limits anticipate the separate effects present in the finite-sample result.

3.2 Finite-sample results: order parameters

We derive an equivalent distribution for the test logit, using fitted parameters $\widehat{\boldsymbol{\gamma}} = \mathrm{vec}(\widehat{\Gamma})$. The full derivation and its assumptions are given in Appendix A. Schematically, we use a cavity approach (El Karoui, 2018): we exclude one of the n prompts from the loss and study the effect of this omission on the fitted parameters. In particular, we study the deterministic limits a_1, a_2, and a_0 of the empirical order parameters $(\widehat{a}_1, \widehat{a}_2, \widehat{a}_0)$, which characterise properties of the optimised weights $\widehat{\boldsymbol{\gamma}}$. Define

$$\mathcal{C}_\beta = \mathbb{E}_{g\sim\mathcal{N}(0,1)}[g\sigma(\beta g)], \quad \Delta = \mathcal{C}_\beta^2 + \frac{1}{2\alpha_{\mathrm{train}}}.$$

Result 1: Self-consistency equations for order parameters

The order parameters are determined implicitly as follows. Fix $\alpha_{\mathrm{train}}, \tau, \beta$ and λ. For any candidate pair (a_1, a_2) such that $a_2 \geq a_1^2$, draw independent $G_{\mathrm{q}}, G_{\mathrm{tr}}, G_\perp \sim \mathcal{N}(0,1)$ and then draw from the label distribution $Y \mid G_{\mathrm{q}} \sim \mathrm{Bernoulli}(\sigma(\beta G_{\mathrm{q}}))$. Use the candidate order parameters to form the trained logit R_s as the unique solution of

$$R_s + s\big(\sigma(R_s) - Y\big) = a_1\mathcal{C}_\beta G_{\mathrm{q}} + \frac{a_1}{\sqrt{2\alpha_{\mathrm{train}}}} G_{\mathrm{tr}} + \sqrt{(a_2 - a_1^2)\Delta}\, G_\perp. \tag{13}$$

Define mean response $\chi(s)$ and susceptibility s jointly by

$$\chi(s) = \mathbb{E}\left[\frac{1}{1 + s\sigma'(R_s)}\right], \quad \chi(s) - \frac{\lambda}{\Delta}s = 1 - \frac{1}{\tau}. \tag{14}$$

Then for positive s, the scalars a_1 and a_2 are desired solutions when

$$a_1 = \frac{\mathbb{E}\big[R_s\left(\mathcal{C}_\beta G_{\mathrm{q}} + G_{\mathrm{tr}}/\sqrt{2\alpha_{\mathrm{train}}}\right)\big]}{\Delta + s\lambda}, \quad a_2 = \frac{\mathbb{E}[R_s^2] - \Delta(1-\chi)a_1^2 - s\lambda a_1^2}{\Delta\chi + s\lambda}. \tag{15}$$

These self-consistently determined scalars characterise the behaviour of the test logit $R = r_{\widehat{\Gamma}}(Z_{\mathrm{new}})$ of a fresh prompt Z_{new}, as given by the following main result. Before proceeding, we need to be careful when discussing the fixed point solutions a_1 and a_2 for a given data configuration determined by $\alpha_{\mathrm{train}}, \tau, \beta$ and λ, as it is not clear if such *finite* solutions exist. This is a well-known problem in logistic regression in the absence of regularisation (Candes & Sur, 2018).

Simply stated, if n data points in d dimensions have binary labels such that the two classes are linearly separable (*i.e.*, by a hyperplane), then there is no finite maximum likelihood estimator: the likelihood can be increased by increasing the norm of the classifier weights without bound. If we assign independent random binary labels 0 and 1 with probability $1/2$ to n d-dimensional points in general position, separability holds with probability tending to 1 as $n, d \to \infty$ if $n/d < 2$ (Cover, 1965; Gardner, 1988), and thus MLE exists with high probability if $n/d > 2$. However, logistic regression makes it more difficult to determine the separability threshold as now the labels are no

longer uniform, but themselves depend on the data through $\boldsymbol{w}^\top \boldsymbol{x}$. Candes & Sur (2018) and later Koriyama & Bellec (2025) show that there is a relationship between the effective signal strength and the transition value of n/d. The zero-signal case corresponds to the Cover transition above where the critical value of n/d is 2.

Result 2: Ridgeless solution existence boundary

A finite-norm optimiser $\widehat{\Gamma}$ exists with high probability for $\tau > \tau_c$, where τ_c is defined by

$$\tau_c^{-1} = \inf_{t \in \mathbb{R}} \mathbb{E}\big[\max(0, G + tU)^2(1 - \sigma(\beta G_{\mathrm{q}})) + \min(0, G + tU)^2 \sigma(\beta G_{\mathrm{q}})\big] \tag{16}$$

for $G, G_{\mathrm{q}}, G_{\mathrm{tr}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $U = (\mathcal{C}_\beta G_{\mathrm{q}} + G_{\mathrm{tr}}/\sqrt{2\alpha_{\mathrm{train}}})/\sqrt{\Delta}$.

We expect a similar phenomenon to occur in the ridgeless case of our theory; we have formalised this intuition in Result 2; see Appendix B for an informal argument. Here the correct order parameter instead is $n/(d(d+1)) \approx n/d^2 = \tau$, as our estimator takes n samples (n sequences) of features $\boldsymbol{h}$ in $d(d+1)$ dimensions. The zero-signal case would correspond to each of the n query labels y_{q} being chosen uniformly at random, and then the MLE should exist for $\tau > 2$. Increasing β, or increasing α_{train} so that the context more accurately represents its latent task, should increase the transition value away from 2. Indeed, empirically, we find a very good match between the location of empirical error blowup and the above predicted τ_c, as seen by Figure 2. Note that τ_c depends on both β and α_{train}, as will be discussed below (see Figure 4a).

Finally, we may use the self-consistency equations in Result 1 to characterise $\widehat{\boldsymbol{\gamma}}$, and to determine the ICL test error. To allow for different context length at testing time, we introduce $\alpha_{\mathrm{test}} = \ell_{\mathrm{test}}/d$.

Result 3: Characterisation of logit distribution

Fix $\alpha_{\mathrm{train}}, \tau, \beta$, and λ. Sample Gaussians $G_{\mathrm{q}}, G_{\mathrm{var}} \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$. Then, assuming the empirical order parameters of the fitted $\widehat{\Gamma}$ concentrate to finite values, their limiting behaviour is

$$(\widehat{a}_1, \widehat{a}_2, \widehat{a}_0) \approx (a_1, a_2, 0). \tag{17}$$

Here a_1 and a_2 are the deterministic self-consistent solutions given by Result 1. Further, the test logit $R = r_{\widehat{\Gamma}}(Z_{\mathrm{new}})$ has marginal distribution given by

$$\text{true query target } y_{\mathrm{q}} \sim \text{Bernoulli}(\sigma(\beta G_{\mathrm{q}})) \tag{18}$$

$$\text{predicted query logit } R = a_1 \mathcal{C}_\beta G_{\mathrm{q}} + G_{\mathrm{var}} \sqrt{\frac{a_2}{2\alpha_{\mathrm{test}}} + (a_2 - a_1^2)\mathcal{C}_\beta^2}. \tag{19}$$

We provide empirical verification of this result in Figure 1 for a range of data parameters. Note that both the theoretical and empirical distributions tend towards the true logit distribution as α_{train} and τ increase, as expected from the population analysis.

This characterisation of the joint distribution of the latent and predicted logits makes it easy to compute the limiting cross-entropy loss and classification error, verified in Figures 2 and 3.

