HiLD 2026: 4th Workshop on High-dimensional Learning Dynamics 1–30

Sequential Correlations Change In-Context Learning: Effective Context Length and Architectural Mismatch

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

Abstract

Modern sequence models have a striking capacity for in-context learning (ICL); they can perform new tasks based only on examples given in the prompt. Understanding how this ability emerges requires theory that capture important properties of natural data. Linear regression has served as a useful sandbox for ICL theory, but existing work has largely focused on prompts with independent examples. In this work, we extend this setting to sequentially correlated data, a basic feature of real sequences, present a solvable model based on linear attention, and test our predictions on realistic transformer architectures. We identify two distinct effects: First, when the query token is independent of the context, within-context correlations induce an effective context length: correlated prompts behave like shorter i.i.d. prompts. Second, when the query is also correlated with its context, test error is reduced, particularly for softmax attention when compared to linear attention. These results suggest that correlated prompts alter not only the effective sample size of in-context learning, but also which attention architectures are best matched to the task.

1. Introduction

In-context learning (ICL) is an important and useful capability of modern sequence models (Brown et al., 2020; Von Oswald et al., 2023; Wei et al., 2022). In this setting, a model performs a task implicitly from few examples without parameter updates. Regression has been a useful toy setting for studying how sequence models can achieve this ability. In-context regression requires a model to predict the label of a final query based on a prompt of input-output examples. The appeal of this setup is that it is simple enough to admit explicit analyses, yet rich enough to distinguish between architectures and learning mechanisms (Akyürek et al., 2023; Letey et al., 2026; Lu et al., 2025; Oko et al., 2024; Von Oswald et al., 2023; Zhang et al., 2024b, 2023).

A standard simplifying assumption in the in-context regression literature is that the examples within a prompt are independent. This assumption is analytically convenient, but not reflective of real datasets. This has prompted the study of sequential correlations in other non-regression in-context data settings, considering for example Markov chains and dynamical systems (Cole et al., 2025; Edelman et al., 2024; Nichani et al., 2024; Park et al., 2025); for the regression setting, sequential correlations have yet to be analysed. We know that sequential correlations matter even for classical ridge regression, where they change the asymptotic risk and invalidate estimators inherited from the i.i.d. setting (Atanasov et al., 2025). Thus it is natural to ask what is the effect of sequential correlations in the

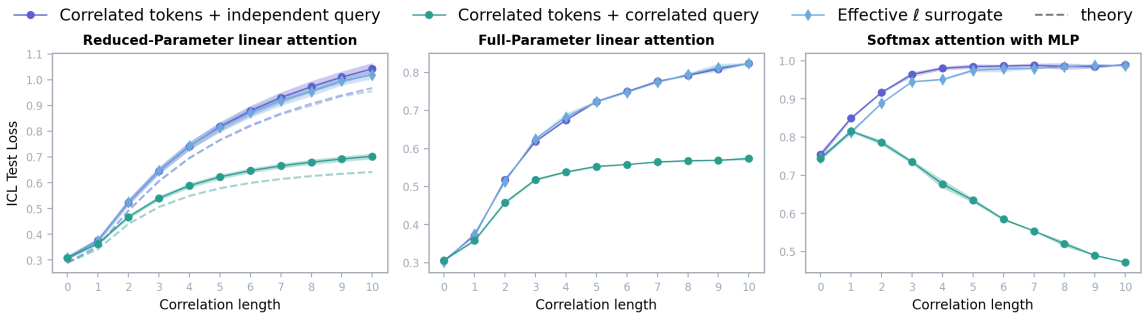

Figure 1: ICL test loss against token correlation strength for three attention architectures. We highlight two findings. (1) ICL performance on correlated data with an independent query (purple) is well approximated by uncorrelated data at an effective context length (blue) given by eq (1). (2) When the query token is equivalently correlated with its corresponding context (green), ICL error is reduced in all models as they utilise this correlation for inference. These predictions are made from our theory curves (dashed) derived for reduced-parameter linear attention. Experimental details given in Appendix A.

in-context regression setting, particularly as attention architectures are designed to operate on sequences in the first place.

In this work, we study a solvable model of in-context linear regression with sequentially correlated tokens. We begin from a reduced linear-attention model analysed in prior works (Bordelon et al., 2025; Letey et al., 2026; Lu et al., 2025; Wu et al., 2024; Zhang et al., 2024a), and perturb the usual i.i.d. setup by introducing a sequence correlation kernel for the context tokens, together with an optional correlation between the test query and the preceding context. This gives a controlled setting in which we can separate the effects of correlations *within* the context, and correlations *between* the query and the context.

Figure 1 summarises our findings. First, when correlations are confined to the context and the query remains independent, their effect is equivalent to an effective context-length reduction: correlated prompts behave like shorter i.i.d. prompts. Second, when the query token is correlated with its preceding context, ICL error decreases substantially, because the model can exploit this extra statistical structure to improve inference. Third, these correlated settings reveal an architectural mismatch: when additional query-dependent statistics enter, the performance gap between attention architectures (linear vs softmax in Figure 1) widens sharply.

An additional technical contribution of this work is distinguishing two cases of the asymptotic theory. Depending on the strength of the correlations (*i.e.*, the "length" over which they persist) relative to the sequence length, the high-dimensional treatment, as originally done by Lu et al. (2025), may fail: for correlations that persist on the same scale as the context length, moments of the reduced-linear attention estimator do not concentrate. We discuss this briefly in the main document and more thoroughly in Appendix D, and leave implications of this analysis to future work.

2. Theory and Data Setup

In-context linear regression. We study a correlated-token version of the solvable linear-regression ICL model induced by a reduced linear-attention block, as studied by Lu et al. (2025); Zhang et al. (2023). A single context consists of a sequence $\{\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_\ell, y_\ell, \boldsymbol{x}_{\text{test}}\}$ of examples, where

$$y_a = \boldsymbol{x}_a^\top \boldsymbol{w} + \varepsilon_a, \qquad y_{\text{test}} = \boldsymbol{x}_{\text{test}}^\top \boldsymbol{w} + \varepsilon_{\text{test}},$$

with task vector $\boldsymbol{w} \in \mathbb{R}^d$, and independent Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \rho I_\ell)$, $\varepsilon_{\text{test}} \sim \mathcal{N}(0, \rho)$. The goal is for the model's output on this sequence to be close to y_{test}, thus performing the correct regression inference given examples in the context.

Sequential correlations. We model sequential structure within the context as linear dependence between the tokens, following previous works by Atanasov et al. (2025); Moniri and Hassani (2025) in the ridge regression setting. Under this model, the mean-$\mathbf{0}$ Gaussian tokens $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell$ have second moments given by

$$\mathbb{E}[\boldsymbol{x}_a \boldsymbol{x}_b^\top] = K_{ab} I_d / d,$$

i.e., $K \in \mathbb{R}^{\ell \times \ell}$ controls correlations across context positions, and the feature-feature covariance is given by I_d/d. The case of independently-sampled tokens is described by $K = I_\ell$.

We will consider both cases where the mean-$\mathbf{0}$ test query $\boldsymbol{x}_{\text{test}}$ is independent of X and correlated with X. We define the query correlations as above through the second moment

$$\mathbb{E}[\boldsymbol{x}_{\text{test}} \boldsymbol{x}_a^\top] = \boldsymbol{k}_{\text{test}\,a} I_d / d$$

where here $\boldsymbol{k}_{\text{test}} \in \mathbb{R}^\ell$ specifies how strongly the query is coupled to each preceding context token; for an independent query, $\boldsymbol{k}_{\text{test}} = \mathbf{0}$.

Correlation summaries and correlation length. Without loss of generality, we take $\operatorname{tr}(K) = \ell$. The remaining correlation-dependent scalars that appear in the theory are

$$k_0 := \boldsymbol{k}_{\text{test}}^\top \boldsymbol{k}_{\text{test}}, \qquad k_1 := \boldsymbol{k}_{\text{test}}^\top K \boldsymbol{k}_{\text{test}}, \qquad k_2 := \operatorname{tr}\big(K^2\big)/\ell.$$

These quantities act as effective summaries of the token correlations seen by the ICL estimator, arising from the structure of the covariates H.

For general K and $\boldsymbol{k}_{\text{test}}$, we can think of k_0, k_1, k_2 as measures of token correlation strength. The quantity k_2 is a bulk measure of context correlations: under our chosen normalisation $\operatorname{tr}(K) = \ell$, it is minimised by the uncorrelated case $K = I_\ell$, where $k_2 = 1$, and increases as the spectrum of K becomes more anisotropic or concentrated. The query-dependent quantity k_0 measures the overall strength of the direct query-context coupling, while the mixed quantity k_1 measures how strongly that query-correlation profile is amplified by the correlated structure of the context itself.

These quantities become more intuitive if we take a specific kernel as an example: the exponential kernel $K_{ab} = \exp(-|a-b|/\xi)$ and $\boldsymbol{k}_{\text{test}\,a} = \exp(-(\ell+1-a)/\xi)$. Because K and $\boldsymbol{k}_{\text{test}}$ only depend on *distances* between tokens, it makes sense to think of correlation "strength" in terms of correlation *length*; here, the correlation length is explicitly ξ. This case provides an intuitive sandbox, as the ICL summary statistics above are directly related to the correlation length by $k_2, k_0 \approx \xi$, $k_1 \approx \xi^2$. Given this clear scaling, we choose to use the exponential kernel for the experiments in Figures 1 and 2. We save detailed interpretation of the summary statistics k_0, k_1, k_2 for non-exponential kernels to future work.

High-dimensional scaling. As standard for theoretical analysis of linear regression, we will work in a proportional regime of context length, *i.e.*, $d \to \infty$ with $\ell \propto d$. We assume the task signal remains identifiable, e.g. $\boldsymbol{w}^\top \boldsymbol{w} = \Theta(d)$.

We will analytically study the case of *weak correlations* with respect to context length ℓ, namely, for the computation of theory curves we will take $k_2, k_0, k_1 = \Theta_\ell(1)$. The case of *strong correlations* corresponds to $k_2, k_0 = \Theta(\ell), k_1 = \Theta(\ell^2)$. This case does not lend itself to an analytical solution in our asymptotic limit as the ICL error does not concentrate. Our theoretical results will thus focus on the former "weak" correlation case, with the "strong" correlation case discussed briefly in Section 3 and Appendices C and D.

Reduced linear-attention predictor. Following prior work on linear attention, the next-token prediction for y_{test} made by linear attention can be well-approximated by

$$\hat{y}_{\text{test}} := \text{tr}\Big(\Gamma H^\top\Big), \quad \Gamma \in \mathbb{R}^{d\times(d+1)}, \quad H := \boldsymbol{x}_{\text{test}} \begin{bmatrix} \frac{d}{\ell}(X\boldsymbol{w}+\boldsymbol{\varepsilon})^\top X & \frac{1}{\ell}(X\boldsymbol{w}+\boldsymbol{\varepsilon})^\top(X\boldsymbol{w}+\boldsymbol{\varepsilon}) \end{bmatrix}.$$

The parameter matrix Γ here is defined in terms of components of the value, key, and query matrices from full-parameter linear attention. This is the predictor that we will study in the theory, and is the "reduced model" referred to on the left-most panel of Figure 1.

Given a training batch of n sequences, corresponding to n such data-matrices H for the reduced model, the optimal parameters and final test loss can be written as

$$\Gamma^* = \left(\frac{n}{d}\lambda I_{d(d+1)} + \sum_{\mu=1}^{n} H^\mu \otimes H^\mu\right)^{-1} \sum_{\mu=1}^{n} y_{\text{test}}^\mu H^\mu, \quad \mathcal{E}_{\text{ICL}}(\Gamma^*) = \mathbb{E}[(y_{\text{test}}^{\text{new}} - \langle H^{\text{new}}, \Gamma^*\rangle)^2].$$

3. Effective in-context sample size

Working in the proportional limit described above, we can analyse the random parameter matrix Γ^*. The full result is given in Appendix E. We present its first implication.

Result: Effective-Sample Size

For prompts with weakly-correlated tokens $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\ell$ and an *independent* query $\boldsymbol{x}_{\text{test}}$, the correlated-ICL performance at context length ℓ is equivalent to **uncorrelated**-ICL performance at lower context length given by

$$\ell_{\text{eff}} := \frac{1+\rho}{k_2+\rho}\ell, \tag{1}$$

where by Jensen's inequality $k_2 = \text{tr}\big(K^2\big)/\ell \geq (\text{tr}(K)/\ell)^2 = 1$, so $\ell_{\text{eff}} \leq \ell$.

Eq (1) makes precise how sequential correlations reduce the usable information content of the prompt. Since k_2 increases as the context covariance becomes more anisotropic, stronger or more persistent correlations decrease the effective context length ℓ_{eff}. For stationary kernels, where k_2 grows with the correlation scale, this recovers the intuitive statement that longer-ranged correlations make fewer context examples effectively independent. The simplicity of the resulting formula is itself notable: the reduction depends linearly on the single scalar k_2, rather than on the full spectrum of K. This contrasts with classical correlated linear regression, where test error typically depends more delicately on the spectrum and eigenvectors of the design covariance.

Eq (1) also clarifies why the strongly-correlated regime is analytically difficult. Our theory here assumes that the correlation summary k_2 remains $\Theta(1)$ as $\ell, d \to \infty$ with $\ell \propto d$. But once correlations persist across more and more of the context, we instead have $k_2 = \Theta(\ell)$. In this regime, ℓ_{eff} no longer scales proportionally with ℓ, but remains only $\Theta(1)$. Thus, even as the nominal context length increases, the model effectively sees only finitely many independent examples. This is precisely the regime in which the standard high-dimensional concentration underlying our asymptotic analysis breaks down, as discussed more thoroughly in Appendix D.

4. Gain from query correlations

We now turn to the second effect in Figure 1: correlations between the test query and its preceding context. To isolate this contribution, define the *query-correlation gain* as

$$\Delta_{\text{query}} := e_{\text{ICL}}(\ell, k_2\,;\, \boldsymbol{k}_{\text{test}} = 0) - e_{\text{ICL}}(\ell, k_2, k_0, k_1),$$

namely, the reduction in ICL error obtained by introducing query-context correlations while holding the bulk context kernel K fixed.

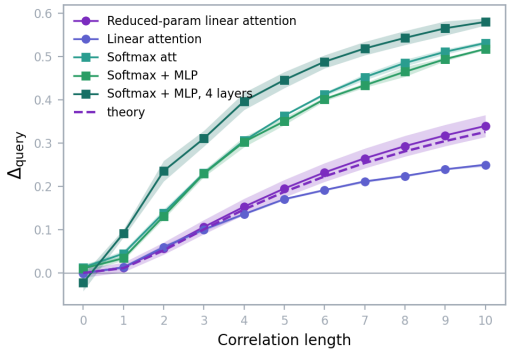

Figure 2: Correlated-query effects depend on architecture.

In Figure 2, this gain is almost always nonnegative, and vanishes only in the i.i.d. case. Thus, unlike bulk context correlations which only reduce effective sample size and hurt ICL performance, query correlations provide additional predictive signal that can in principle be exploited for inference.

In our theory, this effect is controlled not only by the bulk statistic k_2, but also by the query-dependent summaries k_0 and k_1, which increase with stronger correlations. The resulting closed-form expression for Δ_{query} is substantially more complicated than the effective-sample law in eq. (1), and we therefore omit it from the main text. That said, we see that Δ_{query} is relatively modest for linear attention architectures, but much larger for architectures that use softmax attention, indicating that the latter is better able to exploit informative query-context dependence.

5. Conclusion

Prompt correlations change in-context learning in two distinct ways: bulk context correlations reduce effective sample size, while query-context correlations provide useful signal for inference. In correlated linear regression, the first effect is captured by a simple effective-context-length law, while the second reveals a sharp architectural mismatch, with softmax attention benefiting much more than linear attention. These results suggest that understanding ICL on realistic structured data requires going beyond i.i.d. prompt models.

References

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.

Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Risk and cross validation in ridge regression with correlated samples. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=GMwKpJ9TiR.

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.

Frank Cole, Yulong Lu, Tianhao Zhang, and Yuxuan Zhao. In-context learning of linear dynamical systems with transformers: Error bounds and depth-separation, 2025. URL https://arxiv.org/abs/2502.08136.

Benjamin L. Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. The evolution of statistical induction heads: In-context learning markov chains, 2024. URL https://arxiv.org/abs/2402.11004.

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*, 2026. 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.