Result 4: Error formulas

The cross entropy test loss is given by

$$\mathcal{E}_{\mathrm{CE}} = \mathbb{E}_{G_{\mathrm{q}}, G_{\mathrm{var}}}\big[L(R) - \sigma(\beta G_{\mathrm{q}}) R\big] \tag{20}$$

and the classification error is given by

$$\mathcal{E}_{\mathrm{class}} = \mathbb{P}(\mathbf{1}\{R \geq 0\} \neq Y) = \mathbb{E}_{G_{\mathrm{q}}}[\sigma(\beta G_{\mathrm{q}})(1 - \phi) + (1 - \sigma(\beta G_{\mathrm{q}}))\phi] \tag{21}$$

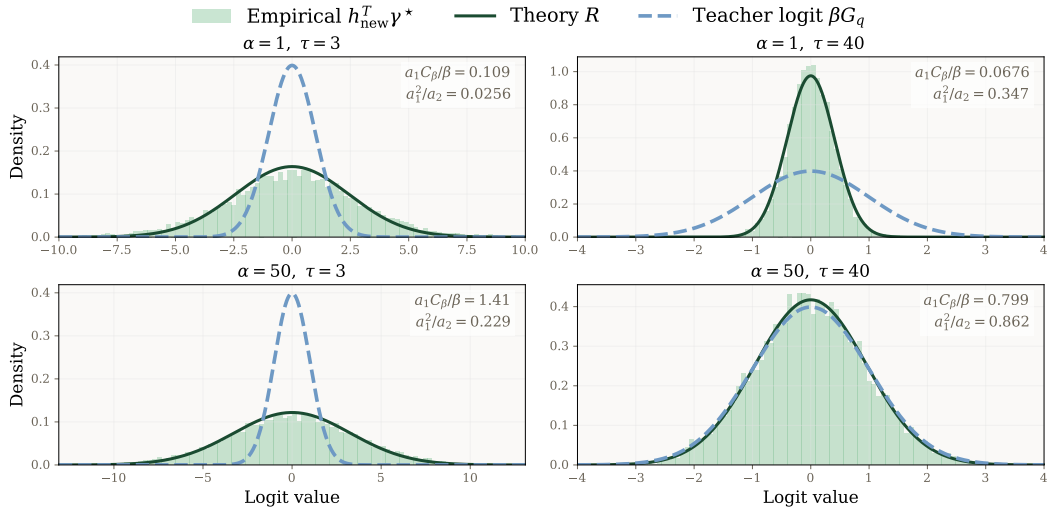
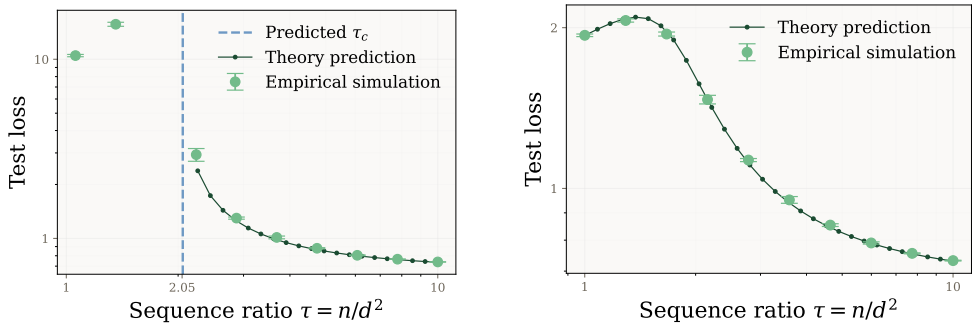


Figure 1: The distribution of empirical $\boldsymbol{h}_{\text{test}}^\top \boldsymbol{\gamma}^*$ matches our derived distribution R. The ridgeless theory predicts recovery of the true logit distribution as $\alpha_{\text{train}}, \tau \to \infty$. Increasing τ (right-hand plots) reduces variance, but alignment with G_{true} requires both variance reduction *and* $a_1 \mathcal{C}_\beta / \beta \to 1$. In all plots, $\beta = 1, \lambda = 10^{-6}, d = 32, \alpha_{\text{test}} = \alpha_{\text{train}}$.

where Φ is the standard normal CDF and

$$\phi = \Phi\left(\frac{a_1 \mathcal{C}_\beta}{\sqrt{V}} G_{\text{q}}\right), \quad V = \frac{a_2}{2\alpha_{\text{test}}} + (a_2 - a_1^2)\mathcal{C}_\beta^2.$$



(a) Ridgeless curve ($\lambda = 0$)

(b) Non-zero ridge curve ($\lambda = 0.001$)

Figure 2: Empirical simulation refers to directly sampling n sequences of length ℓ, computing $\widehat{\boldsymbol{\gamma}}$, and evaluating this parameter matrix on freshly sampled sequences. This is compared to the theoretical prediction in Result 4. The critical ratio is given by eq. (16). $d = 64$, $\alpha_{\text{train}} = 4 = \alpha_{\text{test}}$, $\beta = 1$.

4 Findings

We highlight several interpretable implications of these results.

Effect of test-time context length First note that α_{test} only appears explicitly in V; a_1 and a_2 do not depend on α_{test}. As α_{test} increases, this contribution to non-signal variance will vanish and ICL error will strictly decrease, both as measured by cross-entropy loss or by classification error (see Appendix C for proof). This is an artifact of our choice of linear attention architecture, which explicitly averages its inputs through the $1/\ell$ term that appears in the architecture definition (2); thus,

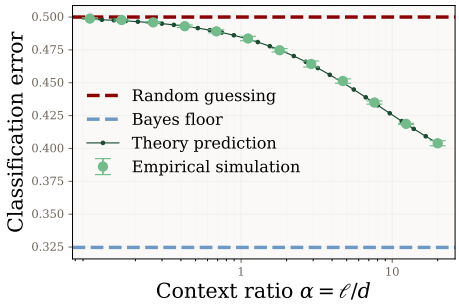


Figure 3: Simulation versus theory for classification error. We can compare to random guessing error 1/2, and Bayes error $1 - \mathbb{E}[\sigma(\beta|G_{\mathrm{q}}|)]$. For short contexts, the model effectively guesses, but nears the Bayes floor as context length increases. $d = 32, \alpha_{\mathrm{train}} = \alpha_{\mathrm{test}}, \tau = 10, \beta = 1, \lambda = 10^{-8}$.

longer test contexts always reduce noise and improve error (Letey et al., 2026b). Whether or not this particular phenomenon generalises to other architectures is a question for future work.

Calibration and noise Under the marginal laws (18) and (19), the target logit is $G_{\mathrm{true}} = \beta G_{\mathrm{q}}$ and the predicted logit is

$$R = \frac{a_1 \mathcal{C}_\beta}{\beta} G_{\mathrm{true}} + \sqrt{V}\, G_{\mathrm{var}}, \quad V = \frac{a_2}{2\alpha_{\mathrm{test}}} + (a_2 - a_1^2)\, \mathcal{C}_\beta^2, \quad G_{\mathrm{var}} \sim \mathcal{N}(0,1).$$

By Result 3, this logit expression makes the various sources of error explicit: the scaling coefficient $a_1\mathcal{C}_\beta/\beta$ in front of G_{true}, the variance due to test context length α_{test}, and the variance due to anisotropy $a_2 - a_1^2$. These interact non-trivially, as

$$G_{\mathrm{true}} \mid R \sim \mathcal{N}\left(\frac{a_1 \beta \mathcal{C}_\beta}{a_1^2 \mathcal{C}_\beta^2 + V} R,\ \frac{\beta^2}{a_1^2 \mathcal{C}_\beta^2 + V} V \right).$$

Removing the noise due to test-time context length then gives

$$G_{\mathrm{true}} \mid R \sim \mathcal{N}\left(\frac{\beta}{\mathcal{C}_\beta} \frac{a_1}{a_2} R,\ \beta^2 \left(1 - \frac{a_1^2}{a_2}\right) \right). \tag{22}$$

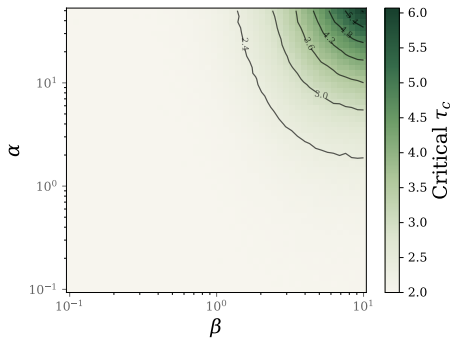
The takeaway here is that while the coefficient of G_{true}, namely $a_1\mathcal{C}_\beta/\beta$, appears like a calibration factor, the actual calibration depends jointly on this factor as well as the isotropy measure a_1^2/a_2 through

$$\frac{\beta}{\mathcal{C}_\beta} \frac{a_1}{a_2} = \frac{a_1^2}{a_2} \Big/ \frac{a_1 \mathcal{C}_\beta}{\beta}.$$

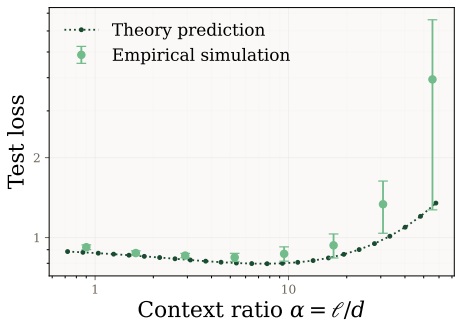
We see that this recovers cleanly all intuitions from the population limit, as well as observations from Figure 1: for the output logit to behave like the latent logit, we require both correct scaling (through $a_1\mathcal{C}_\beta/\beta$) and isotropy (through a_1^2/a_2).