Behrad Moniri and Hamed Hassani. Asymptotics of linear regression with linearly dependent data. In Necmiye Ozay, Laura Balzano, Dimitra Panagou, and Alessandro Abate, editors, *Proceedings of the 7th Annual Learning for Dynamics & Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pages 72–85. PMLR, 04–06 Jun 2025. URL https://proceedings.mlr.press/v283/moniri25a.html.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of*

Machine Learning Research, pages 38018–38070. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/nichani24a.html.

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.

Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning, 2025. URL https://arxiv.org/abs/2412.01003.

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, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 14228–14246. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/2e10b2c2e1aa4f8083c37dfe269873f8-Paper-Conference.pdf.

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, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 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.

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

Ruiqi Zhang, Jingfeng Wu, and Peter L. Bartlett. In-context learning of a linear transformer block: Benefits of the mlp component and one-step gd initialization, 2024b.

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.

Appendix A. Experimental details

The data structure follows the Section 2 setup. For all experiments we choose a stationary kernel, *i.e.*, distance-dependent correlations, given by

$$K_{ab} := \exp(-|a-b|/\xi). \tag{2}$$

Here the correlation length ξ is manifestly part of the definition of K. This definition is convenient to choose due to the clarity of its correlation length parameter, as well as the fact that our Gaussian model with this kernel is equivalent to an AR(1) process which allows for efficient sampling. When query correlations are present, *i.e.*, $\boldsymbol{k}_{\text{test}} \neq \mathbf{0}$, we choose matching structure

$$(\boldsymbol{k}_{\text{test}})_a := \exp(-|\ell+1-a|/\xi). \tag{3}$$

This parameter ξ is what is referred to on the x-axes of Figures 1 and 2.

Data parameters. A batch of data contains sequences of length ℓ with one query each. We train in an offline manner with n total sequences. Each sequence is defined by a task vector $\boldsymbol{w}$. As in Letey et al. (2026); Lu et al. (2025); Raventós et al. (2023), the total number of task vectors in the pretraining batch may differ from n. The total number of unique task vectors in the batch is called k, and each unique vector is sampled i.i.d. from $\mathcal{N}(\mathbf{0}, I_d)$. For all experiments, we take

$$d = 32\,, \quad n = 4096 = 4d^2\,, \quad \ell = 128 = 4d\,, \quad k = 320 = 10d$$

matching the asymptotic scalings of these variables derived in Lu et al. (2025). We take label noise $\rho = 0.01$.

Architecture details. The experiments train sequence models on the above linear regression sequence data with sequentially correlated inputs. The architectures considered include softmax-only layers, softmax+MLP transformer-like architectures, and pure linear attention. Note that we distinguish between the reduced model (where the optimal solution is given directly by Γ^*) from *trainable* fully-parameterised linear attention. Softmax models embed inputs into width $d+1$ with a dedicated label channel, initialise input embeddings with scale proportional to $\sqrt{(d+1)}$, initialise query/key maps on the input channels, and initialise value/output maps to primarily use the label channel with small Gaussian noise. MLP blocks use GELU and are initialised close to zero. Linear-attention models operate directly on concatenated $(\boldsymbol{x}_i, y_i)$ tokens of dimension d+1, with query/key/value matrices initialised near identity. Models are trained by minimising mean-squared error on the held-out query label only. Optimisation uses Optax AdamW-style updates with weight decay `lamb`; the total number of gradient steps is set to T (something large enough to reach training loss convergence), but we employ early stopping and report the test loss of the best checkpointed-model. We use a linear learning-rate warmup from 0 to `max_lr` over the first 10% of training, with learning rate remaining constant after warmup. We use exponential moving average weights for evaluation. Minibatching is used to divide the full dataset of n sequences; we find that smaller batch sizes perform better for softmax attention architectures. Figures 1 and 2 show mean best-test loss over different seeds, where seeds control train data batch, testing samples, and batch divisions. A summary of our training hyperparameters is

$$\texttt{max\_lr} = 0.001, \quad \texttt{lamb} = 0.0001, \quad T = 10000, \quad \text{batch size} = 64 = 2d, \quad \#\text{ seeds} = 5.$$

Appendix B. Detailed setup

Here we re-iterate the data and model setup given in the main document with more details fleshed out. We leave the token design matrix general, *i.e.*,

$$\boldsymbol{x}_a \sim_K \mathcal{N}(0, \Sigma/d)$$

where the sample-sample correlations are given by positive semi-definite K as

$$\mathbb{E}_X[X_{ai}X_{bj}] = \frac{1}{d}K_{ab}\Sigma_{ij} \qquad \text{for } a,b \in [\ell] \text{ and } i,j \in [d]. \tag{4}$$

The label noise we take to be sequentially uncorrelated

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \rho I_{\ell+1}) \qquad \text{independent of all } X \text{ and } \boldsymbol{x}_{\text{test}}.$$

The query $\boldsymbol{x}_{\text{test}}$ may be correlated with the contexts as

$$\mathbb{E}[\boldsymbol{x}_{\text{test}}|X] = X^\top \boldsymbol{m}\,, \qquad \text{Var}(\boldsymbol{x}_{\text{test}}|X) = \frac{1}{d}(1-k_{-1})\Sigma$$

for

$$\boldsymbol{m} \equiv K^{-1}\boldsymbol{k}_{\text{test}} \in \mathbb{R}^\ell\,, \qquad k_{-1} \equiv \boldsymbol{k}_{\text{test}}^\top K^{-1}\boldsymbol{k}_{\text{test}} \leq 1.$$

This is equivalent to sampling, as in eq (4), using $\ell{+}1 \times \ell{+}1$ positive-semidefinite correlation kernel

$$K_{\text{query}} = \begin{bmatrix} K & \boldsymbol{k}_{\text{test}} \\ \boldsymbol{k}_{\text{test}}^\top & 1 \end{bmatrix}.$$

As a final note, this entire data setup is equivalent to

$$X = \frac{1}{\sqrt{d}}\sqrt{K}Z\sqrt{\Sigma}\,, \qquad Z_{si} \sim_{\text{i.i.d.}} \mathcal{N}(0,1).$$

B.1. Notation and Assumptions

We will consider a standard high-dimensional proportional limit, as in (Letey et al., 2026; Lu et al., 2025). The context length scales as

$$\alpha \equiv \frac{\ell}{d} = \Theta_{d,\ell}(1).$$

The usual high dimensional assumptions on Σ are such that

$$\text{tr}(\Sigma),\, \text{tr}\Big(\Sigma \boldsymbol{w}\boldsymbol{w}^\top\Big) = \Theta(d).$$

These are a succinct summary of conditions on both Σ and $\boldsymbol{w}$. The first condition $\text{tr}(\Sigma) = \Theta(d)$ ensures that there is reasonable signal to estimate in the tokens; note that after this section we will always assume $\Sigma = I_d$. The second condition $\text{tr}\big(\Sigma \boldsymbol{w}\boldsymbol{w}^\top\big) = \Theta(d)$ is equivalent to $\|\boldsymbol{w}\|^2 = \Theta(d)$ ensuring that there is enough recoverable task signal. Given these assumptions, everything will be written in normalised $\Theta_d(1)$ quantities

$$\text{tr}[\Sigma] := \frac{1}{d}\,\text{tr}(\Sigma)\,, \qquad \text{tr}\Big[\Sigma \boldsymbol{w}\boldsymbol{w}^\top\Big] := \frac{1}{d}\,\text{tr}\Big(\Sigma \boldsymbol{w}\boldsymbol{w}^\top\Big).$$

These need to be extended to assumptions about K. We will take

$$\text{tr}[K] = \frac{1}{\ell}\,\text{tr}(K) = 1$$

The final thing to reason about is the contributions from K^2 and from $\boldsymbol{k}_{\text{test}}$. This is where the notion of "correlation length" captured in K and $\boldsymbol{k}_{\text{test}}$ becomes important, as the terms

$$\text{tr}\big(K^2\big)\,, \quad k_{-1} \equiv \boldsymbol{k}_{\text{test}}^\top K^{-1}\boldsymbol{k}_{\text{test}}\,, \qquad k_0 \equiv \boldsymbol{k}_{\text{test}}^\top \boldsymbol{k}_{\text{test}}\,, \qquad k_1 \equiv \boldsymbol{k}_{\text{test}}^\top K \boldsymbol{k}_{\text{test}}$$

will appear in the computation.

Working definition of correlation length Suppose we choose K to be PSD and Toeplitz, *i.e.*,

$$K_{st} = c(|s-t|).$$

This means the correlations between $\boldsymbol{x}_s$ and $\boldsymbol{x}_t$ are stationary: they only depend on their relative distance in the sequence. A decent starting definition for the "correlation length" defined by c would be the sum of one of its columns, or the sum over all the lags, *i.e.*,

$$\hat{\xi} = \sum_{\tau=1}^{\ell-1} c(\tau).$$

Example: the exponential kernel Here take $c(\tau) = \exp(-\tau/\xi) \equiv \phi^\tau$. Then $\hat{\xi}$ recovers at least the correct scaling of the correlation length.

- Finite correlations: $\xi = \Theta_\ell(1)$. Then $\hat{\xi} \approx \xi$ (for ξ not too close to 0).
- Proportional correlations: $\xi = \beta\ell$. Then $\hat{\xi} \approx (1 - \exp(-1/\beta))\xi$.

We can compute k_{-1}, k_0, k_1 explicitly in each of these regimes, giving us a surrogate for how these quantities should scale for a general K or $c(\tau)$. Have

$$\frac{1}{\ell}\operatorname{tr}\big(K^2\big) \sim \xi\,, \quad k_{-1} \le 1\,, \quad k_0 \sim \xi\,, \quad k_1 \sim \xi^2$$

and

$$\frac{1}{\ell}\operatorname{tr}\big(K^2\big) \sim \beta\ell\,, \quad k_{-1} = 1\,, \quad k_0 \sim (1-\exp(-2/\beta))\frac{\beta}{2}\ell\,, \quad k_1 \sim \beta^2\ell^2.$$

Final assumptions Given this example, we will distinguish the asymptotic treatment between the two regimes

$$\text{Finite correlations:} \quad k_2 = \frac{1}{\ell}\operatorname{tr}(K^2), \quad k_0, \quad k_1 \quad = \Theta(1)$$

$$\text{Proportional correlations:} \quad \tilde{k}_2 = \frac{1}{\ell^2}\operatorname{tr}(K^2), \quad \tilde{k}_0 = \frac{k_0}{\ell}, \quad \tilde{k}_1 = \frac{k_1}{\ell^2} \quad = \Theta(1)$$

Appendix C. Population losses

We will be considering linear features

$$\boldsymbol{y} = X\boldsymbol{w} + \boldsymbol{\varepsilon}\,, \quad y_{\text{test}} = \boldsymbol{x}_{\text{test}}^\top \boldsymbol{w} + \mathcal{N}(0,\rho).$$

For each such sequence of $X, \boldsymbol{y}, \boldsymbol{x}_{\text{test}}, y_{\text{test}}$, define a data matrix

$$H = \boldsymbol{x}_{\text{test}}\begin{bmatrix}\frac{d}{\ell}(X\boldsymbol{w}+\boldsymbol{\varepsilon})^\top X & \frac{1}{\ell}(X\boldsymbol{w}+\boldsymbol{\varepsilon})^\top(X\boldsymbol{w}+\boldsymbol{\varepsilon})\end{bmatrix} \in \mathbb{R}^{d\times(d+1)}.$$

The predictor we will use for y_{test} is given by

$$\hat{y}_{\text{test}} = \langle \Gamma, H\rangle = \operatorname{tr}\Big(\Gamma H^\top\Big)$$

for parameters $\Gamma \in \mathbb{R}^{d\times(d+1)}$. Before we consider ICL error at optimal parameters Γ^*, we can compute a population formula for

$$\mathcal{E}_{\text{ICL}}(\Gamma) = \mathbb{E}_{\text{new data}}\Big[(y_{\text{test}}^{\text{new}} - \langle \Gamma, H^{\text{new}}\rangle)^2\Big].$$

This formula will be different depending on which asymptotic treatment of k_2, k_0, k_1 we are considering.

Notation. We will use the convention of row-wise vectorisation, and so have

$$\mathrm{Vec}(vu^\top) = v \otimes u \in \mathbb{R}^{\dim(v)\times\dim(u)}\,, \qquad \mathrm{Vec}(vu^\top)\mathrm{Vec}(vu^\top)^\top = (vv^\top)\otimes(\boldsymbol{uu}^\top)$$

Lemma 1 *Suppose we have weak-range correlations, i.e., $k_2, k_0, k_1 = \Theta(1)$. For $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, C)$, writing $\mathrm{tr}[C] = \mathrm{tr}(C)/d$, with*

$$\rho_1 \equiv \mathrm{tr}[C] + \rho\,, \qquad \rho_2 \equiv k_2\,\mathrm{tr}[C] + \rho$$

have ***population ICL risk***

$$\mathcal{E}_{\mathrm{ICL}}(\Gamma) = \rho + \mathrm{tr}[C] - 2\frac{1}{d}\,\mathrm{tr}\Big(\Gamma A^\top\Big) + \frac{1}{d}\,\mathrm{tr}\Big(\Gamma B \Gamma^\top\Big) + \frac{1}{d^2}\,\mathrm{tr}\Big(\mathcal{T}\,\mathrm{vec}(\Gamma)\,\mathrm{vec}(\Gamma)^\top\Big) \tag{5}$$

for

$$A \equiv \begin{bmatrix} C + \frac{1}{\alpha}k_0\,\mathrm{tr}[C] I_d & \mathbf{0} \end{bmatrix}\,, \qquad B \equiv \begin{bmatrix} C + \frac{\rho_2}{\alpha} I_d & \mathbf{0} \\ \mathbf{0} & \rho_1^2 \end{bmatrix}$$

$$\mathcal{T} \equiv \frac{k_1\,\mathrm{tr}[C] + \rho k_0}{\alpha^2}\,\mathrm{vec}\left(\begin{bmatrix} I & \mathbf{0} \end{bmatrix}\right)\mathrm{vec}\left(\begin{bmatrix} I & \mathbf{0} \end{bmatrix}\right)^\top + \frac{k_0}{\alpha}\Big(\mathrm{vec}\left(\begin{bmatrix} I & \mathbf{0} \end{bmatrix}\right)\mathrm{vec}\left(\begin{bmatrix} C & \mathbf{0} \end{bmatrix}\right)^\top + \mathrm{vec}\left(\begin{bmatrix} C & \mathbf{0} \end{bmatrix}\right)\mathrm{vec}\left(\begin{bmatrix} I & \mathbf{0} \end{bmatrix}\right)^\top\Big)$$

Lemma 2 *Suppose we have strong-range correlations, i.e., $k_0, k_2 = \Theta(\ell)$ and $k_1 = \Theta(\ell^2)$. Then writing $\Gamma = [\Gamma_{\mathrm{sq}} \quad \boldsymbol{\gamma}]$ we have*

$$\mathcal{E}_{\mathrm{ICL}}(\Gamma) = \rho + \mathrm{tr}[C] - 2\,\mathrm{tr}[C]\,\mathrm{tr}(\Gamma_{\mathrm{sq}}) + \mathrm{tr}[C]\tilde{k}_2\,\mathrm{tr}\Big(\Gamma_{\mathrm{sq}}\Gamma_{\mathrm{sq}}^\top\Big) + \mathrm{tr}[C]\tilde{k}_1\,\mathrm{tr}(\Gamma_{\mathrm{sq}}\Gamma_{\mathrm{sq}}) + \mathrm{tr}[C]\tilde{k}_1\,\mathrm{tr}(\Gamma_{\mathrm{sq}})^2$$

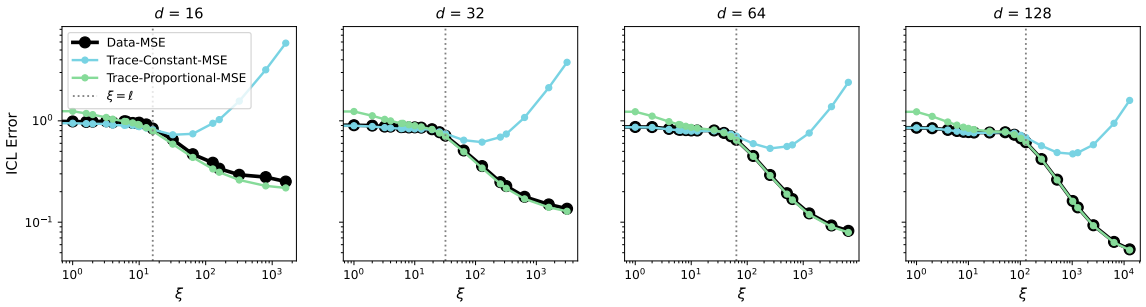

Figure 3: Illustration of transition between these two regimes. Quantities here are ICL MSE for Γ^* computed numerically from data, where the MSE is computed from (1) sampling the test distribution directly (2) using the population average in Lemma 1 (the weakly correlated case) and (3) using the population average in Lemma 2 (the strongly correlated case).

Figure 3 shows clearly the different regimes of correlation strength. Here we use the exponential kernel setting with correlated queries, *i.e.*, K and $\boldsymbol{k}_{\mathrm{test}}$ given by eq.s (2) and (3). We see the Lemma 1 formula tracks the data-sampled MSE for low correlation lengths, and likewise Lemma 2 for proportional correlation lengths, as predicted.

Appendix D. Non-concentration of finite-sample ICL error for strong correlations

We can immediately see in Figure 3 the lack of concentration that appears from the proportional correlation case: as d increases, the MSE losses for $\xi \propto \ell$ keeps decreasing, while the $\xi \ll \ell$ losses show minimal variation across d.

In section we will derive an asymptotically-precise formula for the finite-sample ICL error, i.e. $\mathcal{E}_{\mathrm{ICL}}(\Gamma^*)$ where Γ^* depends on a *finite* selection of training data sequences. Indeed $\mathcal{E}_{\mathrm{ICL}}(\Gamma^*)$ is a random quantity, and the fact that we can write down a deterministic equivalent for it hinges on its concentration in high dimensions. This quantity concentrates for the weak-correlations case, i.e. the case covered in the population analysis by Lemmas 1 and 4, where we have $k_0, k_1, k_2 = \Theta(1)$ as $\ell \to \infty$. Here we will explain why $\mathcal{E}_{\mathrm{ICL}}(\Gamma^*)$ does *not* concentrate in the strong-correlations case, where $k_0, k_2 \propto \ell$, $k_2 \propto \ell^2$.

The optimal parameters for our reduced linear attention model are given by

$$\mathrm{vec}(\Gamma^*) = \left(\frac{n}{d} \lambda I_{d(d+1)} + \sum_{\mu=1}^{n} \mathrm{vec}(H^\mu) \otimes \mathrm{vec}(H^\mu) \right)^{-1} \sum_{\mu=1}^{n} y_{\mathrm{test}}^\mu \, \mathrm{vec}(H^\mu). \tag{6}$$

We can see from this form that a necessary condition of $\mathcal{E}_{\mathrm{ICL}}$ will be the concentration of the data vectors $\mathrm{vec}(H^\mu)$. By concentration, we mean that its norm concentrates as $d, \ell \to \infty$ in the proportional limit. We will thus analyse the asymptotic behaviour of $\|\mathrm{vec}(H)\|$ where we are suppressing the index $\mu \in [n]$.

Recall that we have

$$H = \boldsymbol{x}_{\mathrm{test}} \begin{bmatrix} \frac{d}{\ell}(X\boldsymbol{w} + \boldsymbol{\varepsilon})^\top X & \frac{1}{\ell}(X\boldsymbol{w} + \boldsymbol{\varepsilon})^\top (X\boldsymbol{w} + \boldsymbol{\varepsilon}) \end{bmatrix}.$$

We will simplify our analysis here, as our goal is to focus on intuition, by ignoring the noise terms given by $\boldsymbol{\varepsilon}$; these do not affect the argument. Thus we write

$$\mathrm{vec}(H) = \boldsymbol{v} \otimes \boldsymbol{x}_{\mathrm{test}} + \text{ignored noise contribution}$$

where

$$\boldsymbol{v} = \begin{bmatrix} \frac{d}{\ell} X^\top X \boldsymbol{w} \\ \frac{1}{\ell} \boldsymbol{w}^\top X^\top X \boldsymbol{w} \end{bmatrix}.$$

We thus have that

$$\begin{aligned} \frac{1}{d^2} \|\mathrm{vec}(H)\|^2 &= \frac{1}{d^2} \|\boldsymbol{x}_{\mathrm{test}}\|^2 \|\boldsymbol{v}\|^2 \\ &= \frac{1}{d^2} \|\boldsymbol{x}_{\mathrm{test}}\|^2 \left(\frac{d^2}{\ell^2} \boldsymbol{w}^\top X^\top X X^\top X \boldsymbol{w} + \frac{1}{\ell^2} (\boldsymbol{w}^\top X^\top X \boldsymbol{w})^2 \right) \\ &\overset{d}{=} \left(\frac{1}{d} \boldsymbol{m}^\top K^{1/2} Z Z^\top K^{1/2} \boldsymbol{m} + 2 \frac{\sqrt{1 - k_{-1}}}{d} \boldsymbol{z}^\top Z^\top K^{1/2} \boldsymbol{m} + \frac{1 - k_{-1}}{d} \|\boldsymbol{m} z\|^2 \right) \\ &\quad \times \left(\frac{1}{d^2 \ell^2} \boldsymbol{w}^\top Z^\top K Z Z^\top K Z \boldsymbol{w} + \frac{1}{d^4 \ell^2} (\boldsymbol{w}^\top Z^\top K Z \boldsymbol{w})^2 \right), \end{aligned}$$

where $\overset{d}{=}$ means equivalent in distribution. Here Z is an $\ell \times d$ matrix with i.i.d. $\mathcal{N}(0,1)$ entries and $\boldsymbol{z} \sim \mathcal{N}(0, I_d)$ for the query (independent of Z).

Upon analysis of each of these terms, which we omit for brevity, we find that the problematic term is

$$\begin{aligned}\frac{1}{d^2\ell^2}\boldsymbol{w}^\top Z^\top KZZ^\top KZ\boldsymbol{w} &\overset{d}{=} \|\boldsymbol{w}\|^2 (Z^\top KZZ^\top KZ)_{11} \\ &\overset{d}{=} \frac{\|\boldsymbol{w}\|^2}{d}\frac{1}{\ell^2}\sum_{a,b=1}^{\ell}\kappa_a\kappa_b Z_{a1}Z_{b1}\frac{1}{d}\sum_{i=1}^{d}Z_{ai}Z_{bi} \\ &= \frac{\|\boldsymbol{w}\|^2}{d}\frac{1}{d}\left(\frac{1}{\ell}\sum_{a=1}^{\ell}\kappa_a Z_{a1}^2\right)^2 + \frac{\|\boldsymbol{w}\|^2}{d}\frac{1}{\ell^2}\sum_{a,b=1}^{\ell}\kappa_a\kappa_b Z_{a1}Z_{b1}\frac{1}{d}\sum_{i=2}^{d}Z_{ai}Z_{bi}.\end{aligned}$$

We have arrived at this expression by exploiting the fact that the distribution of the matrix Z is invariant under both left and right rotation to perform two changes of basis: (1) we let K have orthogonal eigendecomposition $K = O\,\mathrm{diag}(\kappa_1,\dots,\kappa_\ell)\,O^\top$, and (2) we choose a basis such that $\boldsymbol{w} = (\|\boldsymbol{w}\|, 0, \dots, 0)^\top$.

Now we must analyse this expression to provide a condition on K, specifically its eigenvalues $\kappa_1,\dots,\kappa_\ell$, for when this term does or does not concentrate. We focus on the second term in the sum

$$A := \frac{\|\boldsymbol{w}\|^2}{d}\frac{1}{\ell^2}\sum_{a,b=1}^{\ell}\kappa_a\kappa_b Z_{a1}Z_{b1}\frac{1}{d}\sum_{i=2}^{d}Z_{ai}Z_{bi}.$$

It has moments given by

$$\begin{aligned}\mathbb{E}[A] &= \frac{d-1}{d}\frac{1}{\ell^2}\sum_{a=1}^{\ell}\kappa_a^2\,, \\ \mathrm{var}[A] &= \frac{2}{\ell^4 d^2}(d-1)\left(\sum_{a=1}^{\ell}\kappa_a^2\right)^2 + \frac{2}{\ell^4 d^2}(d^2-1)\sum_{a=1}^{\ell}\kappa_a^4\,,\end{aligned}$$

and so the square of the coefficient of variation of A is therefore

$$\begin{aligned}\frac{\mathrm{var}[A]}{\mathbb{E}[A]^2} &= 2\left[\frac{1}{d-1} + \frac{d^2-1}{(d-1)^2}\frac{\sum_{a=1}^{\ell}\kappa_a^4}{\left(\sum_{a=1}^{\ell}\kappa_a^2\right)^2}\right] \\ &= 2[1+\mathcal{O}(d^{-1})]\frac{\sum_{a=1}^{\ell}\kappa_a^4}{\left(\sum_{a=1}^{\ell}\kappa_a^2\right)^2} + \mathcal{O}(d^{-1}).\end{aligned}$$

We now notice that the K-dependence here

$$\frac{\sum_{a=1}^{\ell}\kappa_a^4}{\left(\sum_{a=1}^{\ell}\kappa_a^2\right)^2} = \frac{\mathrm{tr}\big(K^4\big)}{\mathrm{tr}(K^2)^2}$$

looks like some sort of participation ratio. Indeed if $K = I_\ell$ (uncorrelated case, we know this concentrates) then

$$\frac{\mathrm{tr}\big(K^4\big)}{\mathrm{tr}(K^2)^2} = \frac{1}{\ell} \to 0,$$

while if $K = \mathbf{1}\mathbf{1}^\top$ (maximally correlated case, here $\boldsymbol{x}_1 = \ldots = \boldsymbol{x}_\ell$) then

$$\frac{\mathrm{tr}\big(K^4\big)}{\mathrm{tr}(K^2)^2} = 1.$$

This gives us an intuitive heuristic for concentration of $\|\,\mathrm{vec}(H)\|$.

Necessary condition for concentration

The random quantity $\mathcal{E}_{\mathrm{ICL}}(\Gamma^*)$, where Γ^* is determined by data sampled with correlation kernel K, concentrates to a deterministic high-dimensional limit only if

$$\frac{\mathrm{tr}\big(K^4\big)}{\mathrm{tr}(K^2)^2} \to 0. \tag{7}$$

We can estimate this in terms of "correlation length" ξ for the exponential case

$$K_{ab} = \exp(-|a-b|/\xi)$$

finding that

$$\frac{\mathrm{tr}\big(K^4\big)}{\mathrm{tr}(K^2)^2} \sim \frac{\xi}{\ell}$$

which gives the condition we expected. We see that for correlations that persist on the same order as the context length, we do not have concentration of $\mathrm{vec}(H)$, and thus we do not have concentration of $\mathcal{E}_{\mathrm{ICL}}(\Gamma^*)$. An extension of great interest to us would be to analyse this nonconcentration for a wider range of K choices, *e.g.* how does this participation ratio behave for other non-exponential (or even non-Toeplitz) choices of K?

Appendix E. Asymptotic formula for ICL error in the weak-correlations case

Here we present a deterministic formula for $\mathcal{E}_{\mathrm{ICL}}(\Gamma^*)$ in the weak-correlations case. Again, the optimal parameters are given by eq. (6). The explicit ridge λ is a parameter that adds regularisation to this solution; it is common to take the ridgeless limit $\lambda \to 0$ in similar works.

Here we have n sample sequences, giving n different $H^\mu \in \mathbb{R}^{d\times(d+1)}$ data embeddings, where the sequence length is of course $\ell + 1$ (ℓ main tokens and one query). We will also introduce a new parameter k that controls the number of regression task vectors $\boldsymbol{w}$ that define the labels in these sequences. The phenomenology surrounding k is not discussed in this work for space reasons, but we include it to best match previous work on this model for consistency (Letey et al., 2026; Lu et al., 2025; Raventós et al., 2023). We sample k task vectors

$$\boldsymbol{w}_j \sim_{\text{i.i.d.}} \mathcal{N}(\mathbf{0}, I_d)$$

and these k vectors are distributed evenly across the n sequences. This serves to limit the "task diversity" of the training batch, allowing us to judge how much of the true task distribution $\mathcal{N}(\mathbf{0}, I_d)$ the model is truly learning in-context.

We consider the same asymptotic treatment as derived in Lu et al. (2025), namely $d \to \infty$ with

$$\alpha := \frac{\ell}{d}\,, \quad \tau := \frac{n}{d^2}\,, \quad \kappa := \frac{k}{d}$$

all taken to be $\Theta(1)$ as $d \to \infty$.

An important quantity to define is the Stieltjes transform of the task sample covariance. This will be given by

$$\mathcal{M}_\kappa(z) := \lim_{d\to\infty, k=\kappa d} \frac{1}{d} \operatorname{tr}\left(\left(\frac{1}{k}\sum_{j=1}^{k} \boldsymbol{w}_j \boldsymbol{w}_j^\top + zI_d\right)^{-1}\right) \tag{8}$$

We will also use

$$\mathcal{M}'_\kappa(z) := \frac{\mathrm{d}}{\mathrm{d}z}\mathcal{M}_\kappa(z).$$

These quantities will appear throughout our formulas.

The explicit ridge parameter λ will be modulated by the fact that we have finite samples, and will appear in our formula through an *effective* ridge defined implicitly by

$$\tilde{\lambda}\mathcal{M}_\kappa\Big(\tilde{\lambda} + \frac{\rho_2}{\alpha}\Big) + \frac{\lambda\tau}{\tilde{\lambda}} = 1 - \tau$$

We will sometimes write $\mathcal{M} := \mathcal{M}_\kappa(\sigma), \quad \mathcal{M}' := \mathcal{M}'_\kappa(\sigma)$ for shorthand, where

$$\sigma = \tilde{\lambda} + \frac{\rho_2}{\alpha}, \quad \tilde{\sigma} = \sigma - \frac{k_0}{\alpha}.$$

Finally, we are only considering the case where the sequential correlations defined by $K, \boldsymbol{k}_{\text{test}}$ are sufficiently weak, *i.e.*,

$$k_2 := \frac{1}{\ell}\operatorname{tr}\big(K^2\big), \quad k_0 := \boldsymbol{k}_{\text{test}}^\top \boldsymbol{k}_{\text{test}}, \quad k_1 := \boldsymbol{k}_{\text{test}}^\top K \boldsymbol{k}_{\text{test}}$$

are all $\Theta(1)$ as $d, \ell \to \infty$. These terms will appears in the constants

$$\rho_1 \equiv 1 + \rho\,, \quad \rho_2 \equiv k_2 + \rho\,, \quad \phi_1 \equiv \frac{1}{\alpha^2}(k_1 + \rho k_0)\,, \quad \phi_2 \equiv \frac{1}{\alpha}k_0.$$

We present the main formula below. All proofs are given in subsequent sections.