Location of critical batch size Figure 4a maps the predicted dependence of τ_c on context length α_{train} and sampling inverse temperature β. We see that increasing either α_{train} or β will increase the critical ratio τ_c. This can be understood from prior intuition: increasing either α_{train} or β will increase the "signal" of the features relative to the zero-signal (uniformly random y_{q}s) case, thus making separability easier to achieve, and pushing the critical sample ratio away from 2.

This has potentially interesting practical consequences, as a natural conclusion from this analysis is that increasing training context length at low batch sizes (low τ) might be harmful, as τ_c increases with α_{train} and so the nominal batch ratio τ is closer to the blow-up regime (see Figure 4b). This is in tension with the general intuition that longer context length at training time should improve in-context ability by providing more accurate moments used by linear attention. This phenomenon



(a) Heatmap of critical τ_c, defined by (16), against α and β.



(b) Increasing α for fixed τ near $\tau_c(\alpha_{\max}, \beta)$ increases $\mathcal{E}_{\text{CE}}$.

Figure 4: Effect of α and β on separability threshold τ_c, and the impact of $\tau_c(\alpha, \beta)$ on ridgeless ICL error (as measured by CE loss).

highlights the importance of understanding where τ_c is for a given task and architecture when trying to improve ICL performance. In particular, the transition in general depends on the ratio of data size (sequences in this case) to effective parameters, which may change with different architectures.

5 Beyond the Reduced Model

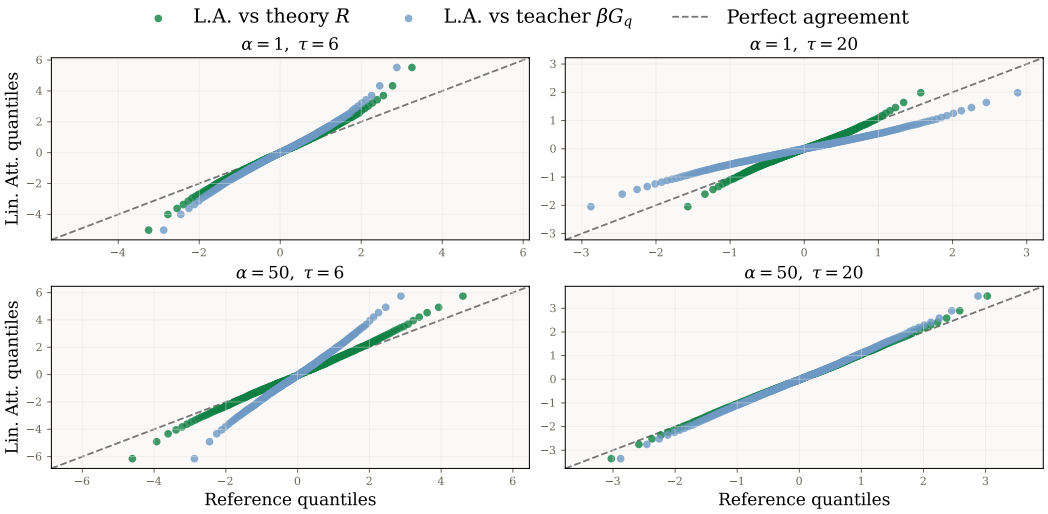


Figure 5: Similarly to Figure 1, we compare the theory to full-parameter linear attention, namely, we train the full K, Q, V matrices. Here we plot a quantile-quantile comparison of the model's output logit on new prompt matrices Z against the theory prediction R and the true latent βG_{q}. In all plots, $d = 16, \lambda = 10^{-6}, \beta = 1$. Regularisation is applied to $\|K\|_F^2 + \|Q\|_F^2 + \|V\|_F^2$.

It is important to characterise how well this model extends beyond the theoretical regime. Recall we began with the fully-parameterised linear attention architecture in terms of K, Q, V matrices, and ultimately studied the reduced model in terms of parameters Γ. Despite this modelling choice for the theory, the prediction in Result 3 also accurately describes the linear-attention architecture with fully trained K, Q, V matrices. This is shown in Figure 5, where we see close agreement between the theory logit R and the trained linear attention output distributions, with small deviation in the tails for small α and τ.

We do not expect the exact theoretical characterisation to transfer directly to other architectures. In particular, conclusions about test-time context length likely depend on the explicit averaging structure of linear attention. Regardless, Figure 6 shows that the trained softmax-attention model exhibits some qualitatively similar behavior. We include the linear-attention theory distribution strictly as a reference; it is not a theoretical prediction for the softmax model.

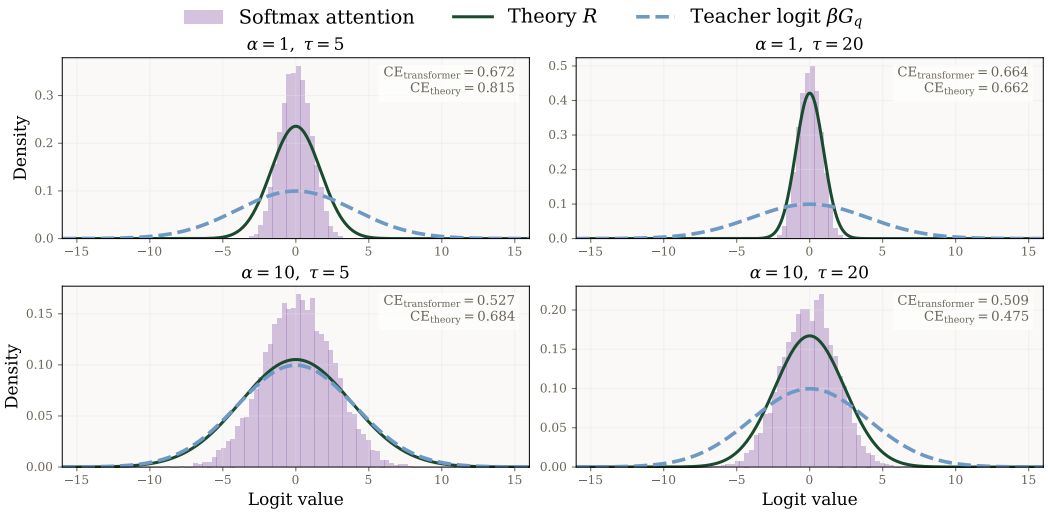


Figure 6: An analogous comparison to Figure 1, but the logits here are from a trained transformer-style softmax attention block. In particular, we use standard scaled dot-product softmax attention with fully trainable query Q, key K, value V, and output O weights. We initialise these at identity to improve training dynamics. The model is trained for 100 epochs with AdamW with weight decay 0.1 and early stopping. We set $d = 32$ and $\beta = 4$.

Across all four settings, the softmax logits have noticeably lower variance than both the linear-attention theory prediction and the teacher logits. Increasing the context ratio α substantially broadens the softmax distribution and improves its alignment with the teacher distribution. The effect of increasing the task ratio τ is less straightforward: it reduces the softmax-logit variance in these experiments rather than producing uniformly closer agreement with the teacher. The softmax model achieves lower cross-entropy than the linear-attention theory reference for $\tau = 5$, whereas the theory reference performs better for $\tau = 20$. These results suggest potentially interesting differences in calibration and confidence between the two architectures, which we leave for future investigation.