Proposition 3 *Consider isotropic tokens and tasks, i.e., $\Sigma = I_d = C$, with sequential correlations in the tokens given by PSD $K \in \mathbb{R}^{\ell\times\ell}$ and $\boldsymbol{k}_{\text{test}} \in \mathbb{R}^\ell$ for $\boldsymbol{k}_{\text{test}}^\top K^{-1}\boldsymbol{k}_{\text{test}} \le 1$. We then have that ICL error for a model defined by Γ^* concentrates as $d \to \infty$ in the above limit, i.e.,*

$$\mathcal{E}_{\text{ICL}}(\Gamma^*) \simeq e_{\text{ICL}}^{\text{corr}}(\alpha, \kappa, \tau, \rho, k_0, k_1, k_2)$$

for

$$e_{\text{ICL}}^{\text{corr}}(\alpha, \kappa, \tau, \rho, k_0, k_1, k_2) = e_{\text{ICL}}^{\text{uncorr}}\left(\frac{1+\rho}{k_2+\rho}\alpha, \kappa, \tau, \rho\right) + e_{\text{query}}(\alpha, \kappa, \tau, \rho, k_0, k_1, k_2)$$

where

$$\begin{aligned} e_{\text{query}} = {} & -2\phi_2 + q_{\text{query}}(\mathcal{M} + \tilde{\lambda}\mathcal{M}') + \phi_2(2\sigma - \phi_2)\mathcal{M}' + 2\phi_2\tilde{\sigma}\mathcal{M} \\ & + \boldsymbol{m}_2^\top S^\top M\, S\boldsymbol{m}_2 - 2\boldsymbol{m}_2^\top S\boldsymbol{m}_3 + 2(1+\phi_2)\boldsymbol{m}_1^\top S\boldsymbol{m}_2 + (\phi_1 + 2\phi_2)\left(1 - \tilde{\sigma}\mathcal{M} - \boldsymbol{m}_1^\top S\boldsymbol{m}_2\right)^2 \\ & + \frac{\rho_2}{\alpha}\Big[q_{\text{query}}(\mathcal{M} + \tilde{\lambda}\mathcal{M}') + \phi_2(2\sigma - \phi_2)\mathcal{M}' + \boldsymbol{m}_2^\top S^\top M\, S\boldsymbol{m}_2 - 2\boldsymbol{m}_2^\top S\boldsymbol{m}_3 + 2\phi_2\mathcal{M}\Big]. \end{aligned}$$

for

$$\boldsymbol{m}_1 = \begin{bmatrix} \mathcal{M} \\ \phi_2(1 - \sigma\mathcal{M}) \end{bmatrix}, \quad \boldsymbol{m}_2 = \begin{bmatrix} 1 - \tilde{\sigma}\mathcal{M} \\ \phi_2(1 - \tilde{\sigma} + \sigma\tilde{\sigma}\mathcal{M}) \end{bmatrix}, \quad \boldsymbol{m}_3 = \begin{bmatrix} \mathcal{M} + \tilde{\sigma}\mathcal{M}' \\ \phi_2(1 - \sigma\mathcal{M} - \tilde{\sigma}\mathcal{M} - \sigma\tilde{\sigma}\mathcal{M}') \end{bmatrix}$$

$$M = \begin{bmatrix} -\mathcal{M}' & \phi_2(\mathcal{M} + \sigma\mathcal{M}') \\ \phi_2(\mathcal{M} + \sigma\mathcal{M}') & \phi_2^2(1 - 2\sigma\mathcal{M} - \sigma^2\mathcal{M}') \end{bmatrix}, \quad S = \begin{bmatrix} \mathcal{M} & 1 + \phi_2(1 - \sigma\mathcal{M}) \\ 1 + \phi_2(1 - \sigma\mathcal{M}) & -\phi_1 + \phi_2^2(1 - \sigma + \sigma^2\mathcal{M}) \end{bmatrix}^{-1}$$

$$q_{\text{query}} = \frac{1}{\tau - (1 - 2\tilde{\lambda}\mathcal{M} - \tilde{\lambda}^2\mathcal{M}')}((2\frac{k_0}{\alpha}\Big(\tilde{\lambda}(\mathcal{M} + \sigma\mathcal{M}') + (1 - 2\sigma\mathcal{M})\Big) + \frac{k_0^2}{\alpha^2}\Big(\mathcal{M} - \tilde{\lambda}\mathcal{M}'\Big) + \boldsymbol{m}_2^\top S\boldsymbol{m}_2 - \tilde{\lambda}\Big(-2\boldsymbol{m}_2^\top S\boldsymbol{m}_3 + \boldsymbol{m}_2^\top S^\top MS\boldsymbol{m}_2\Big))$$

Appendix F. Proofs of Population Averages

We begin for necessary y_{test}, H averages needed for terms in the population loss formulas in Lemmas 1 and 2. *A language model was used at various stages of these proofs to help with the Wick expansions of the 6th order moments in terms of traces. All expressions were subsequently carefully verified by the authors both analytically and numerically.*

Lemma 4 *Suppose we have $k_2, k_0, k_1 = \Theta(1)$ i.e., weak correlations. Write $\boldsymbol{v} = \Sigma\boldsymbol{w}$. Then we have*

$$\mathbb{E}[y_{\text{test}}^2] = \operatorname{tr}\Big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big] + \rho$$

$$\mathbb{E}[y_{\text{test}}H] \approx \frac{1}{d}\left[\boldsymbol{v}\boldsymbol{v}^\top + \tfrac{1}{\alpha}k_0\operatorname{tr}\big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\big]\Sigma \quad \big(\operatorname{tr}\big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\big] + \rho\big)\boldsymbol{v}\right]$$

$$\mathbb{E}[\operatorname{vec}(H)\operatorname{vec}(H)^\top] \approx \frac{1}{d}\Sigma \otimes \begin{bmatrix} \boldsymbol{v}\boldsymbol{v}^\top + \big(k_2\operatorname{tr}\big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\big] + \rho\big)\Sigma/\alpha & \big(\operatorname{tr}\big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\big] + \rho\big)\boldsymbol{v} \\ \big(\operatorname{tr}\big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\big] + \rho\big)\boldsymbol{v}^\top & \big(\operatorname{tr}\big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\big] + \rho\big))^2 \end{bmatrix} + \frac{1}{d^2}\mathcal{X}$$

for $d(d+1) \times d(d+1)$ tensor $\mathcal{X}$ given by

$$\begin{aligned}\mathcal{X} = \frac{1}{\alpha^2}\Big(k_1\operatorname{tr}[\Sigma W] + \rho k_0\Big)\operatorname{vec}\left(\begin{bmatrix}\Sigma & \mathbf{0}\end{bmatrix}\right)\operatorname{vec}\left(\begin{bmatrix}\Sigma & \mathbf{0}\end{bmatrix}\right)^\top \\ + \frac{k_0}{\alpha}\Big(\operatorname{vec}\left(\begin{bmatrix}\Sigma & \mathbf{0}\end{bmatrix}\right)\operatorname{vec}\left(\begin{bmatrix}S & \Big(\operatorname{tr}\Big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big] + \rho\Big)\boldsymbol{v}\end{bmatrix}\right)^\top \\ + \operatorname{vec}\left(\begin{bmatrix}S & \Big(\operatorname{tr}\Big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big] + \rho\Big)\boldsymbol{v}\end{bmatrix}\right)\operatorname{vec}\left(\begin{bmatrix}\Sigma & \mathbf{0}\end{bmatrix}\right)^\top\Big).\end{aligned}$$

The "$\approx$" means subleading terms in ℓ, d are neglected.

Proof Let's go term by term.

Label-label term. Conditioning on X, we have

$$\mathbb{E}[\boldsymbol{x}_{\text{test}}\boldsymbol{x}_{\text{test}}^\top \mid X] = \frac{1}{d}(1 - k_{-1})\Sigma + (X^\top\boldsymbol{m})(X^\top\boldsymbol{m})^\top$$

and so

$$\begin{aligned}\mathbb{E}[y_{\text{test}}^2] &= \mathbb{E}[\operatorname{tr}\Big(\boldsymbol{x}_{\text{test}}\boldsymbol{x}_{\text{test}}^\top\boldsymbol{w}\boldsymbol{w}^\top\Big) + \rho] \\ &= \frac{1}{d}(1 - k_{-1})\mathbb{E}_X\Big[\operatorname{tr}\Big(\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big)\Big] + \mathbb{E}_X\Big[\operatorname{tr}\Big((X^\top\boldsymbol{m})(X^\top\boldsymbol{m})^\top\boldsymbol{w}\boldsymbol{w}^\top\Big)\Big] + \rho \\ &= (1 - k_{-1} + k_{-1})\operatorname{tr}\Big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big] + \rho.\end{aligned}$$

Label-feature term. For

$$\boldsymbol{b} = X^\top(X\boldsymbol{w} + \boldsymbol{\varepsilon}), \quad c = (X\boldsymbol{w} + \boldsymbol{\varepsilon})^\top(X\boldsymbol{w} + \boldsymbol{\varepsilon})$$

we have

$$y_{\text{test}}H = \boldsymbol{x}_{\text{test}}(\boldsymbol{x}_{\text{test}}^\top\boldsymbol{w} + \varepsilon_{\text{test}})\left[\tfrac{d}{\ell}\boldsymbol{b}^\top \quad \tfrac{1}{\ell}c\right]$$

and so

$$\mathbb{E}[y_{\text{test}}H] = \frac{1}{d}(1 - k_{-1})\Sigma\boldsymbol{w}\begin{bmatrix}\frac{d}{\ell}\mathbb{E}_X[\boldsymbol{b}] \\ \frac{1}{\ell}\mathbb{E}_X[c]\end{bmatrix}^\top + \left[\tfrac{d}{\ell}\mathbb{E}_X[(X^\top\boldsymbol{m})(X^\top\boldsymbol{m})^\top\boldsymbol{w}\,\boldsymbol{b}^\top] \quad \tfrac{1}{\ell}\mathbb{E}_X[(X^\top\boldsymbol{m})(X^\top\boldsymbol{m})^\top\boldsymbol{w}\,c]\right]$$

Have

$$\begin{aligned}
\frac{d}{\ell}\mathbb{E}_X[\boldsymbol{b}] &= \operatorname{tr}[K]\boldsymbol{v} \\
\frac{1}{\ell}\mathbb{E}_X[c] &= \operatorname{tr}[K]\operatorname{tr}\Big[\Sigma \boldsymbol{w}\boldsymbol{w}^\top\Big] + \rho \\
\frac{d}{\ell}\mathbb{E}_X[(X^\top \boldsymbol{m})(X^\top \boldsymbol{m})^\top \boldsymbol{w}\boldsymbol{b}^\top] &= \frac{1}{d}\cdot\frac{1}{\ell}(k_{-1}\operatorname{tr}(K)+k_0)\boldsymbol{v}\boldsymbol{v}^\top + \frac{1}{d}\cdot\frac{1}{\ell}k_0\operatorname{tr}\Big(\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big)\Sigma \\
&\approx \frac{1}{d}\cdot\left(k_{-1}\operatorname{tr}[K]\boldsymbol{v}\boldsymbol{v}^\top + \frac{1}{\alpha}k_0\operatorname{tr}\Big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big]\Sigma\right) \\
\frac{1}{\ell}\mathbb{E}_X[(X^\top \boldsymbol{m})(X^\top \boldsymbol{m})^\top \boldsymbol{w}\, c] &= \frac{1}{d}\cdot\rho k_{-1}\boldsymbol{v} + \frac{1}{d}\cdot k_{-1}\operatorname{tr}[K]\operatorname{tr}\Big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big]\boldsymbol{v} + \frac{2}{d^2}\frac{k_0}{\ell}\boldsymbol{v}\boldsymbol{v}^\top\boldsymbol{w} \\
&\approx \frac{1}{d}\cdot k_{-1}\Big(\operatorname{tr}[K]\operatorname{tr}\Big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big] + \rho\Big)\boldsymbol{v}
\end{aligned}$$

Combining gives

$$\mathbb{E}[y_{\text{test}}H] = \frac{1}{d}\left[\operatorname{tr}[K]\boldsymbol{v}\boldsymbol{v}^\top + \tfrac{1}{\alpha}k_0\operatorname{tr}\big[\Sigma\boldsymbol{w}\boldsymbol{w}^\top\big]\Sigma \quad \rho_1\boldsymbol{v}\right]$$

Feature-feature term Using $\boldsymbol{b}, c$ as above we can write more easily

$$\operatorname{Vec}(H)\operatorname{Vec}(H)^\top = (\boldsymbol{x}_{\text{test}}\boldsymbol{x}_{\text{test}}^\top)\otimes\left(\begin{bmatrix}\frac{d}{\ell}\boldsymbol{b}\\ \frac{1}{\ell}c\end{bmatrix}\begin{bmatrix}\frac{d}{\ell}\boldsymbol{b}\\ \frac{1}{\ell}c\end{bmatrix}^\top\right)$$

Taking conditional expectation over $\boldsymbol{x}_{\text{test}}$ for fixed X, our expression simplifies as

$$\mathbb{E}\Big[\operatorname{Vec}(H)\operatorname{Vec}(H)^\top\Big] = \frac{1}{d}(1-k_{-1})\Sigma\otimes\mathbb{E}_{X,\epsilon}\Big[\boldsymbol{u}\boldsymbol{u}^\top\Big] + \mathbb{E}_{X,\epsilon}\Big[(X^\top\boldsymbol{m})(X^\top\boldsymbol{m})^\top\otimes(\boldsymbol{u}\boldsymbol{u}^\top)\Big].$$

where I'm writing $\boldsymbol{u}^\top = \begin{bmatrix}\frac{d}{\ell}\boldsymbol{b}^\top & \frac{1}{\ell}c\end{bmatrix}$ for convenience. Now let's start with the $\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\top]$ term.

We have

$$\begin{aligned}
\mathbb{E}[\boldsymbol{b}\boldsymbol{b}^\top] &= \mathbb{E}[X^\top X\boldsymbol{w}\boldsymbol{w}^\top X^\top X] + \mathbb{E}[X^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X] \\
&= \frac{1}{d^2}\Big(\operatorname{tr}\big(K^2\big)+\operatorname{tr}(K)^2\Big)\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Sigma + \frac{1}{d^2}\operatorname{tr}\big(K^2\big)\operatorname{tr}\Big(\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big)\Sigma + \frac{1}{d}\rho\operatorname{tr}(K)\Sigma \\
\mathbb{E}[c\boldsymbol{b}] &= \mathbb{E}[\boldsymbol{w}^\top X^\top X\boldsymbol{w}X^\top X\boldsymbol{w}] + 2\mathbb{E}[\boldsymbol{\varepsilon}^\top X w X^\top\boldsymbol{\varepsilon}] + \mathbb{E}[\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}X^\top X\boldsymbol{w}] \\
&= \left(\frac{1}{d^2}\Big(\operatorname{tr}(K)^2+2\operatorname{tr}\big(K^2\big)\Big)\operatorname{tr}\Big(\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big) + \frac{1}{d}\rho(\ell+2)\operatorname{tr}(K)\right)\Sigma\boldsymbol{w} \\
\mathbb{E}[c^2] &= \mathbb{E}[\boldsymbol{w}^\top X^\top X\boldsymbol{w}\boldsymbol{w}^\top X^\top X\boldsymbol{w}] + 2\mathbb{E}[\boldsymbol{w}^\top X^\top X\boldsymbol{w}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}] + 4\mathbb{E}[\boldsymbol{w}^\top X^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top X\boldsymbol{w}] + \mathbb{E}[\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{\varepsilon}] \\
&= \frac{1}{d^2}\Big(\operatorname{tr}(K)^2+2\operatorname{tr}\big(K^2\big)\Big)\operatorname{tr}\Big(\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big)^2 + \frac{1}{d}\rho(2\ell+4)\operatorname{tr}(K)\operatorname{tr}\Big(\Sigma\boldsymbol{w}\boldsymbol{w}^\top\Big) + \rho^2(\ell^2+2\ell).
\end{aligned}$$

This simplifies as

$$\begin{aligned}
\frac{d^2}{\ell^2}\mathbb{E}[\boldsymbol{b}\boldsymbol{b}^\top] &\approx \operatorname{tr}[K]^2(\Sigma\boldsymbol{w})(\Sigma\boldsymbol{w})^\top + \frac{1}{\alpha}\big(\rho\operatorname{tr}[K] + \operatorname{tr}\big[K^2\big]\operatorname{tr}[\Sigma W]\big)\Sigma \\
\frac{d}{\ell^2}\mathbb{E}[c\boldsymbol{b}] &\approx \operatorname{tr}[K](\operatorname{tr}[K]\operatorname{tr}[\Sigma W] + \rho)\Sigma\boldsymbol{w} \\
\frac{1}{\ell^2}\mathbb{E}[c^2] &\approx (\rho + \operatorname{tr}[K]\operatorname{tr}[\Sigma W])^2
\end{aligned}$$

upon making the above high-dimensional assumptions.