6 Conclusion

Motivated by a large body of work on how linear attention performs simple in-context regressions, we have characterised the high-dimensional asymptotics of in-context binary classification for a linear attention model with a softmax readout. We derive an interpretable joint law for the latent and predicted logits: the output logits are the sum of a scaled version of the true logit, with an additional variance-expanding term that captures finite-size effects from test-time context length and finite-pretraining-batch anisotropy. We show that both anisotropy and scaling of the logit affect calibration with the true logit. This gives a clear taxonomy of how ICL can be improved by increasing the context length or the pretraining data diversity, which could be tested in experiments with real data and different architectures. Further, since we have a formula for the output logit, this opens the door for many additional computations of interest. For example, one may differentiate the error functions with respect to ridge to derive optimality conditions for the ridge parameter in terms of data. Empirically we have found that optimal ridge depends interestingly on α_{test} and α_{train}; this warrants further investigation.

AI USE STATEMENT

We used generative AI tools for various steps throughout the research and writing processes. We describe this thoroughly as follows.

- Synthetic dataset generation. No.
- Develop theoretical models. No.
- Formulate mathematical claims. AI was used to help state our formulas more clearly in the document; it was not used to formulate the actual mathematics.
- Provide critical components for proofs. No.
- Assist in the writing of proofs. Yes. GPT 5.6 and 6 were used to verify and check our proofs and formulas; GPT 6 suggested a simplification of the inverse-Hessian expression in Appendix A such that (14) does not involve the inverse Hessian, which helps with numerical stability.
- Propose or refine hypotheses. No.
- Design research methodology. No.
- Provide feedback on research methodology. Yes, namely verification of our proof and method.
- Implement methods. Yes. GPT 5.6 and GPT 6 were used to generate substantial blocks of the code, particularly for simulations of the theory; all implementations and code were commented, rewritten, and verified by the lead authors. The most significant use of AI to generate code was for the final softmax experiments, which we offer as a preliminary investigation.
- Assist with translation. No.
- Clean and reformat datasets. No.
- Support data analysis. No.
- Interpret results. AI was used in the writing process, with specific prompts, where we asked specific questions in order to improve the clarity, readability, and logical flow of our results. Some of this naturally includes interpretation.
- Brainstorming. Yes. The conception of this paper occurred through a brainstorming session between the senior author and an agent model. In particular, this included the suggestion that the softmax-readout linear attention architecture might be analysable at the population level.

AI was not used to draft the manuscript or generate the literature review; it was used to polish prose and for suggestions on overall presentation. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artefacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All code for simulations will be provided upon acceptance.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *International Conference on Learning Representations*, volume 2024, pp. 17686–17698, 2024.

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.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 57125–57211. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/b2e63e36c57e153b9015fece2352a9f9-Paper-Conference.pdf.

Jean Barbier, Florent Krzakala, Nicolas Macris, Léo Miolane, and Lenka Zdeborová. Optimal errors and phase transitions in high-dimensional generalized linear models. *Proceedings of the National Academy of Sciences*, 116(12):5451–5460, 2019.

Blake Bordelon, Mary I. Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time, 2025. URL https://arxiv.org/abs/2510.01098.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Emmanuel J. Candes and Pragya Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression, 2018. URL https://arxiv.org/abs/1804.09753.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers, 2022.

Thomas M Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

Noureddine El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175, 2018.

Elizabeth Gardner. The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257, jan 1988. doi: 10.1088/0305-4470/21/1/030. URL https://doi.org/10.1088/0305-4470/21/1/030.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pp. 3452–3462. PMLR, 2020.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020.

Takuya Koriyama and Pierre C. Bellec. Phase transitions for the existence of unregularized m-estimators in single index models, 2025. URL https://arxiv.org/abs/2501.03163.

Mary Letey, Yue M. Lu, Cengiz Pehlevan, and Jacob A Zavatone-Veth. Sequential correlations change in-context learning: Effective context length and architectural mismatch. In *High-dimensional Learning Dynamics 2026*, 2026a. URL https://openreview.net/forum?id=AYprbQxAbI.

Mary Letey, Jacob A Zavatone-Veth, Yue M. Lu, and Cengiz Pehlevan. Pretrain-test task alignment governs generalization in in-context learning. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=KZLeg0MQ2r.

Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025. doi: 10.1073/pnas.2502599122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2502599122.

Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3357–3361, 2019. doi: 10.1109/ICASSP.2019.8683376.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75 (4):667–766, 2022.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Pretrained transformer efficiently learns low-dimensional target functions in-context, 2024. URL https://arxiv.org/abs/2411.02544.

Luca Pesce, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Are gaussian data all you need? extents and limits of universality in high-dimensional generalized linear estimation, 2023. URL https://arxiv.org/abs/2302.08923.

Pragya Sur and Emmanuel J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019. ISSN 1091-6490. doi: 10.1073/pnas.1810420116. URL http://dx.doi.org/10.1073/pnas.1810420116.

Bhavya Vasudeva, Deqing Fu, Tianyi Zhou, Elliott Kau, Youqi Huang, and Vatsal Sharan. Transformers learn low sensitivity functions: Investigations and implications. In *International Conference on Learning Representations*, volume 2025, pp. 7805–7834, 2025.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vSh5ePa0ph.

Chenyang Zhang and Yuan Cao. Transformers efficiently perform in-context logistic regression via normalized gradient descent. *arXiv preprint arXiv:2605.06609*, 2026.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL http://jmlr.org/papers/v25/23-1042.html.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization, 2023. URL https://arxiv.org/abs/2305.19420.

A Derivation of the cavity result

The following provides a derivation of Result 1 and Result 3 together. Note that Result 4 follows immediately from 3.

Proof. We will decompose the proof into steps.

Step 1. Proximal equation. Write $f_y(z) = L(z) - yz$ and let $\widehat{\boldsymbol{\gamma}}_{[\nu]}$ minimise the objective with prompt ν removed. Let

$$\mathcal{H}_{[\nu]} = \frac{1}{n}\sum_{\mu\neq\nu} \sigma'(R_{\mu,[\nu]})\boldsymbol{h}_\mu\boldsymbol{h}_\mu^\top + \lambda d I_{d(d+1)}, \quad R_{\mu,[\nu]} = \boldsymbol{h}_\mu^\top \widehat{\boldsymbol{\gamma}}_{[\nu]}. \tag{23}$$

A quadratic expansion about $\widehat{\boldsymbol{\gamma}}_{[\nu]}$ gives

$$\widehat{\boldsymbol{\gamma}} \approx \arg\min_{\boldsymbol{\gamma}} \left\{ \frac{1}{n} f_{y_\nu}(\boldsymbol{h}_\nu^\top \boldsymbol{\gamma}) + \frac{1}{2}(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{[\nu]})^\top \mathcal{H}_{[\nu]}(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}}_{[\nu]}) \right\}. \tag{24}$$

Minimizing at fixed $z = \boldsymbol{h}_\nu^\top \boldsymbol{\gamma}$ yields

$$R_\nu = \mathrm{prox}_{s_\nu f_{y_\nu}}(R_{\nu,[\nu]}), \qquad R_{\nu,[\nu]} = R_\nu + s_\nu(\sigma(R_\nu) - y_\nu), \qquad s_\nu = \frac{1}{n}\boldsymbol{h}_\nu^\top \mathcal{H}_{[\nu]}^{-1}\boldsymbol{h}_\nu. \tag{25}$$

Logit convention. Before taking the high-dimensional limit, $R_\mu := \boldsymbol{h}_\mu^\top \widehat{\boldsymbol{\gamma}}$ denotes the full-sample fitted logit, while $R_{\mu,[\nu]} := \boldsymbol{h}_\mu^\top \widehat{\boldsymbol{\gamma}}_{[\nu]}$ denotes the logit of prompt μ when prompt ν is omitted from training. In particular, $R_{\mu,[\mu]}$ is the cavity logit for the held-out prompt. Only after invoking concentration and exchangeability do we identify the limiting joint law of $(R_{\mu,[\mu]}, R_\mu, y_\mu)$ with $(R_{\mathrm{loo}}, R_s, Y)$.

Step 2. Hessian prefactor. We are interested in writing a characterisation for the left-out Hessian prefactor terms that appear in the proximal equation

$$s_\nu := \frac{1}{n}\boldsymbol{h}_\nu^\top \mathcal{H}_{[\nu]}^{-1}\boldsymbol{h}_\nu.$$

This is a heuristic argument; we mark the assumptions used below explicitly. We have

$$\begin{aligned}
\mathcal{L}_{[\nu]} &= \frac{1}{n}\sum_{\mu\neq\nu} \big(L(\boldsymbol{\gamma}^\top \boldsymbol{h}_\mu) - y_\mu \boldsymbol{\gamma}^\top \boldsymbol{h}_\mu\big) + \frac{\lambda d}{2}\|\boldsymbol{\gamma}\|^2 \\
\frac{\mathrm{d}\mathcal{L}_{[\nu]}}{\mathrm{d}\boldsymbol{\gamma}} &= \frac{1}{n}\sum_{\mu\neq\nu} \big(L'(\boldsymbol{\gamma}^\top \boldsymbol{h}_\mu) - y_\mu\big)\boldsymbol{h}_\mu + \lambda d\boldsymbol{\gamma} \\
\mathcal{H}_{[\nu]} &= \frac{1}{n}\sum_{\mu\neq\nu} L''(R_{\mu,[\nu]})\boldsymbol{h}_\mu\boldsymbol{h}_\mu^\top + \lambda d I_{d(d+1)}
\end{aligned}$$

As in the quadratic expansion, we evaluate

$$\mathcal{H}_{[\nu]} := \frac{\mathrm{d}^2\mathcal{L}_{[\nu]}}{\mathrm{d}\boldsymbol{\gamma}^2}\Big(\boldsymbol{\gamma} = \boldsymbol{\gamma}^*_{[\nu]}\Big).$$

Eventually we will compare this to the Hessian of the full loss. We have two choices of where to evaluate such an object, at $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*$ or at $\boldsymbol{\gamma} = \boldsymbol{\gamma}^*_{[\nu]}$. In the latter, they differ by a rank-one term

$$\mathcal{H}(\boldsymbol{\gamma}^*_{[\nu]}) := \frac{\mathrm{d}^2\mathcal{L}}{\mathrm{d}\boldsymbol{\gamma}^2}\Big(\boldsymbol{\gamma} = \boldsymbol{\gamma}^*_{[\nu]}\Big) = \mathcal{H}_{[\nu]} + \frac{1}{n}L''(R_{\nu,[\nu]})\boldsymbol{h}_\nu\boldsymbol{h}_\nu^\top.$$

For the former, we must bound $R_\mu - R_{\mu,[\nu]}$ induced differences, as

$$\mathcal{H}(\boldsymbol{\gamma}^*) := \frac{\mathrm{d}^2\mathcal{L}}{\mathrm{d}\boldsymbol{\gamma}^2}(\boldsymbol{\gamma} = \boldsymbol{\gamma}^*) = \mathcal{H}_{[\nu]} + \frac{1}{n}L''(R_\nu)\boldsymbol{h}_\nu\boldsymbol{h}_\nu^\top + \frac{1}{n}\sum_{\mu\neq\nu}\big(L''(R_\mu) - L''(R_{\mu,[\nu]})\big)\boldsymbol{h}_\mu\boldsymbol{h}_\mu^\top.$$

We assume, without proof, that the matrix remainder

$$\frac{1}{n}\sum_{\mu\neq\nu}\big(L''(R_\mu)-L''(R_{\mu,[\nu]})\big)\boldsymbol{h}_\mu\boldsymbol{h}_\mu^\top$$

can be neglected in the normalised inverse-Hessian quadratic forms and traces used below. Expand

$$\begin{aligned}\mathcal{H}(\boldsymbol{\gamma}^*)^{-1}:=\mathcal{H}^{-1}&\approx\left(\mathcal{H}_{[\nu]}+\frac{1}{n}L''(R_\nu)\boldsymbol{h}_\nu\boldsymbol{h}_\nu^\top\right)^{-1}\\&=\mathcal{H}_{[\nu]}^{-1}-\frac{1}{n}\frac{L''(R_\nu)}{1+L''(R_\nu)s_\nu}\mathcal{H}_{[\nu]}^{-1}\boldsymbol{h}_\nu\boldsymbol{h}_\nu^\top\mathcal{H}_{[\nu]}^{-1}\\ \text{so}\quad\frac{1}{n}\boldsymbol{h}_\nu^\top\mathcal{H}^{-1}\boldsymbol{h}_\nu&\approx\frac{s_\nu}{1+L''(R_\nu)s_\nu}.\end{aligned}$$

We have

$$I=\mathcal{H}^{-1}\mathcal{H}=\lambda d\mathcal{H}^{-1}+\frac{1}{n}\sum_\mu L''(R_\mu)\mathcal{H}^{-1}\boldsymbol{h}_\mu\boldsymbol{h}_\mu^\top$$

and so

$$\frac{1}{n}\sum_{\mu=1}^n\frac{1}{1+L''(R_\mu)s_\mu}-\frac{\lambda d}{n}\operatorname{tr}\big(\mathcal{H}^{-1}\big)\approx 1-\frac{1}{\tau}.$$

We can simplify the s_μ quadratic forms as follows. Recall the structure of $\boldsymbol{h}=\operatorname{vec}(\boldsymbol{x}_{\text{test}}\boldsymbol{u}^\top)$ and so for any matrix $A\in\mathbb{R}^{d(d+1)\times d(d+1)}$, we have

$$\frac{1}{n}\boldsymbol{h}^\top A\boldsymbol{h}=\frac{1}{n}x_iu_aA_{ia,jb}x_ju_b.$$

The $\boldsymbol{x}$s are mean-zero, sub-Gaussian, with independent coordinates, and so if we write matrices $B_{ij}=u_aA_{ia,jb}u_b$ then

$$\frac{1}{n}\boldsymbol{h}^\top A\boldsymbol{h}\approx\frac{1}{n}\operatorname{tr}\big(B\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\big)=\frac{1}{n}\operatorname{tr}(B).$$

For the $\boldsymbol{u}$ terms, conditional on $\boldsymbol{w}$,

$$\begin{aligned}\mathbb{E}[\boldsymbol{u}^\top\mid\boldsymbol{w}]&=\begin{bmatrix}d\mathcal{C}_\beta\boldsymbol{w}^\top & 1/2\end{bmatrix}\\ \mathbb{E}\left[\boldsymbol{u}\boldsymbol{u}^\top\mid\boldsymbol{w}\right]&=\begin{bmatrix}d^2\mathcal{C}_\beta^2\boldsymbol{w}\boldsymbol{w}^\top+\dfrac{d}{\alpha_{\text{train}}}\left(\dfrac{1}{2}I_d-\mathcal{C}_\beta^2\boldsymbol{w}\boldsymbol{w}^\top\right) & \left(\dfrac{d\,\mathcal{C}_\beta}{2}+\dfrac{\mathcal{C}_\beta}{2\alpha_{\text{train}}}\right)\boldsymbol{w}\\ \left(\dfrac{d\,\mathcal{C}_\beta}{2}+\dfrac{\mathcal{C}_\beta}{2\alpha_{\text{train}}}\right)\boldsymbol{w}^\top & \dfrac{1}{4}+\dfrac{1}{4\alpha_{\text{train}}d}\end{bmatrix}\end{aligned}$$

where

$$\mathcal{C}_\beta:=\mathbb{E}_{g\sim\mathcal{N}(0,1)}[g\sigma(\beta g)].$$

Averaging over the $\|\boldsymbol{w}\|=1$ spherically uniform task distribution gives

$$\mathbb{E}[\boldsymbol{u}^\top]=[\mathbf{0}\quad 1/2]\tag{26}$$

$$\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top]\approx\begin{bmatrix}d\Big(\mathcal{C}_\beta^2+\frac{1}{2\alpha_{\text{train}}}\Big)I_d & \mathbf{0}\\ \mathbf{0}^\top & \frac{1}{4}\end{bmatrix}:=\Sigma_\beta\tag{27}$$

For the leave-one-out inverse Hessians considered below, we assume

$$\frac{1}{n}\boldsymbol{h}^\top A\boldsymbol{h}\approx\frac{1}{n}\operatorname{tr}(A(I_d\otimes\Sigma_\beta)).$$

So we will write

$$s_\nu\approx\frac{1}{n}\operatorname{tr}\Big(\mathcal{H}_{[\nu]}^{-1}(I_d\otimes\Sigma_\beta)\Big).$$