For the higher order term,

$$\mathbb{E}\left[(X^\top\boldsymbol{m})(X^\top\boldsymbol{m})^\top\otimes(\boldsymbol{u}\boldsymbol{u}^\top)\right] = \begin{bmatrix}\frac{d^2}{\ell^2}\mathbb{E}[(X^\top\boldsymbol{m})(X^\top\boldsymbol{m})^\top\otimes\boldsymbol{b}\boldsymbol{b}^T] & \frac{d}{\ell^2}\mathbb{E}[(X^\top\boldsymbol{m})(X^\top\boldsymbol{m})^\top\otimes c\boldsymbol{b}] \\ \frac{d}{\ell^2}\mathbb{E}[(X^\top\boldsymbol{m})(X^\top\boldsymbol{m})^\top\otimes c\boldsymbol{b}^T] & \frac{1}{\ell^2}\mathbb{E}[(X^\top\boldsymbol{m})(X^\top\boldsymbol{m})^\top\otimes c^2]\end{bmatrix}$$

Write

$$\boldsymbol{v} := \Sigma \boldsymbol{w}, \qquad S := \boldsymbol{v}\boldsymbol{v}^\top, \qquad k_{-1} \equiv \operatorname{tr}(KM) = \boldsymbol{k}_{\text{test}}^\top K^{-1} \boldsymbol{k}_{\text{test}}, \quad k_0 \equiv \boldsymbol{k}_{\text{test}}^\top \boldsymbol{k}_{\text{test}}, \quad k_1 \equiv \boldsymbol{k}_{\text{test}}^\top K \boldsymbol{k}_{\text{test}}$$

For large ℓ, d and the chosen high-dimensional assumptions, the Isserlis expansion gives

$$\frac{d^2}{\ell^2}\mathbb{E}\Big[\big((X^\top m)(X^\top m)^\top\big) \otimes (\boldsymbol{b}\boldsymbol{b}^\top)\Big]_{(I,i),(J,j)} \approx \frac{1}{d} k_{-1}\bigg(\Sigma_{IJ} S_{ij} \operatorname{tr}[K]^2 + \frac{1}{\alpha}\big(\operatorname{tr}[\Sigma W]\operatorname{tr}\big[K^2\big] + \rho \operatorname{tr}[K]\big)\Sigma_{IJ}\Sigma_{ij}\bigg) + \frac{1}{d^2}\mathcal{X}_{(I,i)(J,j)}$$

with

$$\begin{aligned}\mathcal{X}_{(I,i)(J,j)} &\equiv \frac{1}{\alpha^2} k_1 \operatorname{tr}[\Sigma W]\Sigma_{Ii}\Sigma_{Jj} + \frac{1}{\alpha}\operatorname{tr}[K] k_0 (\Sigma_{Ii} S_{Jj} + \Sigma_{Jj} S_{Ii}) + \rho \frac{1}{\alpha^2} k_0 \Sigma_{Ii}\Sigma_{Jj} \\ &= \frac{1}{\alpha^2}(k_1 \operatorname{tr}[\Sigma W] + \rho k_0)\Sigma_{Ii}\Sigma_{Jj} + \frac{1}{\alpha} k_0 \operatorname{tr}[K](\Sigma_{Ii} S_{Jj} + \Sigma_{Jj} S_{Ii}).\end{aligned}$$

Now for the second term,

$$\begin{aligned}\mathbb{E}[(X^\top \boldsymbol{m})(X^\top \boldsymbol{m})^\top \otimes c\boldsymbol{b}]_{(I,i),J} &= \mathbb{E}[c (X^\top \boldsymbol{m})_I (X^\top \boldsymbol{m})_J b_i] \\ &= \mathbb{E}[(X_{sk} w_k + \varepsilon_s)(X_{sn} w_n + \varepsilon_s) X_{tI} X_{pJ} M_{tp} X_{qi}(X_{qj} w_j + \varepsilon_q)] \\ &= M_{tp} W_{kn} w_j \mathbb{E}[X_{sk} X_{sn} X_{tI} X_{pJ} X_{qi} X_{qj}] + (\ell + 2)\rho M_{tp} w_j \mathbb{E}[X_{tI} X_{pJ} X_{qi} X_{qj}]\end{aligned}$$

Expanding all the Wick terms gives

$$\frac{d}{\ell^2}\mathbb{E}\Big[\big((X^\top m)(X^\top m)^\top\big) \otimes (cb)\Big]_{(I,i),J} \approx \frac{1}{d} k_{-1}(\operatorname{tr}[\Sigma W] + \rho)\Sigma_{IJ} v_i + \frac{1}{d^2}\mathcal{X}_{(I,i)(J,d+1)}$$

where

$$\mathcal{X}_{(I,i)(J,d+1)} \equiv \frac{1}{\alpha} k_0 (\operatorname{tr}[\Sigma W] + \rho)\Sigma_{Ii} v_J.$$

Finally, the last term gives

$$\frac{1}{\ell^2}\mathbb{E}\Big[((X^\top m)(X^\top m)) \otimes c^2\Big]_{IJ} \approx \frac{1}{d} k_{-1}(\operatorname{tr}[\Sigma W] + \rho)^2 \Sigma_{IJ}$$

Gathering everything together gives the required formula. ■

Lemma 5 *We have*

$$\mathbb{E}[y_{\text{test}}^2] = \operatorname{tr}[\Sigma W] + \rho \tag{9}$$

$$\mathbb{E}[y_{\text{test}} H] \approx \begin{bmatrix}\tilde{k}_0 \operatorname{tr}[\Sigma W]\Sigma & \mathbf{0}\end{bmatrix} \tag{10}$$

$$\mathbb{E}[\operatorname{vec}(H)\operatorname{vec}(H)^\top] \approx \begin{bmatrix}\operatorname{tr}[\Sigma W]\Big(\tilde{k}_2 \Sigma \otimes \Sigma + \tilde{k}_1 \Sigma \,\tilde{\otimes}\, \Sigma + \tilde{k}_1 \operatorname{vec}(\Sigma)\operatorname{vec}(\Sigma)^\top\Big) & \mathbf{0} \\ \mathbf{0} & 0\end{bmatrix} \tag{11}$$

where I'm using $\tilde{\otimes}$ to refer to the transposition+Kronecker operation, i.e., for $A, B \in \mathbb{R}^{d\times d}$, have

$$(A \otimes B)_{(I,i),(J,j)} = A_{IJ} B_{ij}, \qquad (A \tilde{\otimes} B)_{(I,i),(J,j)} = A_{Ij} B_{Ji}.$$

Proof Let's go term by term.

Label-label term. As before $\mathbb{E}[y_{\text{test}}^2] = \operatorname{tr}[\Sigma W] + \rho$.

Label-feature term. Have

$$\begin{aligned}
\frac{d}{\ell}\mathbb{E}_X[\boldsymbol{b}] &= \boldsymbol{v} \\
\frac{1}{\ell}\mathbb{E}_X[c] &= \operatorname{tr}[\Sigma W] + \rho \\
\frac{d}{\ell}\mathbb{E}_X[(X^\top \boldsymbol{m})(X^\top \boldsymbol{m})^\top \boldsymbol{w}\boldsymbol{b}^\top] &= \frac{1}{d}\Big(k_{-1} + \tilde{k}_0\Big) S + \tilde{k}_0 \operatorname{tr}[\Sigma W]\Sigma \\
\frac{1}{\ell}\mathbb{E}_X[(X^\top \boldsymbol{m})(X^\top \boldsymbol{m})^\top \boldsymbol{w}\, c] &= \frac{1}{d}\cdot k_{-1}(\operatorname{tr}[\Sigma W] + \rho)\boldsymbol{v} + \frac{2}{d^2}\tilde{k}_0 S\boldsymbol{w}
\end{aligned}$$

and so

$$\mathbb{E}[y_{\text{test}} H] \approx \begin{bmatrix} \tilde{k}_0 \operatorname{tr}[\Sigma W]\Sigma & \mathbf{0} \end{bmatrix}$$

Feature-feature term Again

$$\mathbb{E}\Big[\mathrm{Vec}(H)\mathrm{Vec}(H)^\top\Big] = \frac{1}{d}(1-k_{-1})\Sigma \otimes \mathbb{E}_{X,\epsilon}\Big[\boldsymbol{u}\boldsymbol{u}^\top\Big] + \mathbb{E}_{X,\epsilon}\Big[(X^\top \boldsymbol{m})(X^\top \boldsymbol{m})^\top \otimes (\boldsymbol{u}\boldsymbol{u}^\top)\Big].$$

where $\boldsymbol{u}^\top = \begin{bmatrix} \frac{d}{\ell}\boldsymbol{b}^\top & \frac{1}{\ell}c \end{bmatrix}$. The order-4 terms are

$$\begin{aligned}
\frac{d^2}{\ell^2}\mathbb{E}[\boldsymbol{b}\boldsymbol{b}^\top] &= (1+\tilde{k}_2)S + d\tilde{k}_2 \operatorname{tr}[\Sigma W]\Sigma + \frac{1}{\alpha}\rho\Sigma \\
\frac{d}{l^2}\mathbb{E}[c\boldsymbol{b}] &= \Big(\Big(1+2\tilde{k}_2\Big)\operatorname{tr}[\Sigma W] + \rho\Big)\boldsymbol{v} \\
\frac{1}{\ell^2}\mathbb{E}[c^2] &= \Big(1+2\tilde{k}_2\Big)\operatorname{tr}[\Sigma W]^2 + 2\rho\operatorname{tr}[\Sigma W] + \rho^2.
\end{aligned}$$

Expanding the order-6 terms, the first is

$$\begin{aligned}
\frac{d^2}{\ell^2}\,\mathbb{E}\Big[\big((X^\top m)(X^\top m)^\top\big)\otimes(\boldsymbol{b}\boldsymbol{b}^\top)\Big]_{(I,i),(J,j)} &= \frac{1}{d}\Sigma_{IJ}\,S_{ij}\,k_{-1}\,(1+\tilde{k}_2) \;+\; \Sigma_{IJ}\,\Sigma_{ij}\,k_{-1}\,\tilde{k}_2\operatorname{tr}[\Sigma W] \\
&+\; \tilde{k}_1\operatorname{tr}[\Sigma W]\Big(\Sigma_{Ii}\Sigma_{Jj} + \Sigma_{Ij}\Sigma_{Ji}\Big) \\
&+\; \frac{1}{d}(\tilde{k}_0+\tilde{k}_1)\Big(\Sigma_{Ii}S_{Jj} + \Sigma_{Ji}S_{Ij} + \Sigma_{Jj}S_{Ii} + \Sigma_{Ij}S_{Ji}\Big) \\
&+\; 2\frac{1}{d}\tilde{k}_1\,\Sigma_{ij}\,S_{IJ} \\
&+\; \frac{1}{d}\frac{1}{\alpha}\rho\Big(\Sigma_{IJ}\Sigma_{ij}k_{-1} + \tilde{k}_0(\Sigma_{Ii}\Sigma_{Jj} + \Sigma_{Ij}\Sigma_{Ji})\Big)
\end{aligned}$$

Now for the second term,

$$\begin{aligned}
\frac{d}{\ell^2}\mathbb{E}\Big[\big((X^\top m)(X^\top m)^\top\big)\otimes(c\boldsymbol{b})\Big]_{(I,i),J} &= \frac{1}{d}\cdot\operatorname{tr}[\Sigma W]\,\Sigma_{IJ}\,v_i\;k_{-1}(1+2\tilde{k}_2) \\
&+\; \frac{1}{d}\cdot\operatorname{tr}[\Sigma W]\big(\Sigma_{Ii}v_J + \Sigma_{Ji}v_I\big)\,(\tilde{k}_0 + 2\tilde{k}_1) \\
&+\; \frac{1}{d^2}(v_I v_J v_i)\,(2\tilde{k}_0 + 4\tilde{k}_1) \\
&+\; \frac{1}{d}\cdot\rho\Big(k_{-1}\Sigma_{IJ}\,v_i + \tilde{k}_0\big(\Sigma_{Ii}v_J + \Sigma_{Ji}v_I\big)\Big)
\end{aligned}$$

Finally, the last term

$$\frac{1}{\ell^2}\mathbb{E}\Big[((X^\top m)(X^\top m))\otimes c^2\Big]_{IJ} = \frac{1}{d}\cdot\Big[\Sigma_{IJ}\ \mathrm{tr}[\Sigma W]^2\, k_{-1}(1+2\tilde{k}_2)\ +\ \frac{1}{d}\,\mathrm{tr}[\Sigma W]\, v_I v_J(4\tilde{k}_0+8\tilde{k}_1)\Big]$$
$$+\frac{2}{d}\rho\cdot\Big[\Sigma_{IJ}\ \mathrm{tr}[\Sigma W]\, k_{-1}\ +\ \frac{2}{d}\tilde{k}_0\, v_I v_J\Big]\ +\ \frac{1}{d}\cdot\rho^2\,\Sigma_{IJ}\, k_{-1}.$$

Thus we have

$$\approx (1-k_{-1})\Sigma\otimes\begin{bmatrix}\tilde{k}_2\,\mathrm{tr}[\Sigma W]\Sigma & \mathbf{0}\\ \mathbf{0}^\top & 0\end{bmatrix}+\mathrm{tr}[\Sigma W]\begin{bmatrix}k_{-1}\tilde{k}_2\Sigma\otimes\Sigma+\tilde{k}_1\Sigma\tilde{\otimes}\Sigma+\tilde{k}_1\,\mathrm{vec}(\Sigma)\,\mathrm{vec}(\Sigma)^\top & \mathbf{0}\\ \mathbf{0} & 0\end{bmatrix}$$

and the result follows. ■

Appendix G. Proof of Proposition 3

This entire section is an extended calculation that serves as a complete proof of the deterministic formula for $\mathcal{E}_{\text{ICL}}(\Gamma^*)$ given in Proposition 3. The methodology in this section is to analyse the random object Γ^*, which contains randomness through the particular training batch $(H^\mu, y^\mu_{\text{test}})_{\mu=1}^n$, as a resolvent of a larger random matrix, and the necessary $\text{tr}\big(\Gamma A^\top\big)$ and $\text{tr}\big(\Gamma^\top B\Gamma\big)$ terms as traces of a resolvent against deterministic test matrices. Throughout this entire section we will take $\Sigma = I_d$.

This section will be relatively light on detail as the full calculation setup can be found in both Lu et al. (2025) and Letey et al. (2026). For a more rigorous analysis, with error terms bounded properly, see Lu et al. (2025); here we will simply write $\approx$ when terms can be dropped due to the negligibility in high dimensions.

We care about the MSE loss (computed by Theorem 1) for the optimal parameters

$$\text{vec}(\Gamma^*) = \left(\frac{n}{d}\lambda I + \sum_{\mu=1}^n \text{vec}(H^\mu)\text{vec}(H^\mu)^\top\right)^{-1} \sum_{\mu=1}^n y^\mu_{\text{test}}\text{vec}(H^\mu) \tag{12}$$

for $\mu \in [n]$ denoting the index for the n training sample sequences. Define

$$\boldsymbol{z}_\mu = \begin{bmatrix} y^\mu_{\text{test}}/d \\ \text{vec}(H^\mu)/\sqrt{d}\end{bmatrix} \in \mathbb{R}^{d(d+1)+1}$$

and construct extended resolvent

$$G_{\text{ext}}(\pi) = \frac{1}{\sum_{\mu\in[n]} \boldsymbol{z}_\mu\boldsymbol{z}_\mu^\top + \pi B_{\text{ext}} + \tau\lambda I}\,. \tag{13}$$

for some test matrix $B_{\text{ext}} \in \mathbb{R}^{(d(d+1)+1)\times(d(d+1)+1)}$.

It's important to note that even though we're studying sequentially-correlated data, the vectors $\boldsymbol{z}_\mu$s are still independent of each other. We can thus apply the "leave-one-out" or cavity method here over the μ index. We end up with

$$\sum_{\mu\in[n]} \frac{1}{1+\boldsymbol{z}_\mu^\top G^{[\mu]}_{\text{ext}}\boldsymbol{z}_\mu} G^{[\mu]}_{\text{ext}}\boldsymbol{z}_\mu\boldsymbol{z}_\mu^\top + G_{\text{ext}}(\pi B_{\text{ext}} + \tau\lambda I) = I. \tag{14}$$

For fixed task $\boldsymbol{w}_\mu$ we can apply Lemma 4 to compute

$$\mathbb{E}_{X,\boldsymbol{x}_{\text{test}},\boldsymbol{\epsilon}}[\boldsymbol{z}\boldsymbol{z}^\top] = \frac{1}{d^2}\Upsilon(\boldsymbol{w})$$

for

$$\Upsilon(\boldsymbol{w}) \equiv \begin{bmatrix} \text{tr}\big[\boldsymbol{w}\boldsymbol{w}^\top\big] + \rho & \frac{1}{\sqrt{d}}\,\text{vec}\left(\big[\boldsymbol{w}\boldsymbol{w}^\top + \frac{1}{\alpha}k_0\,\text{tr}\big[\boldsymbol{w}\boldsymbol{w}^\top\big]I_d \quad \rho_1(\boldsymbol{w})\boldsymbol{w}\big]\right) \\ \frac{1}{\sqrt{d}}\,\text{vec}\left(\big[\boldsymbol{w}\boldsymbol{w}^\top + \frac{1}{\alpha}k_0\,\text{tr}\big[\boldsymbol{w}\boldsymbol{w}^\top\big]I_d \quad \rho_1(\boldsymbol{w})\boldsymbol{w}\big]\right)^\top & I_d\otimes E(\boldsymbol{w}) + \frac{1}{d}\mathcal{X}(\boldsymbol{w}) \end{bmatrix}$$

where

$$E(\boldsymbol{w}) \equiv \begin{bmatrix} \boldsymbol{w}\boldsymbol{w}^\top + \rho_2(\boldsymbol{w})I_d/\alpha & \rho_1(\boldsymbol{w})\boldsymbol{w} \\ \rho_1(\boldsymbol{w})\boldsymbol{w}^\top & \rho_1(\boldsymbol{w})^2\end{bmatrix}, \quad \rho_1(\boldsymbol{w}) = \text{tr}\Big[\boldsymbol{w}\boldsymbol{w}^\top\Big] + \rho, \quad \rho_2(\boldsymbol{w}) = \text{tr}\Big[\boldsymbol{w}\boldsymbol{w}^\top\Big]k_2 + \rho$$

and $d(d+1) \times d(d+1)$ tensor

$$\begin{aligned}\mathcal{X}(\boldsymbol{w}) \equiv \frac{1}{\alpha^2}\Big(\mathrm{tr}\Big[\boldsymbol{w}\boldsymbol{w}^\top\Big]k_1 + \rho k_0\Big)\,\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)^\top \\ + \frac{1}{\alpha}k_0\Big(\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)\mathrm{vec}\left(\begin{bmatrix} \boldsymbol{w}\boldsymbol{w}^\top & \mathbf{0}\end{bmatrix}\right)^\top + \mathrm{vec}\left(\begin{bmatrix} \boldsymbol{w}\boldsymbol{w}^\top & \mathbf{0}\end{bmatrix}\right)\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)^\top\Big) \\ + \frac{1}{\alpha}\rho_1(\boldsymbol{w})k_0\Big(\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)\mathrm{vec}\Big(\boldsymbol{w}\boldsymbol{e}_{d+1}^\top\Big)^\top + \mathrm{vec}\Big(\boldsymbol{w}\boldsymbol{e}_{d+1}^\top\Big)\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)^\top\Big).\end{aligned}$$

The quadratic form $\boldsymbol{z}_\mu^\top G_{\mathrm{ext}}^{[\mu]}\boldsymbol{z}_\mu$ concentrates for fixed tasks $\boldsymbol{w}_\mu$ by standard arguments, so we have

$$\boldsymbol{z}_\mu^\top G_{\mathrm{ext}}^{[\mu]}\boldsymbol{z}_\mu \simeq \chi^\mu(\boldsymbol{w}_\mu) \tag{15}$$

where

$$\chi^\mu(\boldsymbol{w}_\mu) \equiv \frac{1}{d^2}\,\mathrm{tr}\left([G_{\mathrm{ext}}^\mu]_{\setminus 0}\cdot\left[I\otimes E(\boldsymbol{w}_\mu) + \frac{1}{d}\mathcal{X}(\boldsymbol{w}_\mu)\right]\right). \tag{16}$$

Replacing $\boldsymbol{z}_\mu^\top G_{\mathrm{ext}}^{[\mu]}\boldsymbol{z}_\mu$ in (14) with $\chi^\mu(\boldsymbol{w}_\mu)$ gives

$$\sum_{\mu\in[n]}\frac{1}{1+\chi^\mu(\boldsymbol{w}_\mu)}G_{\mathrm{ext}}^{[\mu]}\boldsymbol{z}_\mu\boldsymbol{z}_\mu^\top + G_{\mathrm{ext}}(\pi B_{\mathrm{ext}} + \tau\lambda I) \simeq I\,. \tag{17}$$

In this equation we will also replace $\boldsymbol{z}_\mu\boldsymbol{z}_\mu^\top$ with its conditional expectation over $X, \boldsymbol{x}_{\mathrm{test}}, \boldsymbol{\varepsilon}$, giving

$$\frac{\tau}{n}\sum_{\mu\in[n]}\frac{1}{1+\chi^\mu(\boldsymbol{w}_\mu)}G_{\mathrm{ext}}^{[\mu]}\Upsilon(\boldsymbol{w}_\mu) + G_{\mathrm{ext}}(\pi B_{\mathrm{ext}} + \tau\lambda I) \simeq I \tag{18}$$

where recall $\tau = n/d^2$.

In high dimensions and for large n, there is negligible difference between $\sum_{\nu\neq\mu}$ and $\sum_\mu$. Thus, we replace G_{ext}^μ by G_{ext}, and $\chi^\mu(\boldsymbol{w}_\mu)$ by

$$\chi(\boldsymbol{w}_\mu) \equiv \frac{1}{d^2}\,\mathrm{tr}\left([G_{\mathrm{ext}}]_{\setminus 0}\cdot\left[I\otimes E(\boldsymbol{w}_\mu) + \frac{1}{d}\mathcal{X}(\boldsymbol{w}_\mu)\right]\right). \tag{19}$$

So finally we have the expression for G_{ext}

$$G_{\mathrm{ext}}\left(\frac{\tau}{n}\sum_{\mu\in[n]}\frac{1}{1+\chi(\boldsymbol{w}_\mu)}\Upsilon(\boldsymbol{w}_\mu) + \pi B_{\mathrm{ext}} + \tau\lambda I\right) \simeq I\,. \tag{20}$$

Exploit finiteness of training task set. So far we are summing over n task vectors, but really only n/k of these are unique. Thus, we can simplify (20) as

$$G_{\mathrm{ext}}\left(\frac{\tau}{k}\sum_{j\in[k]}\frac{1}{1+\chi(\boldsymbol{w}_j)}\Upsilon(\boldsymbol{w}_j) + \pi B_{\mathrm{ext}} + \tau\lambda I\right) \simeq I\,. \tag{21}$$

We replace $\chi(\boldsymbol{w}_j)$, which is self-averaging in $\boldsymbol{w}_j$, with its mean

$$\widehat{\chi}_{\mathrm{ave}} \equiv \frac{1}{k}\sum_{j\in[k]}\chi(\boldsymbol{w}_j). \tag{22}$$

To clean up the sums over the tasks $\boldsymbol{w}_1, ...\boldsymbol{w}_k$ we use the that

$$\frac{1}{k}\sum_{j\in[k]} \mathrm{tr}\Big[\boldsymbol{w}_j\boldsymbol{w}_j^\top\Big] = \mathrm{tr}[R_{\mathrm{tr}}]$$

$$\frac{1}{k}\sum_{j\in[k]} \Big(\rho + \mathrm{tr}\Big[\boldsymbol{w}_j\boldsymbol{w}_j^\top\Big]\Big)\boldsymbol{w}_j \simeq (\rho + \mathrm{tr}[R_{\mathrm{tr}}])\,\boldsymbol{b}_{\mathrm{tr}}$$

$$\frac{1}{k}\sum_{j\in[k]} \Big(\rho + \mathrm{tr}\Big[\boldsymbol{w}_j\boldsymbol{w}_j^\top\Big]\Big)^2 \simeq (\rho + \mathrm{tr}[R_{\mathrm{tr}}])^2$$

for

$$\boldsymbol{b}_{\mathrm{tr}} \equiv \frac{1}{k}\sum_{j\in[k]} \boldsymbol{w}_j\,, \qquad R_{\mathrm{tr}} \equiv \frac{1}{k}\sum_{j\in[k]} \boldsymbol{w}_j\boldsymbol{w}_j^\top.$$

We thus have that

$$\frac{1}{k}\sum_{j\in[k]} E(\boldsymbol{w}_j) \approx B_{\mathrm{tr}} \equiv \begin{bmatrix} R_{\mathrm{tr}} + \rho_2 I_d/\alpha & \rho_1\boldsymbol{b}_{\mathrm{tr}} \\ \rho_1\boldsymbol{b}_{\mathrm{tr}}^\top & \rho_1^2 \end{bmatrix}$$

$$\frac{1}{k}\sum_{j\in[k]} \mathcal{X}(\boldsymbol{w}_j) \approx \Phi$$

for

$$\begin{aligned}\Phi \equiv\ & \frac{1}{\alpha^2}(\mathrm{tr}[R_{\mathrm{tr}}]k_1 + \rho k_0)\,\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)^\top \\ &+\ \frac{1}{\alpha}k_0\Big(\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)\mathrm{vec}\left(\begin{bmatrix} R_{\mathrm{tr}} & \mathbf{0}\end{bmatrix}\right)^\top + \mathrm{vec}\left(\begin{bmatrix} R_{\mathrm{tr}} & \mathbf{0}\end{bmatrix}\right)\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)^\top\Big) \\ &+\ \frac{1}{\alpha}\rho_1 k_0\Big(\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)\mathrm{vec}\left(\boldsymbol{b}_{\mathrm{tr}}\boldsymbol{e}_{d+1}^\top\right)^\top + \mathrm{vec}\left(\boldsymbol{b}_{\mathrm{tr}}\boldsymbol{e}_{d+1}^\top\right)\mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0}\end{bmatrix}\right)^\top\Big)\end{aligned}$$

where

$$\rho_1 = \mathrm{tr}[R_{\mathrm{tr}}] + \rho\,, \qquad \rho_2 = \mathrm{tr}\big[K^2\big]\,\mathrm{tr}[R_{\mathrm{tr}}] + \rho.$$

Finally, we have

$$\widehat{\chi}_{\mathrm{ave}} = \frac{1}{d^2}\,\mathrm{tr}\left([G_{\mathrm{ext}}]_{\setminus 0}\cdot\left[I\otimes B_{\mathrm{tr}} + \frac{1}{d}\Phi\right]\right). \tag{23}$$

Thus, after averaging over $X, \boldsymbol{x}_{\mathrm{test}}, \boldsymbol{\varepsilon}$ in the extended resolvent G_{ext}, we have a deterministic equivalent $G_{\mathrm{ext}} \simeq \mathcal{G}_{\mathrm{ext}}$ (still depending on random task quantities $R_{\mathrm{tr}}, \boldsymbol{b}_{\mathrm{tr}}$) defined by self-consistent equations

$$[\mathcal{G}_{\mathrm{ext}}]_{\setminus 0} = \left(\frac{\tau}{1+\chi_\pi}I_d\otimes B_{\mathrm{tr}} + \frac{1}{d}\frac{\tau}{1+\chi_\pi}\Phi + \pi\Pi + \tau\lambda I_d\otimes I_{d+1}\right)^{-1} \tag{24}$$

$$\chi_\pi = \frac{1}{d^2}\,\mathrm{tr}\left([\mathcal{G}_{\mathrm{ext}}]_{\setminus 0}\left(I_d\otimes B_{\mathrm{tr}} + \frac{1}{d}\Phi\right)\right) \tag{25}$$

where the full matrix is given by

$$\begin{aligned}\mathcal{G}_{\mathrm{ext}}(\pi)^{-1} \equiv\ & \frac{\tau}{1+\chi_\pi}\begin{bmatrix} \rho_1 & \frac{1}{\sqrt{d}}\,\mathrm{vec}\left(\begin{bmatrix} R_{\mathrm{tr}} + \frac{1}{\alpha}k_0\,\mathrm{tr}[R_{\mathrm{tr}}]I_d & \rho_1\boldsymbol{b}_{\mathrm{tr}}\end{bmatrix}\right)^\top \\ \frac{1}{\sqrt{d}}\,\mathrm{vec}\left(\begin{bmatrix} R_{\mathrm{tr}} + \frac{1}{\alpha}k_0\,\mathrm{tr}[R_{\mathrm{tr}}]I_d & \rho_1\boldsymbol{b}_{\mathrm{tr}}\end{bmatrix}\right) & I_d\otimes B_{\mathrm{tr}} + \frac{1}{d}\Phi\end{bmatrix} \\ &+\ \pi B_{\mathrm{ext}} + \tau\lambda I\end{aligned} \tag{26}$$

Note that up until this point, the results have matched the previous analysis from Lu et al. (2025) and Letey et al. (2026) with the exception of ρ_2 (depending on k_2) and the low-rank term Φ (depending on k_0 and k_1).

The low-rank term is given by

$$\begin{aligned}\Phi &= \phi_1 1_+ 1_+^\top + \phi_2(1_+ r^\top + r 1_+^\top) + \phi_3(1_+ \mu^\top + \mu 1_+^\top) \\ &= \begin{bmatrix} 1_+ & v \end{bmatrix} \begin{bmatrix} \phi_1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1_+^\top \\ v^\top \end{bmatrix} \\ &= T \phi T^\top\end{aligned}$$

for

$$1_+ \equiv \mathrm{vec}\left(\begin{bmatrix} I_d & \mathbf{0} \end{bmatrix}\right), \qquad r \equiv \mathrm{vec}\left(\begin{bmatrix} R_{\mathrm{tr}} & \mathbf{0} \end{bmatrix}\right), \qquad \mu \equiv \mathrm{vec}\left(\boldsymbol{b}_{\mathrm{tr}} \boldsymbol{e}_{d+1}^\top\right), \qquad v \equiv \mathrm{vec}\left(\begin{bmatrix} \phi_2 R_{\mathrm{tr}} & \phi_3 \boldsymbol{b}_{\mathrm{tr}} \end{bmatrix}\right)$$

and

$$\phi_1 \equiv \frac{1}{\alpha^2}(\mathrm{tr}[R_{\mathrm{tr}}] k_1 + \rho k_0), \qquad \phi_2 \equiv \frac{1}{\alpha} k_0, \qquad \phi_3 \equiv \frac{1}{\alpha}(\mathrm{tr}[R_{\mathrm{tr}}] + \rho) k_0.$$

G.1. Effective ridge

The effective ridge will be $\tilde{\lambda} = \lambda(1 + \chi_0)$ where χ_0 is defined as the solution to the implicit equations

$$[\mathcal{G}_{\mathrm{ext}}]_{\backslash 0} = \left(\frac{\tau}{1+\chi_0} I_d \otimes B_{\mathrm{tr}} + \frac{1}{d}\frac{\tau}{1+\chi_0}\Phi + \tau\lambda I_d \otimes I_{d+1}\right)^{-1} \tag{27}$$

$$\chi_0 = \frac{1}{d^2}\,\mathrm{tr}\left([\mathcal{G}_{\mathrm{ext}}]_{\backslash 0}\left(I_d \otimes B_{\mathrm{tr}} + \frac{1}{d}\Phi\right)\right). \tag{28}$$

We can recast the above implicit equations using Woodbury, as

$$\frac{\tau\chi_0}{1+\chi_0} = \frac{1}{d^2}\,\mathrm{tr}\left(\left(I_d \otimes F_0 - \frac{1}{d}\Phi_C\right)\left(I_d \otimes B_{\mathrm{tr}} + \frac{1}{d}\Phi\right)\right) \tag{29}$$

where

$$F_0 = (B_{\mathrm{tr}} + \lambda(1+\chi_0) I_{d+1})^{-1}$$

$$\Phi_C = (I_d \otimes F_0)\begin{bmatrix} 1_+ & v \end{bmatrix} S \begin{bmatrix} 1_+^\top \\ v^\top \end{bmatrix}(I_d \otimes F_0) \qquad \text{for some } 2\times 2 \text{ mtx } S \text{ with } \Theta(1) \text{ elements.}$$

However, all the $1/d$ terms in (29) are subleading, e.g.

$$\frac{1}{d^3}\,\mathrm{tr}\left((I_d \otimes F_0) r r^\top (I_d \otimes F_0 B_{\mathrm{tr}})\right) = \frac{1}{d^3}\,\mathrm{tr}(R_{\mathrm{tr}} F_0 B_{\mathrm{tr}} F_0 R_{\mathrm{tr}}) = \mathcal{O}\left(\frac{1}{d^2}\right)$$

and so we write

$$\frac{\tau\chi_0}{1+\chi_0} \approx \frac{1}{d^2}\,\mathrm{tr}\left(I_d \otimes F_0 B_{\mathrm{tr}}\right) = \mathrm{tr}\left[(B_{\mathrm{tr}} + \tilde{\lambda} I_{d+1})^{-1} B_{\mathrm{tr}}\right]. \tag{30}$$

This can be simplified as

$$\begin{aligned}\frac{\tau\chi_0}{1+\chi_0} &= 1 - \lambda(1+\chi_0)\,\mathrm{tr}\left[(B_{\mathrm{tr}} + \lambda(1+\chi_0) I_{d+1})^{-1}\right] \\ &\approx 1 - \lambda(1+\chi_0)\,\mathrm{tr}\left[\left(R_{\mathrm{tr}} + \left(\lambda(1+\chi_0) + \frac{\rho_2}{\alpha}\right) I_d\right)^{-1}\right]\end{aligned}$$

after ignoring the $\boldsymbol{b}_{\mathrm{tr}}$ terms in B_{tr}. This is where the Stieltjes transform of R_{tr} is first introduced:

$$\mathcal{M}_\kappa(w) = \lim_{d\to\infty, k\to\infty, k/d=\kappa} \frac{1}{d}\,\mathrm{tr}\left((R_{\mathrm{tr}} + w I_d)^{-1}\right).$$

Using this, we find the effective ridge self-consistency equation

$$\tilde{\lambda}\mathcal{M}_\kappa\Big(\tilde{\lambda}+\frac{\rho_2}{\alpha}\Big)+\frac{\lambda\tau}{\tilde{\lambda}}=1-\tau \tag{31}$$