Now we assume that $\mathcal{L}, \mathcal{L}_{[\nu]}$ are sufficiently well behaved so that the relevant normalised traces of the inverse Hessians at their respective minimisers are comparable, i.e.

$$\frac{1}{n}\operatorname{tr}\left(\mathcal{H}_{[\nu]}^{-1}\Sigma_h\right) \approx \frac{1}{n}\operatorname{tr}\left(\mathcal{H}^{-1}\Sigma_h\right).$$

Thus for $\Sigma_h = I_d \otimes \Sigma_\beta$, we write

$$s_\nu \approx \frac{1}{n}\operatorname{tr}\Big(\mathcal{H}_{[\nu]}^{-1}\Sigma_h\Big) \approx \frac{1}{n}\operatorname{tr}\big(\mathcal{H}^{-1}\Sigma_h\big) := s.$$

We further assume that this quantity concentrates to a deterministic scalar s. We then have

$$\frac{1}{n}\sum_{\mu=1}^{n}\frac{1}{1+L''(R_\mu)s} - \frac{\lambda d}{n}\operatorname{tr}\big(\mathcal{H}^{-1}\big) = 1 - \frac{1}{\tau}.$$

Estimating $\operatorname{tr}\big(\mathcal{H}^{-1}\big)$ will be difficult and we wish to avoid it. For $\lambda > 0$, we make the additional approximation, without proof, that the contribution of the final vector block of Σ_β to the normalised inverse-Hessian trace is asymptotically negligible, so that the main block dominates. Thus,

$$s \approx \frac{1}{n}\operatorname{tr}\big(\mathcal{H}^{-1}(I_d \otimes \Sigma_\beta)\big) \approx \frac{d\Delta}{n}\operatorname{tr}\big(\mathcal{H}^{-1}\big).$$

For $\lambda = 0$, this approximation is not required, since the term involving $\operatorname{tr}\big(\mathcal{H}^{-1}\big)$ in the preceding equation vanishes.

To summarise, we have that

$$s_\nu = \frac{1}{n}\boldsymbol{h}_\nu^\top \mathcal{H}_{[\nu]}^{-1}\boldsymbol{h}_\nu \approx s$$

for deterministic scalar s defined self-consistently by

$$1 - \frac{1}{\tau} = \mathbb{E}\left[\frac{1}{1+L''(R_s)s}\right] - \frac{\lambda}{\Delta}s.$$

In the limit as $\lambda \to 0$, we have the nicer ridgeless characterisation

$$1 - \frac{1}{\tau} = \mathbb{E}\left[\frac{1}{1+L''(R_s)s}\right].$$

Step 3. Gaussian equivalent. We will construct a logit that is statistically equivalent to the left-out logits R_{loo}. To do this, we make a standard assumption commonly used in cavity arguments, that asymptotically there is no difference between $\boldsymbol{h}_\mu^\top \boldsymbol{\gamma}_{[\mu]}^*$ and $\boldsymbol{h}_{\text{new}}^\top \boldsymbol{\gamma}^*$. Moreover, $\boldsymbol{\gamma}_{[\mu]}^*$ is independent of $\boldsymbol{h}_\mu$, since prompt μ is excluded from its training set. Here $\boldsymbol{h}_{\text{new}}$ is drawn from the same distribution as the training distribution, namely context length is still ℓ and is not ℓ_{test}. This distinction is important, as $\boldsymbol{h}_{\text{new}}^\top \boldsymbol{\gamma}^*$ here is not the test logit, which would describe sequences of length ℓ_{test}.

Thus we have that

$$R_{\text{loo}} \overset{d}{=} \boldsymbol{x}_q^\top\left(A^*\boldsymbol{v} + c\boldsymbol{a}^*\right), \qquad \boldsymbol{v} := \frac{d}{\ell}\sum_{i=1}^{\ell} y_i \boldsymbol{x}_i, \qquad c := \frac{1}{\ell}\sum_{i=1}^{\ell} y_i.$$

Write

$$A^* = \frac{\widehat{a}_1}{d} I_d + \bar{A}^*, \qquad \widehat{a}_1 := \operatorname{tr}(A^*), \qquad \operatorname{tr}\big(\bar{A}^*\big) = 0.$$

Then

$$R_{\text{loo}} = \frac{\widehat{a}_1}{d}\boldsymbol{x}_q^\top \boldsymbol{v} + \boldsymbol{x}_q^\top \bar{A}^* \boldsymbol{v} + c\,\boldsymbol{x}_q^\top \boldsymbol{a}^*.$$

We analyse these three terms separately.

We make a standard cavity self-averaging assumption: conditional on the leave-one-out estimator, the relevant overlaps and quadratic forms of a fresh prompt concentrate around some deterministic limits. Since $\boldsymbol{x}_q$ is Gaussian and independent of context variables and the leave-one-out estimator, the terms

below are conditionally jointly Gaussian. As their conditional covariances concentrate, their limiting joint law is derived by the corresponding covariance calculations below.

First, write

$$\boldsymbol{x}_q = G_q \boldsymbol{w} + \boldsymbol{x}_{\text{perp}}, \qquad G_q := \boldsymbol{x}_q^\top \boldsymbol{w} \sim \mathcal{N}(0,1), \qquad \boldsymbol{x}_{\text{perp}} \sim \mathcal{N}(0, I_d - \boldsymbol{w}\boldsymbol{w}^\top).$$

Conditional on $\boldsymbol{w}$, we have

$$\mathbb{E}[\boldsymbol{v} \mid \boldsymbol{w}] = d\mathcal{C}_\beta \boldsymbol{w}$$

and

$$\mathrm{Cov}(\boldsymbol{v} \mid \boldsymbol{w}) = \frac{d}{2\alpha_{\text{train}}} I_d - \frac{d\mathcal{C}_\beta^2}{\alpha_{\text{train}}} \boldsymbol{w}\boldsymbol{w}^\top.$$

Consequently,

$$\frac{1}{d}\boldsymbol{x}_q^\top \boldsymbol{v} = \mathcal{C}_\beta G_q + \frac{1}{d}\boldsymbol{x}_q^\top \left(\boldsymbol{v} - d\mathcal{C}_\beta \boldsymbol{w}\right).$$

Using concentration of quadratic forms,

$$\mathrm{Var}\left(\frac{1}{d}\boldsymbol{x}_q^\top \left(\boldsymbol{v} - d\mathcal{C}_\beta \boldsymbol{w}\right)\right) \approx \frac{1}{2\alpha_{\text{train}}}.$$

Moreover, its covariance with G_q is asymptotically zero. Under the self-averaging assumption above, it results in

$$\frac{\widehat{a}_1}{d}\boldsymbol{x}_q^\top \boldsymbol{v} \overset{d}{\approx} \widehat{a}_1 \mathcal{C}_\beta G_q + \frac{\widehat{a}_1}{\sqrt{2\alpha_{\text{train}}}} G_{\text{tr}},$$

where $G_{\text{tr}} \sim \mathcal{N}(0,1)$ is independent of G_q.

We next consider the traceless contribution. Since

$$\mathrm{tr}\big(\bar{A}^*\big) = 0,$$

we have

$$\mathbb{E}\left[\boldsymbol{x}_q^\top \bar{A}^* \boldsymbol{v}\right] \approx d\mathcal{C}_\beta \mathbb{E}\left[\boldsymbol{x}_q^\top \bar{A}^* \boldsymbol{w}\right] = 0.$$

We have

$$\mathbb{E}\left[\left(\boldsymbol{x}_q^\top \bar{A}^* \boldsymbol{v}\right)^2\right] \approx d^2\mathcal{C}_\beta^2 \mathbb{E}\left[\left(\boldsymbol{x}_q^\top \bar{A}^* \boldsymbol{w}\right)^2\right] + \frac{d}{2\alpha_{\text{train}}} \mathbb{E}\left[\boldsymbol{x}_q^\top \bar{A}^* \bar{A}^{*\top} \boldsymbol{x}_q\right].$$

Averaging over $\boldsymbol{x}_q$ and $\|\boldsymbol{w}\|^2 = 1$ gives

$$\mathbb{E}\left[\left(\boldsymbol{x}_q^\top \bar{A}^* \boldsymbol{w}\right)^2\right] = \frac{1}{d}\|\bar{A}^*\|_F^2$$

and

$$\mathbb{E}\left[\boldsymbol{x}_q^\top \bar{A}^* \bar{A}^{*\top} \boldsymbol{x}_q\right] = \|\bar{A}^*\|_F^2.$$