G.2. Relating Γ^* to G

Recall that

$$\mathrm{vec}(\Gamma^*)=\left(\frac{n}{d}\lambda I+\sum_{\mu=1}^{n}\mathrm{vec}(H^\mu)\mathrm{vec}(H^\mu)^\top\right)^{-1}\sum_{\mu=1}^{n}y^\mu_{\mathrm{test}}\mathrm{vec}(H^\mu)$$

and so we have

$$G_{\mathrm{ext}}(0)=d\begin{bmatrix}\frac{1}{d}\sum_\mu (y^\mu_{\mathrm{test}})^2+\frac{n}{d}\lambda & \frac{1}{\sqrt{d}}\sum_\mu y^\mu_{\mathrm{test}}\,\mathrm{vec}(H^\mu)^\top\\ \frac{1}{\sqrt{d}}\sum_\mu y^\mu_{\mathrm{test}}\,\mathrm{vec}(H^\mu) & \sum_\mu \mathrm{vec}(H^\mu)\,\mathrm{vec}(H^\mu)^\top+\frac{n}{d}\lambda I\end{bmatrix}^{-1}$$

Using that

$$\begin{bmatrix}a & \boldsymbol{b}^\top\\ \boldsymbol{b} & D\end{bmatrix}^{-1}=\begin{bmatrix}c & -c\boldsymbol{q}^\top\\ -c\boldsymbol{q} & D^{-1}+c\boldsymbol{q}\boldsymbol{q}^\top\end{bmatrix},\qquad c=\frac{1}{a-\boldsymbol{b}^\top D^{-1}\boldsymbol{b}},\quad \boldsymbol{q}=D^{-1}\boldsymbol{b}$$

we have a formula for Γ^* in terms of $G_{\mathrm{ext}}(0)$ as

$$\frac{\sqrt{d}}{\boldsymbol{e}_1^\top G_{\mathrm{ext}}(0)\boldsymbol{e}_1}G_{\mathrm{ext}}(0)\boldsymbol{e}_1=\begin{bmatrix}\sqrt{d}\\ \mathrm{vec}(\Gamma)^*\end{bmatrix}$$

Using our above formulas, we also have

$$\frac{\sqrt{d}}{\boldsymbol{e}_1^\top \mathcal{G}_{\mathrm{ext}}(0)\boldsymbol{e}_1}\mathcal{G}_{\mathrm{ext}}(0)\boldsymbol{e}_1=\begin{bmatrix}\sqrt{d}\\ \Big(I_d\otimes B_{\mathrm{tr}}+\frac{1}{d}\Phi+\tilde{\lambda}I_{d(d+1)}\Big)^{-1}\mathrm{vec}\left(\begin{bmatrix}R_{\mathrm{tr}}+\frac{k_0}{\alpha}\,\mathrm{tr}[R_{\mathrm{tr}}]I_d & \rho_1\boldsymbol{b}_{\mathrm{tr}}\end{bmatrix}\right)\end{bmatrix}$$

and so

$$\mathrm{vec}(\Gamma^*)\simeq\mathrm{vec}(\Gamma_{\mathrm{de}})\equiv\Big(I_d\otimes B_{\mathrm{tr}}+\frac{1}{d}\Phi+\tilde{\lambda}I_{d(d+1)}\Big)^{-1}\mathrm{vec}\left(\begin{bmatrix}R_{\mathrm{tr}}+\frac{k_0}{\alpha}\,\mathrm{tr}[R_{\mathrm{tr}}]I_d & \rho_1\boldsymbol{b}_{\mathrm{tr}}\end{bmatrix}\right)=F_\Phi\boldsymbol{g}$$

for

$$F_\Phi=\Big(I_d\otimes B_{\mathrm{tr}}+\frac{1}{d}\Phi+\tilde{\lambda}I_{d(d+1)}\Big)^{-1}$$
$$\boldsymbol{g}=\mathrm{vec}\left(\begin{bmatrix}R_{\mathrm{tr}}+\frac{k_0}{\alpha}\,\mathrm{tr}[R_{\mathrm{tr}}]I_d & \rho_1\boldsymbol{b}_{\mathrm{tr}}\end{bmatrix}\right).$$

This immediately gives the linear term in the MSE error expression as

$$\mathrm{tr}\Big[\Gamma^*A^\top\Big]=\frac{1}{d}\,\mathrm{tr}\Big(\Gamma^*A^\top\Big)=\frac{1}{d}\,\mathrm{tr}\Big(\mathrm{vec}(\Gamma^*)\,\mathrm{vec}(A)^\top\Big)\simeq\frac{1}{d}\,\mathrm{tr}\Big(\mathrm{vec}(\Gamma_{\mathrm{de}})\,\mathrm{vec}(A)^\top\Big)$$

where we will specifically use $\mathrm{vec}(A)=(1+\phi_2)1_+$ from Lemma 1.

For the quadratic terms we need to be a bit more careful, as the correct object to work with for high-dimensional equivalence is technically the components of $\mathcal{G}$ and not this linear representation of $\mathrm{vec}(\Gamma_{\mathrm{de}})$. This is what the parameterisation πB_{ext} is for. Take

$$B_{\mathrm{ext}}=\begin{bmatrix}0 & 0\\ 0 & \Pi\end{bmatrix}$$

for some $\Pi\in\mathbb{R}^{d(d+1)\times d(d+1)}$. Then

$$\frac{\mathrm{d}}{\mathrm{d}\pi}\frac{1}{c(\pi)}(\pi=0)=\frac{1}{d}\,\mathrm{vec}(\Gamma^*)^\top\Pi\,\mathrm{vec}(\Gamma^*)$$

where $c(\pi) = \boldsymbol{e}_1^\top G_{\text{ext}}(\pi)\boldsymbol{e}_1$. We can safely replace

$$c(\pi) = \boldsymbol{e}_1^\top G_{\text{ext}}(\pi)\boldsymbol{e}_1 \simeq \boldsymbol{e}_1\mathcal{G}_{\text{ext}}(\pi)\boldsymbol{e}_1.$$

By Schur complement on $\mathcal{G}_{\text{ext}}(\pi)$ we have

$$\begin{aligned}\frac{1}{\boldsymbol{e}_1\mathcal{G}_{\text{ext}}(\pi)\boldsymbol{e}_1} &= \frac{\tau}{1+\chi_\pi}\rho_1 + \tau\lambda - \frac{1}{d}\frac{\tau^2}{(1+\chi_\pi)^2}\boldsymbol{g}^\top\left(\frac{\tau}{1+\chi_\pi}\left(I_d\otimes B_{\text{tr}} + \frac{1}{d}\Phi\right) + \pi\Pi + \tau\lambda I\right)^{-1}\boldsymbol{g} \\ &= \frac{\tau}{1+\chi_\pi}\rho_1 + \tau\lambda - \frac{1}{d}\frac{\tau}{1+\chi_\pi}\boldsymbol{g}^\top\left(I_d\otimes B_{\text{tr}} + \frac{1}{d}\Phi + \pi\frac{1+\chi_\pi}{\tau}\Pi + \lambda(1+\chi_\pi)I\right)^{-1}\boldsymbol{g}.\end{aligned}$$

Given the eventual MSE term we want from Lemma 1, we choose

$$\Pi = I_d\otimes B + \frac{1}{d}\Psi$$

for

$$\Psi = \begin{bmatrix}1_+ & \text{vec}\left(\begin{bmatrix}I_d & \mathbf{0}\end{bmatrix}\right)\end{bmatrix}\begin{bmatrix}\phi_1 & \phi_2\\ \phi_2 & 0\end{bmatrix}\begin{bmatrix}1_+^\top \\ \text{vec}\left(\begin{bmatrix}I_d & \mathbf{0}\end{bmatrix}\right)^\top\end{bmatrix} = (\phi_1+2\phi_2)1_+1_+^\top.$$

Using the same approximation as before, we will take

$$[\mathcal{G}_{\text{ext}}]_{\setminus 0} = I_d\otimes\left(\frac{\tau}{1+\chi_\pi}B_{\text{tr}} + \pi B + \tau\lambda I\right)^{-1} + \frac{1}{d}\text{ low rank terms from }\Phi\text{ and }\Psi$$

and approximate χ_π as

$$\chi_\pi = \text{tr}\left[\left(\frac{\tau}{1+\chi_\pi}B_{\text{tr}} + \pi B + \tau\lambda I\right)^{-1}B_{\text{tr}}\right].$$

As before, we find that

$$\frac{\tau\chi_0'}{(1+\chi_0)^2} = \frac{\text{tr}[F_0BF_0B_{\text{tr}}]}{\text{tr}[F_0B_{\text{tr}}F_0B_{\text{tr}}]-\tau}, \qquad F_0 = (B_{\text{tr}}+\tilde{\lambda}I)^{-1}.$$

Thus,

$$\frac{1}{d}\,\text{vec}(\Gamma^*)^\top\Pi\,\text{vec}(\Gamma^*) \simeq \frac{1}{d}\,\text{vec}(\Gamma_{\text{de}})^\top\Pi\,\text{vec}(\Gamma_{\text{de}}) - \frac{\tau\chi_0'}{(1+\chi_0)^2}\left(\rho_1 - \frac{1}{d}\boldsymbol{g}^\top\,\text{vec}(\Gamma_{\text{de}}) - \tilde{\lambda}\frac{1}{d}\,\text{vec}(\Gamma_{\text{de}})^\top\,\text{vec}(\Gamma_{\text{de}})\right).$$

Note that

$$\text{vec}(\Gamma_{\text{de}}) = (I_d\otimes F_0)\boldsymbol{g} - \frac{1}{d}\Phi_C\boldsymbol{g}$$

$$\Pi = I_d\otimes B_{\text{test}} + \frac{1}{d}\Psi$$

where

$$\begin{aligned}\Phi_C &= ((I_d\otimes F_0)T)\begin{bmatrix}\mathcal{M} & 1+\phi_2(1-\sigma\mathcal{M})\\ 1+\phi_2(1-\sigma\mathcal{M}) & -\phi_1+\phi_2^2(1-\sigma+\sigma^2\mathcal{M})\end{bmatrix}^{-1}((I_d\otimes F_0)T)^\top \\ &= (I_d\otimes F_0)\begin{bmatrix}1_+ & \phi_2 r+\phi_3\mu\end{bmatrix}S\begin{bmatrix}1_+^\top \\ \phi_2 r^\top + \phi_3\mu^\top\end{bmatrix}(I_d\otimes F_0)\end{aligned}$$

This comes from expanding

$$F_0 = \begin{bmatrix} F & \boldsymbol{f} \\ \boldsymbol{f}^\top & f \end{bmatrix}$$

$$\frac{1}{d}\begin{bmatrix} 1_+^\top \\ r^\top \\ \mu^\top \end{bmatrix}\begin{bmatrix} 1_+ & r & \mu \end{bmatrix} = \begin{bmatrix} \operatorname{tr}[F] & \operatorname{tr}[FR_{\rm tr}] & \operatorname{tr}\left[\boldsymbol{f}\boldsymbol{b}_{\rm tr}^\top\right] \\ \operatorname{tr}[FR_{\rm tr}] & \operatorname{tr}\left[FR_{\rm tr}^2\right] & \operatorname{tr}\left[F\boldsymbol{f}\boldsymbol{b}_{\rm tr}^\top\right] \\ \operatorname{tr}\left[\boldsymbol{f}\boldsymbol{b}_{\rm tr}^\top\right] & \operatorname{tr}\left[F\boldsymbol{f}\boldsymbol{b}_{\rm tr}^\top\right] & f\operatorname{tr}\left[\boldsymbol{b}_{\rm tr}\boldsymbol{b}_{\rm tr}^\top\right] \end{bmatrix}$$

$$\simeq \begin{bmatrix} \mathcal{M}_\kappa(\sigma) & 1-\sigma\mathcal{M}_\kappa(\sigma) & 0 \\ 1-\sigma\mathcal{M}_\kappa(\sigma) & 1-\sigma+\sigma^2\mathcal{M}_\kappa(\sigma) & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \sigma \equiv \frac{\rho_2}{\alpha} + \tilde{\lambda}.$$

As we can see, all the $\boldsymbol{b}_{\rm tr}, \boldsymbol{f}$ contributions are negligible compared to the resolvent $(R_{\rm tr} + \sigma I_d)^{-1}$ contributions. Using similar intuition, in the expansions of the linear and quadratic error terms, we will neglect the $\rho_1\boldsymbol{b}_{\rm tr}$ component of $\boldsymbol{g}$, as well as $\mu = \operatorname{vec}(\boldsymbol{b}_{\rm tr}\boldsymbol{e}_{d+1}^\top)$ in Φ_C.

Linear error term Following this, and writing the matrix component of $\boldsymbol{g}$ as

$$\tilde{R} = R_{\rm tr} + \frac{k_0}{\alpha}\operatorname{tr}[R_{\rm tr}]I_d$$

we have

$$\frac{1}{d}\operatorname{vec}(A)^\top\operatorname{vec}(\Gamma_{\rm de}) = (1+\phi_2)\frac{1}{d}\left(1_+^\top(I\otimes F)\boldsymbol{g} - \frac{1}{d}1_+^\top\Phi_C\boldsymbol{g}\right)$$

$$\simeq (1+\phi_2)\left(\operatorname{tr}\left[F\tilde{R}\right] - \begin{bmatrix}\operatorname{tr}[F] & \phi_2\operatorname{tr}[FR_{\rm tr}]\end{bmatrix} S \begin{bmatrix} \operatorname{tr}\left[F\tilde{R}\right] \\ \phi_2\operatorname{tr}\left[R_{\rm tr}F\tilde{R}\right]\end{bmatrix}\right)$$

Quadratic error term Recall

$$B = \begin{bmatrix} (1+\frac{\rho_2}{\alpha})I_d & \boldsymbol{0} \\ \boldsymbol{0}^\top & \rho_1^2 \end{bmatrix}, \qquad \Psi = (\phi_1+2\phi_2)1_+1_+^\top.$$

Have

$$\frac{1}{d}\boldsymbol{g}^\top\operatorname{vec}(\Gamma_{\rm de}) \approx \operatorname{tr}\left[\tilde{R}F\tilde{R}\right] - \begin{bmatrix}\operatorname{tr}\left[\tilde{R}F\right] & \phi_2\operatorname{tr}\left[\tilde{R}FR_{\rm tr}\right]\end{bmatrix} S \begin{bmatrix} \operatorname{tr}\left[F\tilde{R}\right] \\ \phi_2\operatorname{tr}\left[R_{\rm tr}F\tilde{R}\right]\end{bmatrix}$$

$$\frac{1}{d}\operatorname{vec}(\Gamma_{\rm de})^\top\operatorname{vec}(\Gamma_{\rm de}) \approx \operatorname{tr}\left[\tilde{R}F^2\tilde{R}\right] - 2\begin{bmatrix}\operatorname{tr}\left[\tilde{R}F\right] & \phi_2\operatorname{tr}\left[\tilde{R}FR_{\rm tr}\right]\end{bmatrix} S \begin{bmatrix} \operatorname{tr}\left[F^2\tilde{R}\right] \\ \phi_2\operatorname{tr}\left[R_{\rm tr}F^2\tilde{R}\right]\end{bmatrix}$$

$$+ \begin{bmatrix}\operatorname{tr}\left[\tilde{R}F\right] & \phi_2\operatorname{tr}\left[\tilde{R}FR_{\rm tr}\right]\end{bmatrix} S^\top \begin{bmatrix} \operatorname{tr}\left[F^2\right] & \phi_2\operatorname{tr}\left[F^2R_{\rm tr}\right] \\ \phi_2\operatorname{tr}\left[F^2R_{\rm tr}\right] & \phi_2^2\operatorname{tr}\left[R_{\rm tr}F^2R_{\rm tr}\right]\end{bmatrix} S \begin{bmatrix} \operatorname{tr}\left[F\tilde{R}\right] \\ \phi_2\operatorname{tr}\left[R_{\rm tr}F\tilde{R}\right]\end{bmatrix}$$

$$\frac{1}{d}\operatorname{vec}(\Gamma_{\rm de})^\top(I_d\otimes B)\operatorname{vec}(\Gamma_{\rm de}) \approx \left(1+\frac{\rho_2}{\alpha}\right)\cdot\frac{1}{d}\operatorname{vec}(\Gamma_{\rm de})^\top\operatorname{vec}(\Gamma_{\rm de})$$

$$\frac{1}{\phi_1+2\phi_2}\frac{1}{d^2}\operatorname{vec}(\Gamma_{\rm de})^\top\Psi\operatorname{vec}(\Gamma_{\rm de}) \approx \operatorname{tr}\left[\tilde{R}F\right]^2 - 2\operatorname{tr}\left[\tilde{R}F\right]\begin{bmatrix}\operatorname{tr}[F] & \phi_2\operatorname{tr}[R_{\rm tr}F]\end{bmatrix} S \begin{bmatrix} \operatorname{tr}\left[F\tilde{R}\right] \\ \phi_2\operatorname{tr}\left[R_{\rm tr}F\tilde{R}\right]\end{bmatrix}$$

$$+ \left(\begin{bmatrix}\operatorname{tr}[F] & \phi_2\operatorname{tr}[FR_{\rm tr}]\end{bmatrix} S \begin{bmatrix} \operatorname{tr}\left[F\tilde{R}\right] \\ \phi_2\operatorname{tr}\left[R_{\rm tr}F\tilde{R}\right]\end{bmatrix}\right)^2.$$