Therefore

$$\mathbb{E}\left[\left(\boldsymbol{x}_q^\top \bar{A}^* \boldsymbol{v}\right)^2\right] \approx d\|\bar{A}^*\|_F^2 \left(\mathcal{C}_\beta^2 + \frac{1}{2\alpha_{\text{train}}}\right).$$

Note that

$$d\|\bar{A}^*\|_F^2 = \widehat{a}_2 - \widehat{a}_1^2$$

and so

$$\boldsymbol{x}_q^\top \bar{A}^* \boldsymbol{v} \overset{d}{\approx} G_{\text{traceless}} \sqrt{(\widehat{a}_2 - \widehat{a}_1^2)\left(\mathcal{C}_\beta^2 + \frac{1}{2\alpha_{\text{train}}}\right)}.$$

Finally, since $c \approx 1/2$, we have

$$c\,\boldsymbol{x}_q^\top \boldsymbol{a}^* \overset{d}{\approx} G_{\text{vec}} \sqrt{\frac{\widehat{a}_0}{4}}, \qquad \widehat{a}_0 := \|\boldsymbol{a}^*\|^2.$$

Combining everything gives a joint statistically-equivalent characterisation

$$R_{\text{loo}} \overset{d}{=} \widehat{a}_1 \mathcal{C}_\beta G_{\text{q}} + \widehat{a}_1 \frac{1}{\sqrt{2\alpha_{\text{train}}}} G_{\text{tr}} + G_{\text{traceless}} \sqrt{(\widehat{a}_2 - \widehat{a}_1^2)\left(\mathcal{C}_\beta^2 + \frac{1}{2\alpha_{\text{train}}}\right)} + G_{\text{vec}} \sqrt{\frac{\widehat{a}_0}{4}},$$

where

$$G_{\text{q}}, G_{\text{tr}}, G_{\text{traceless}}, G_{\text{vec}} \sim_{\text{i.i.d.}} \mathcal{N}(0,1).$$

Step 4. Closing the proximal equation. We now assume that the empirical order parameters $(\widehat{a}_1, \widehat{a}_2, \widehat{a}_0)$ concentrate around the deterministic limits (a_1, a_2, a_0). We also assume that $s_\mu \to s$. Exchangeability then identifies the limiting joint law of $(R_{\mu,[\mu]}, R_\mu, y_\mu)$ with (R_{loo}, R_s, Y), where $R_{\text{loo}} = R_s + s(\sigma(R_s) - Y)$. From this point onward we use the scalar variables (R_{loo}, R_s, Y) after converting empirical averages to expectations. The proof here will proceed in a heuristic fashion. For $\lambda \to 0$, we assume without proof that the limiting order parameters are $O(1)$ with respect to d; we have numerically validated this. For $\lambda \neq 0$, additional control of the scaling of the order parameters is required, so we retain the corresponding terms in the final expression.

To proceed, we can differentiate $\mathcal{L}$ with respect to our desired parameters to get

$$\frac{\partial \mathcal{L}}{\partial \widehat{a}_1} = \frac{1}{n}\sum_\mu (\sigma(R_\mu) - y_\mu)\frac{\boldsymbol{x}_{\text{q},\mu}^\top \boldsymbol{v}_\mu}{d} + \lambda \widehat{a}_1 = 0$$
$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{a}} = \frac{1}{n}\sum_\mu (\sigma(R_\mu) - y_\mu) c_\mu \boldsymbol{x}_{\text{q},\mu} + \lambda d \boldsymbol{a} = \boldsymbol{0}.$$

We focus first on the $\widehat{a}_1$ equation. For $\lambda = 0$ this expression simplifies; for general λ, we retain the $\lambda\widehat{a}_1$ term. Using the proximal equation, concentration $\widehat{a}_1 \to a_1$, and approximating $1/n\sum_\mu$ by an expectation, we have

$$\mathbb{E}\left[(R_{\text{loo}} - R_s)\frac{1}{d}\boldsymbol{x}_{\text{q}}^\top \boldsymbol{v}\right] + s\lambda a_1 = 0.$$

We can expand

$$\begin{aligned}\mathbb{E}[R_{\text{loo}}\frac{1}{d}\boldsymbol{x}_{\text{q}}^\top \boldsymbol{v}] &= \frac{1}{d}\mathbb{E}\big[\text{tr}\left((A\boldsymbol{v}\boldsymbol{v}^\top + \boldsymbol{a}(c\boldsymbol{v})^\top)\boldsymbol{x}_{\text{q}}\boldsymbol{x}_{\text{q}}^\top\right)\big] \\ &= \frac{1}{d}\mathbb{E}\big[\text{tr}\left(A\boldsymbol{v}\boldsymbol{v}^\top + \boldsymbol{a}(c\boldsymbol{v})^\top\right)\big] \\ &= a_1\left(\mathcal{C}_\beta^2 + \frac{1}{2\alpha_{\text{train}}}\right).\end{aligned}$$

Furthermore, using the same tricks as previously, we can write the marginal for $\boldsymbol{x}_{\text{q}}^\top \boldsymbol{v}$ as

$$\frac{1}{d}\boldsymbol{x}_{\text{q}}^\top \boldsymbol{v} \overset{d}{=} \mathcal{C}_\beta G_{\text{q}} + \frac{1}{\sqrt{2\alpha_{\text{train}}}}G_{\text{tr}}.$$

Thus

$$a_1 = \frac{\mathbb{E}\Big[R_s\Big(\mathcal{C}_\beta G_{\text{q}} + \frac{1}{\sqrt{2\alpha_{\text{train}}}}G_{\text{tr}}\Big)\Big]}{\Delta + s\lambda}.$$

Now for $\boldsymbol{a}$, set

$$W := c\,\boldsymbol{a}^{*\top}\boldsymbol{x}_{\text{q}}.$$

The vector first-order condition and the proximal equation give

$$\begin{aligned}0 &= \mathbb{E}\big[(\sigma(R_s) - Y)W\big] + \lambda d\widehat{a}_0 \\ &= \frac{1}{s}\mathbb{E}\big[(R_{\text{loo}} - R_s)W\big] + \lambda d\widehat{a}_0.\end{aligned}$$

Thus

$$\mathbb{E}[R_{\text{loo}}W] - \mathbb{E}[R_s W] + s\lambda d\widehat{a}_0 = 0.$$

Under the Gaussian equivalent above, $W \overset{d}{\approx} \sqrt{\widehat{a}_0/4}\,G_{\text{vec}}$ and G_{vec} is independent of the other cavity variables. Hence Stein's lemma gives

$$\mathbb{E}[R_{\text{loo}}W] \approx \frac{\widehat{a}_0}{4}, \qquad \mathbb{E}[R_s W] \approx \frac{\widehat{a}_0}{4}\chi(s),$$

where $\chi(s) = \mathbb{E}[(1 + s\sigma'(R_s))^{-1}]$. Therefore, asymptotically,

$$\widehat{a}_0\left(\frac{1 - \chi(s)}{4} + s\lambda d\right) \longrightarrow 0.$$

This forces $a_0 = 0$ and, moreover, $s\lambda d\widehat{a}_0 \to 0$.