The final step is to remember that $F = (R_{\rm tr} + \sigma)^{-1}$ and $\tilde{R} = R_{\rm tr} + k_0 I_d/\alpha$ and so we have the following dictionary of terms

$$\begin{aligned}
\mathrm{tr}[F] &\simeq \mathcal{M}_\kappa(\sigma) \\
\mathrm{tr}[FR_{\rm tr}] &\simeq 1 - \sigma\mathcal{M}_\kappa(\sigma) \\
\mathrm{tr}\Big[F\tilde{R}\Big] &\simeq 1 - \sigma\mathcal{M}_\kappa(\sigma) + \frac{k_0}{\alpha}\mathcal{M}_\kappa(\sigma) \\
\mathrm{tr}\Big[FR_{\rm tr}\tilde{R}\Big] &\simeq \frac{k_0}{\alpha}(1 - \sigma\mathcal{M}_\kappa(\sigma)) + 1 - \sigma + \sigma^2\mathcal{M}_\kappa(\sigma) \\
\mathrm{tr}\Big[F\tilde{R}\tilde{R}\Big] &\simeq \left(\frac{k_0}{\alpha}\right)^2\mathcal{M}_\kappa(\sigma) + 2\frac{k_0}{\alpha}(1 - \sigma\mathcal{M}_\kappa(\sigma))) + 1 - \sigma + \sigma^2\mathcal{M}_\kappa(\sigma) \\
\mathrm{tr}\big[F^2\big] &\simeq -\mathcal{M}'_\kappa(\sigma) \\
\mathrm{tr}\big[F^2R_{\rm tr}\big] &\simeq \mathcal{M}_\kappa(\sigma) + \sigma\mathcal{M}'_\kappa(\sigma) \\
\mathrm{tr}\Big[F^2\tilde{R}\Big] &\simeq \mathcal{M}_\kappa(\sigma) + \sigma\mathcal{M}'_\kappa(\sigma) - \frac{k_0}{\alpha}\mathcal{M}'_\kappa(\sigma) \\
\mathrm{tr}\big[F^2R_{\rm tr}R_{\rm tr}\big] &\simeq 1 - 2\sigma\mathcal{M}_\kappa(\sigma) - \sigma^2\mathcal{M}'_\kappa(\sigma) \\
\mathrm{tr}\Big[F^2R_{\rm tr}\tilde{R}\Big] &\simeq 1 - 2\sigma\mathcal{M}_\kappa(\sigma) - \sigma^2\mathcal{M}'_\kappa(\sigma) + \frac{k_0}{\alpha}(\mathcal{M}_\kappa(\sigma) + \sigma\mathcal{M}'_\kappa(\sigma)) \\
\mathrm{tr}\Big[F^2\tilde{R}\tilde{R}\Big] &\simeq 1 - 2\sigma\mathcal{M}_\kappa(\sigma) - \sigma^2\mathcal{M}'_\kappa(\sigma) + 2\frac{k_0}{\alpha}(\mathcal{M}_\kappa(\sigma) + \sigma\mathcal{M}'_\kappa(\sigma)) - \left(\frac{k_0}{\alpha}\right)^2\mathcal{M}'_\kappa(\sigma)
\end{aligned}$$

So finally, if we use some shorthand

$$\tilde{\sigma} = \sigma - \frac{k_0}{\alpha}\,, \quad \boldsymbol{m}_1 = \begin{bmatrix} \mathcal{M} \\ \phi_2(1 - \sigma\mathcal{M}) \end{bmatrix}, \quad \boldsymbol{m}_2 = \begin{bmatrix} 1 - \tilde{\sigma}\mathcal{M} \\ \phi_2(1 - \tilde{\sigma} + \sigma\tilde{\sigma}\mathcal{M}) \end{bmatrix}, \quad \boldsymbol{m}_3 = \begin{bmatrix} \mathcal{M} + \tilde{\sigma}\mathcal{M}' \\ \phi_2(1 - \sigma\mathcal{M} - \tilde{\sigma}\mathcal{M} - \sigma\tilde{\sigma}\mathcal{M}') \end{bmatrix}$$

$$M = \begin{bmatrix} -\mathcal{M}' & \phi_2(\mathcal{M} + \sigma\mathcal{M}') \\ \phi_2(\mathcal{M} + \sigma\mathcal{M}') & \phi_2^2(1 - 2\sigma\mathcal{M} - \sigma^2\mathcal{M}') \end{bmatrix}$$

and so

$$\begin{aligned}
\frac{1}{d}\,\mathrm{vec}(A)^\top\,\mathrm{vec}(\Gamma_{\rm de}) &\simeq (1 + \phi_2)\Big(1 - \tilde{\sigma}\mathcal{M} - \boldsymbol{m}_1^\top S\boldsymbol{m}_2\Big) \\
\frac{1}{d}\boldsymbol{g}^\top\,\mathrm{vec}(\Gamma_{\rm de}) &\simeq 1 + \sigma - 2\tilde{\sigma} + \tilde{\sigma}^2\mathcal{M} - \boldsymbol{m}_2^\top S\boldsymbol{m}_2 \\
\frac{1}{d}\,\mathrm{vec}(\Gamma_{\rm de})^\top\,\mathrm{vec}(\Gamma_{\rm de}) &\simeq 1 - 2\tilde{\sigma}\mathcal{M} - \tilde{\sigma}^2\mathcal{M}' - 2\boldsymbol{m}_2^\top S\boldsymbol{m}_3 + \boldsymbol{m}_2^\top S^\top MS\boldsymbol{m}_2 \\
\frac{1}{d^2}\,\mathrm{vec}(\Gamma_{\rm de})^\top\Psi\,\mathrm{vec}(\Gamma_{\rm de}) &\simeq (\phi_1 + 2\phi_2)\Big((1 - \tilde{\sigma}\mathcal{M}) - \boldsymbol{m}_1^\top S\boldsymbol{m}_2\Big)^2
\end{aligned}$$

with

$$c = \frac{\tau\chi'_0}{(1 + \chi_0)^2} \simeq (1 + \frac{\rho_2}{\alpha})\frac{\mathcal{M} + \tilde{\lambda}\mathcal{M}'}{1 - 2\tilde{\lambda}\mathcal{M} - \tilde{\lambda}^2\mathcal{M}' - \tau}.$$

So finally, we have

$$\begin{aligned}
e_{\rm ICL} = {} & (1 - c)\rho_1 - 2(1 + \phi_2)(1 - \tilde{\sigma}\mathcal{M} - \boldsymbol{m}_1^\top S\boldsymbol{m}_2) + (\phi_1 + 2\phi_2)(1 - \tilde{\sigma}\mathcal{M} - \boldsymbol{m}_1 S\boldsymbol{m}_2)^2 \\
& + \Big(1 + \frac{\rho_2}{\alpha} + c\tilde{\lambda}\Big)(1 - 2\tilde{\sigma}\mathcal{M} - \tilde{\sigma}^2\mathcal{M}' - 2\boldsymbol{m}_2^\top S\boldsymbol{m}_3 + \boldsymbol{m}_2^\top S^\top MS\boldsymbol{m}_2) \\
& + c(1 + \sigma - 2\tilde{\sigma} + \tilde{\sigma}^2\mathcal{M} - \boldsymbol{m}_2^\top S\boldsymbol{m}_2)
\end{aligned}$$

for

$$\rho_1 \equiv 1 + \rho\,, \qquad \rho_2 \equiv \mathrm{tr}[K^2] + \rho$$

$$\sigma = \tilde{\lambda}+\frac{\rho_2}{\alpha}\,, \quad \tilde{\sigma} = \sigma-\frac{k_0}{\alpha}\,, \quad \tilde{\lambda}\mathcal{M}_\kappa\Big(\tilde{\lambda}+\frac{\rho_2}{\alpha}\Big)+\frac{\lambda\tau}{\tilde{\lambda}} = 1-\tau\,, \quad c = (1+\frac{\rho_2}{\alpha})\frac{\mathcal{M}+\tilde{\lambda}\mathcal{M}'}{1-2\tilde{\lambda}\mathcal{M}-\tilde{\lambda}^2\mathcal{M}'-\tau}$$

$$\boldsymbol{m}_1 = \begin{bmatrix}\mathcal{M}\\ \phi_2(1-\sigma\mathcal{M})\end{bmatrix}\,, \quad \boldsymbol{m}_2 = \begin{bmatrix}1-\tilde{\sigma}\mathcal{M}\\ \phi_2(1-\tilde{\sigma}+\sigma\tilde{\sigma}\mathcal{M})\end{bmatrix}\,, \quad \boldsymbol{m}_3 = \begin{bmatrix}\mathcal{M}+\tilde{\sigma}\mathcal{M}'\\ \phi_2(1-\sigma\mathcal{M}-\tilde{\sigma}\mathcal{M}-\sigma\tilde{\sigma}\mathcal{M}')\end{bmatrix}$$

$$M = \begin{bmatrix}-\mathcal{M}' & \phi_2(\mathcal{M}+\sigma\mathcal{M}')\\ \phi_2(\mathcal{M}+\sigma\mathcal{M}') & \phi_2^2(1-2\sigma\mathcal{M}-\sigma^2\mathcal{M}')\end{bmatrix}\,, \quad S = \begin{bmatrix}\mathcal{M} & 1+\phi_2(1-\sigma\mathcal{M})\\ 1+\phi_2(1-\sigma\mathcal{M}) & -\phi_1+\phi_2^2(1-\sigma+\sigma^2\mathcal{M})\end{bmatrix}^{-1}$$

$$\phi_1 \equiv \frac{1}{\alpha^2}(k_1+\rho k_0) = \phi_1\,, \quad \phi_2 \equiv \frac{1}{\alpha}k_0 = \phi_2$$

Sanity check For $K = I_d$ have $\rho_1 = \rho_2$, $\sigma = \tilde{\sigma}$, $\phi_1 = \phi_2 = \phi_1 = \phi_2 = 0$, and all $\boldsymbol{m}, S, M$ terms don't contribute. Get

$$\begin{aligned}
e_{\text{ICL}} &= 1+\rho-2(1-\sigma\mathcal{M})+\Big(1+\frac{1+\rho}{\alpha}\Big)(1-2\sigma\mathcal{M}-\sigma^2\mathcal{M}')\\
&\quad - \; c(\rho+\sigma-\sigma^2\mathcal{M}-\tilde{\lambda}(1-2\sigma\mathcal{M}-\sigma^2\mathcal{M}'))\\
&= 1+\rho-2(1-\sigma\mathcal{M})+\Big(1+\frac{1+\rho}{\alpha}\Big)(1-2\sigma\mathcal{M}-\sigma^2\mathcal{M}')+\Big(1+\frac{1+\rho}{\alpha}\Big)(\mathcal{M}+\tilde{\lambda}\mathcal{M}')q\\
&= \rho+q\mathcal{M}+(\tilde{\lambda}q-\sigma^2)\mathcal{M}'+\frac{1+\rho}{\alpha}(1-(q-2\sigma)\mathcal{M}+(\tilde{\lambda}q-\sigma^2)\mathcal{M}')
\end{aligned}$$

as previously in Letey et al. (2026).

G.3. Dependence on query terms

This formula can be cleaned up a fair bit by gathering terms that depend on k_0, k_1 *i.e.*, the query correlation terms. First we write all the terms depending on c as

$$\begin{aligned}
&(-c)\Big(\rho_1-\tilde{\lambda}\Big(1-2\tilde{\sigma}\mathcal{M}-\tilde{\sigma}^2\mathcal{M}'-2\boldsymbol{m}_2^\top S\boldsymbol{m}_3+\boldsymbol{m}_2^\top S^\top MS\boldsymbol{m}_2\Big)-\Big(1+\sigma-2\tilde{\sigma}+\tilde{\sigma}^2\mathcal{M}-\boldsymbol{m}_2^\top S\boldsymbol{m}_2\Big)\Big)\\
&= \Big(1+\frac{\rho_2}{\alpha}\Big)\frac{\mathcal{M}+\tilde{\lambda}\mathcal{M}'}{\tau-(1-2\tilde{\lambda}\mathcal{M}-\tilde{\lambda}^2\mathcal{M}')}\Big(\rho+\sigma-\sigma^2\mathcal{M}-\tilde{\lambda}(1-2\sigma\mathcal{M}-\sigma^2\mathcal{M}')\Big)\\
&\quad - \Big(1+\frac{\rho_2}{\alpha}\Big)\frac{\mathcal{M}+\tilde{\lambda}\mathcal{M}'}{\tau-(1-2\tilde{\lambda}\mathcal{M}-\tilde{\lambda}^2\mathcal{M}')}\Big(2\frac{k_0}{\alpha}\Big(\tilde{\lambda}(\mathcal{M}+\sigma\mathcal{M}')+(1-2\sigma\mathcal{M})\Big)+\frac{k_0^2}{\alpha^2}\Big(\mathcal{M}-\tilde{\lambda}\mathcal{M}'\Big)\Big)\\
&\quad + \Big(1+\frac{\rho_2}{\alpha}\Big)\frac{\mathcal{M}+\tilde{\lambda}\mathcal{M}'}{\tau-(1-2\tilde{\lambda}\mathcal{M}-\tilde{\lambda}^2\mathcal{M}')}\Big(\boldsymbol{m}_2^\top S\boldsymbol{m}_2-\tilde{\lambda}\Big(-2\boldsymbol{m}_2^\top S\boldsymbol{m}_3+\boldsymbol{m}_2^\top S^\top MS\boldsymbol{m}_2\Big)\Big)\\
&= \Big(1+\frac{\rho_2}{\alpha}\Big)(\mathcal{M}+\tilde{\lambda}\mathcal{M}')(q_{\text{old}}+q_{\text{query}})
\end{aligned}$$

and the remaining ICL error terms can be expressed as

$$\begin{aligned}
&\rho_1-2(1-\sigma\mathcal{M})+\Big(1+\frac{\rho_2}{\alpha}\Big)(1-2\sigma\mathcal{M}-\sigma^2\mathcal{M}')\\
&\quad - \frac{k_0}{\alpha}\Big(2\mathcal{M}-2\Big(1+\frac{\rho_2}{\alpha}\Big)(\mathcal{M}+\sigma\mathcal{M}')\Big)-\frac{k_0^2}{\alpha^2}\Big(1+\frac{\rho_2}{\alpha}\Big)\mathcal{M}'\\
&\quad - 2\phi_2(1-\tilde{\sigma}\mathcal{M})+2(1+\phi_2)\boldsymbol{m}_1^\top S\boldsymbol{m}_2\\
&\quad + \Big(1+\frac{\rho_2}{\alpha}\Big)(\boldsymbol{m}_2^\top S^\top MS\boldsymbol{m}_2-2\boldsymbol{m}_2^\top S\boldsymbol{m}_3)\\
&\quad + (\phi_1+2\phi_2)(1-\tilde{\sigma}\mathcal{M}-\boldsymbol{m}_1^\top S\boldsymbol{m_2})^2.
\end{aligned}$$

Thus we can write

$$
\begin{aligned}
e_{\mathrm{ICL}}(k_0, k_1) = e_{\mathrm{ICL}}(\text{independent query}) &- 2\phi_2 + q_{\mathrm{query}}(\mathcal{M} + \tilde{\lambda}\mathcal{M}') + \phi_2(2\sigma - \phi_2)\mathcal{M}' \\
&+ \ \boldsymbol{m}_2^\top S^\top M\, S\boldsymbol{m}_2 - 2\boldsymbol{m}_2^\top S\boldsymbol{m}_3 + 2\phi_2\tilde{\sigma}\mathcal{M} + 2(1+\phi_2)\boldsymbol{m}_1^\top S\boldsymbol{m}_2 \\
&+ \ (\phi_1 + 2\phi_2)\left(1 - \tilde{\sigma}\mathcal{M} - \boldsymbol{m}_1^\top S\boldsymbol{m}_2\right)^2 \\
&+ \frac{\rho_2}{\alpha}\Big[q_{\mathrm{query}}(\mathcal{M} + \tilde{\lambda}\mathcal{M}') + \phi_2(2\sigma - \phi_2)\mathcal{M}' + \boldsymbol{m}_2^\top S^\top M\, S\boldsymbol{m}_2 - 2\boldsymbol{m}_2^\top S\boldsymbol{m}_3 + 2\phi_2\mathcal{M}\Big].
\end{aligned}
$$