From the full derivative equation for $\boldsymbol{\gamma}$ we have

$$\mathbb{E}[(\sigma(R_s) - Y)R_s] + \lambda d\left(\frac{\widehat{a}_2}{d} + \widehat{a}_0\right) = 0.$$

Using the proximal equation, this becomes

$$\mathbb{E}[R_{\text{loo}}R_s] - \mathbb{E}[R_s^2] + s\lambda d\left(\frac{\widehat{a}_2}{d} + \widehat{a}_0\right) = 0.$$

Taking the limit and using $s\lambda d\widehat{a}_0 \to 0$ gives

$$\mathbb{E}[R_{\text{loo}}R_s] - \mathbb{E}[R_s^2] + s\lambda a_2 = 0.$$

In the ridgeless case, the final term vanishes. For finite ridge, we have the following self-consistency equation after using $a_0 = 0$:

$$a_2 = \frac{\mathbb{E}[R_s^2] - \Delta(1-\chi)a_1^2 - s\lambda a_1^2}{\Delta\chi + s\lambda}.$$

In the explicitly ridgeless case, we may expand $\mathbb{E}[R_{\text{loo}}R_s]$ at leading order, giving

$$\begin{aligned}\mathbb{E}[R_{\text{loo}}R_s] &= a_1\mathbb{E}\left[R_s\left(\mathcal{C}_\beta G_{\text{q}} + \frac{1}{\sqrt{2\alpha_{\text{train}}}}G_{\text{tr}}\right)\right] + \mathbb{E}\left[R_s G_{\text{traceless}}\sqrt{(a_2 - a_1^2)\left(\mathcal{C}_\beta^2 + \frac{1}{2\alpha_{\text{train}}}\right)}\right] \\ &= a_1^2\left(\mathcal{C}_\beta^2 + \frac{1}{2\alpha_{\text{train}}}\right) + (a_2 - a_1^2)\left(\mathcal{C}_\beta^2 + \frac{1}{2\alpha_{\text{train}}}\right)\mathbb{E}\left[\frac{\partial R_s}{\partial R_{\text{loo}}}\right] \quad \text{by Stein's lemma.}\end{aligned}$$

Now

$$\frac{\partial R_s}{\partial R_{\text{loo}}} = \frac{1}{1 + s\sigma'(R_s)}$$

from differentiating the proximal equation, and so

$$a_2 = \left(1 - 1/\mathbb{E}\left[\frac{1}{1 + s\sigma'(R_s)}\right]\right)a_1^2 + \frac{\mathbb{E}[R_s^2]}{\left(\mathcal{C}_\beta^2 + \frac{1}{2\alpha_{\text{train}}}\right)\mathbb{E}\left[\frac{1}{1+s\sigma'(R_s)}\right]}.$$

Step 5. Characterisation of the test logit. Here we apply a similar argument to one used in Step 3. We wish to characterise the test logit $\boldsymbol{h}_{\text{test}}^\top\boldsymbol{\gamma}^*$. Thus,

$$R_{\text{test}} = \boldsymbol{x}_q^\top\left(A^*\boldsymbol{v} + c\boldsymbol{a}^*\right), \qquad \boldsymbol{v} := \frac{d}{\ell_{\text{test}}}\sum_{i=1}^{\ell_{\text{test}}} y_i\boldsymbol{x}_i, \qquad c := \frac{1}{\ell_{\text{test}}}\sum_{i=1}^{\ell_{\text{test}}} y_i.$$

The argument for a statistically equivalent characterisation proceeds exactly the same as in Step 3, except the moments involve α_{test} now instead of α_{train}. This gives the desired marginal distribution

$$\begin{aligned}&y_{\text{q}} \sim \text{Bernoulli}(\sigma(\beta G_{\text{q}})) \\ &R_{\text{test}} = a_1\mathcal{C}_\beta G_{\text{q}} + a_1\frac{1}{\sqrt{2\alpha_{\text{test}}}}G_{\text{tr}} + G_{\text{traceless}}\sqrt{(a_2 - a_1^2)\Delta_{\text{test}}}, \quad \Delta_{\text{test}} := \mathcal{C}_\beta^2 + \frac{1}{2\alpha_{\text{test}}}.\end{aligned}$$

This concludes the proof of Result 1 and Result 3.

□

B SCHEMATIC DERIVATION OF SOLUTION BOUNDARY RESULT

This claim follows by applying the result of Koriyama & Bellec (2025) to our problem setting, which requires an additional heuristic transfer argument. The setting considered by Koriyama & Bellec (2025) is the classic teacher-student setting where the model being trained has Gaussian covariates $\boldsymbol{x}$. This assumption clearly does not hold in our setting, where not only do the model's covariates $\boldsymbol{h}$

not match the teacher covariates $\boldsymbol{x}$, but the $\boldsymbol{h}$s are not Gaussian. That said, we present the following schematic argument for the applicability of their result to our setting: each $\boldsymbol{h}$ takes the form

$$\boldsymbol{h} = \mathrm{vec}(H), \quad H = \boldsymbol{x}_{\mathrm{q}}\boldsymbol{u}^\top, \quad \boldsymbol{u}^\top = \left[\frac{d}{\ell}\sum_{i=1}^{\ell} y_i \boldsymbol{x}_i^\top \;\; \frac{1}{\ell}\sum_{i=1}^{\ell} y_i^2\right].$$

We see that $\boldsymbol{x}_{\mathrm{q}}$ is Gaussian, and $\boldsymbol{u}$ is nearly Gaussian (conditional on $\boldsymbol{w}$), allowing us to exploit the tensorised form of $\boldsymbol{h}$ to transfer the result of Koriyama & Bellec (2025) (similar to how we exploited this form in arguing the concentration of quadratic forms $\boldsymbol{h}^\top R\boldsymbol{h}$). We stress that this is a heuristic argument.

The second caveat is that the statement in Result 2 is quite informal. The correct statement is

$$\mathbb{P}(\text{finite solution to ridgeless problem exists}) \to \begin{cases} 0 & \tau < \tau_c \\ 1 & \tau > \tau_c \end{cases}.$$

Given the above, all that remains is to directly apply the result. We take $G \sim \mathcal{N}(0,1)$ independently, and define (U, Y) by the limiting joint distribution of $(\boldsymbol{x}_{\mathrm{q}}^\top \boldsymbol{v}/(d\sqrt{\Delta}), y_{\mathrm{q}})$. We know from the previous marginal characterisation that

$$U \propto \mathcal{C}_\beta G_{\mathrm{q}} + \frac{1}{\sqrt{2\alpha_{\mathrm{train}}}} G_{\mathrm{tr}}$$

where $G_{\mathrm{q}}, G_{\mathrm{tr}} \sim \mathcal{N}(0,1)$ with $Y \sim \text{Bernoulli}(\sigma(\beta G_{\mathrm{q}}))$. Normalising the variance of U gives the required result.

C Properties of loss functions with context length

Here we justify our claim that increasing test-time context length strictly decreases both losses. Intuitively, this follows because α_{test} only appears in the variance term V that contributes randomness that is independent of the actual signal G_{q}.

To be concrete, for the cross-entropy loss, we can write

$$\begin{aligned}\mathcal{E}_{\mathrm{CE}} &= \mathbb{E}[L(R)] - \mathbb{E}[\sigma(\beta G_{\mathrm{q}}) a_1 \mathcal{C}_\beta G_{\mathrm{q}}] \\ &= \mathbb{E}_{G\sim\mathcal{N}(0,1)}[\log(1+\exp(cG))] - a_1\mathcal{C}_\beta^2\end{aligned}$$

where

$$c = \sqrt{a_2\left(\frac{1}{2\alpha_{\mathrm{test}}} + \mathcal{C}_\beta^2\right)}$$

does not depend on G. Thus

$$\frac{\partial \mathcal{E}_{\mathrm{CE}}}{\partial c} = \mathbb{E}\left[\frac{G\exp(cG)}{1+\exp(cG)}\right] = \mathbb{E}\left[\frac{c\exp(cG)}{(1+\exp(cG))^2}\right] \geq 0$$

and the result follows as $\partial c/\partial \alpha_{\mathrm{test}} < 0$.

For the classification loss, write

$$\mathcal{E}_{\mathrm{class}} = \mathbb{E}[\sigma(\beta G_{\mathrm{q}})] + \mathbb{E}[(1 - 2\sigma(\beta G_{\mathrm{q}}))\phi]$$

where again

$$\phi = \Phi(sG_{\mathrm{q}}), \qquad s = \frac{a_1\mathcal{C}_\beta}{\sqrt{\frac{a_2}{2\alpha_{\mathrm{test}}} + (a_2 - a_1^2)\mathcal{C}_\beta^2}}.$$

Thus ϕ is a decreasing function in s for $G < 0$ and increasing in s for $G > 0$. We see that s is an increasing function of α_{test}, and so ϕ is decreasing in α_{test} for $G < 0$ and increasing in α_{test} for $G > 0$. These cases exactly correspond to $1 - 2\sigma(\beta G_{\mathrm{q}}) > 0$ and $1 - 2\sigma(\beta G_{\mathrm{q}}) < 0$, and so $\mathcal{E}_{\mathrm{class}}$ is always a decreasing function of α_{test}.